INTEGRATING TASK-SPECIFIC MODELS WITH LARGE LANGUAGE MODELS FOR CROSS-LINGUAL TRANSFER TO LOW-RESOURCE LANGUAGES

VIPUL KUMAR RATHORE

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY DELHI
JULY 2026

© Indian Institute of Technology Delhi – 2026

All rights reserved.

INTEGRATING TASK-SPECIFIC MODELS WITH LARGE LANGUAGE MODELS FOR CROSS-LINGUAL TRANSFER TO LOW-RESOURCE LANGUAGES

by

VIPUL KUMAR RATHORE
Department of Computer Science and Engineering

Submitted
in fulfillment of the requirements of the degree of
Doctor of Philosophy

to the

INDIAN INSTITUTE OF TECHNOLOGY DELHI
JULY 2026

Certificate

This is to certify that the thesis titled, **Integrating Task-Specific Models with Large Language Models for Cross-Lingual Transfer to Low-Resource Languages**, being submitted by **Mr. Vipul Kumar Rathore**, for the award of **Doctor of Philosophy** in Department of Computer Science and Engineering is a record of bonafide work carried out by him under our guidance and supervision at the **Department of Computer Science and Engineering, Indian Institute of Technology Delhi**. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma. In particular, work done in Chapters 3, 4, 5 and 6 were done jointly with undergraduate students. In each case, the part done by the collaborators appeared in their respective bachelor's thesis.

Prof. Parag Singla
Professor
Department of Computer Science and Engg.
Head, Yardi School of Artificial Intelligence
Indian Institute of Technology Delhi
New Delhi - 110016

Prof. Mausam
Professor
Department of Computer Science and Engg.
Indian Institute of Technology Delhi
New Delhi - 110016

Acknowledgements

As I am about to graduate, I want to thank everyone who consistently supported me throughout this journey.

First of all, I owe an immense debt of gratitude to my advisors, Prof. Parag Singla and Prof. Mausam, for their unwavering guidance and rigorous mentorship over the course of this PhD. Their ability to ask the right questions at the right time, their commitment to research integrity, and their confidence in my work – even when mine wavered – have shaped me as a researcher in ways I am only beginning to appreciate.

None of this would have been possible without my parents. Their quiet sacrifices and steady belief in me carried me through the toughest phases of this work, often without my saying a word about how hard those phases were. I am more grateful than I can express here.

I am equally grateful to my wife, Swati, for her patience through the late nights and lost weekends, for absorbing so much of what a PhD quietly takes from the people closest to it, and for never once letting me feel that I was asking too much.

I have made some lifelong friends in this journey – Soumen, Aritra, Naman, Prashant, Kunal, and Shubham. Thank you all for the amazing food, jokes, and ideas we shared over these years. I also thank my lab colleagues, especially Vishal, with whom I had countless technical discussions over coffee, for their wonderful support along the way.

It was a pleasure to have worked with some of the brightest undergraduates in the department – Kartikeya Badola, Rajdeep Singh Dhingra, Aniruddha Deb, Ankish Chandresh, and Malik Hammad Faisal. I learned a great deal from them.

I gratefully acknowledge the Prime Minister's Research Fellowship (PMRF) for supporting me financially during my PhD, and the alumni grants from the CSE department along with the IBM travel grant for supporting my conference travel. I would also like to thank the administrative staff at the Computer Science Department at IIT Delhi – especially Rekha ji – for their continuous help throughout these years.

A PhD is as much about personal growth as it is about academic contribution. It demands patience, teaches resilience, and rewards persistence. I leave this journey changed by it, and grateful for everyone who walked part of it with me.

Vipul Kumar Rathore

Abstract

Cross-lingual transfer (CLT) enables NLP systems for low-resource languages (LRLs) by leveraging labeled data from high-resource languages (HRLs) without requiring target-language annotation. This paradigm avoids costly corpus creation for each new language, offering a path toward scaling NLP support to the world's ~7,000 languages, most of which remain digitally underrepresented. Despite remarkable progress with multilingual pre-trained language models (mPLMs) and large language models (LLMs), both paradigms fail to adequately serve truly low-resource languages – particularly for *structured prediction* tasks requiring fine-grained, token-level reasoning. We identify two key dimensions along which existing CLT methods fall short: the *modeling paradigm* (task-specific fine-tuning vs. hybrid integration with LLMs) and the *task structure* (sequence labeling vs. distantly supervised relation extraction). Using these two dimensions, we organize our contributions into two halves of the thesis.

In the first half, we study CLT for *sequence labeling* tasks (NER and POS tagging). Most existing fine-tuning approaches rely on a single source language and require unlabeled target-language data for adaptation, limiting their applicability to truly unseen languages. We propose ZGUL, a multi-source language adapter ensembling framework that incorporates both local (token-level) and global (language-level) attention mechanisms to dynamically fuse knowledge from multiple source languages, enabling zero-shot generalization to unseen languages. At inference, ZGUL applies entropy minimization to adapt ensemble weights for the target language. Moving beyond fine-tuning, we propose SSP, a two-stage self-supervised prompting framework that first uses an off-the-shelf CLT model to generate noisy pseudo-labels, and then applies an Integer Linear Programming (ILP)-based exemplar selection strategy for monolingual in-context learning using these pseudo-labels. This bridges fine-tuned CLT models with LLMs, enabling effective prompting without any target-language annotations.

In the second half, we turn to *distantly supervised relation extraction* (DSRE), where even source-language supervision is inherently noisy. We propose PARE, which jointly encodes all sentences for an entity pair as a single passage, outperforming prior bag-level aggregation methods that encode sentences independently. However, PARE's bag-level training objective limits its precision at sentence-level inference and its robustness under cross-lingual transfer. We address these limitations with HYDRE, a hybrid inference-time framework that augments PARE with an LLM-as-selector mechanism guided by a novel relation-aware exemplar retrieval strategy from noisy training data. HYDRE delivers substantial gains in both English and cross-lingual RE. To support rigorous cross-lingual evaluation, we also curate INDICRE4, a new human-verified DSRE benchmark covering four typologically diverse, low-resource Indic languages.

We conclude by summarizing our contributions and proposing directions for future work that builds on the methods introduced in this thesis.

सार

क्रॉस–लिंगुअल ट्रांसफर (CLT) का उद्देश्य हाई–रिसोर्स भाषाओं (HRLs) के लेबल्ड डेटा का उपयोग करके लो–रिसोर्स भाषाओं (LRLs) के लिए एनएलपी सिस्टम बनाना है, वह भी टारगेट भाषा में किसी एनोटेशन के बिना। यह तरीका हर नई भाषा के लिए महँगा कॉर्पस बनाने की ज़रूरत से बचाता है और दुनिया की लगभग 7,000 भाषाओं तक एनएलपी की पहुँच बढ़ाने का रास्ता खोलता है, जिनमें से ज़्यादातर भाषाएँ डिजिटल रूप से कम प्रतिनिधित्व वाली हैं। मल्टिलिंगुअल प्री–ट्रेन्ड लैंग्वेज मॉडल (mPLMs) और लार्ज लैंग्वेज मॉडल (LLMs) में उल्लेखनीय प्रगति के बावजूद, ये दोनों पैराडाइम वास्तव में लो–रिसोर्स भाषाओं के लिए पर्याप्त काम नहीं करते – खास तौर पर उन स्ट्रक्चर्ड प्रेडिक्शन टास्क में जिनमें बारीक, टोकन–लेवल रीज़निंग की ज़रूरत होती है। हम दो मुख्य आयामों की पहचान करते हैं जिनमें मौजूदा CLT तरीके कमज़ोर पड़ते हैं: मॉडलिंग पैराडाइम (टास्क–स्पेसिफिक फाइन–ट्यूनिंग बनाम LLMs के साथ हाइब्रिड इंटीग्रेशन) और टास्क स्ट्रक्चर (सीक्वेंस लेबलिंग बनाम डिस्टेंटली सुपरवाइज्ड रिलेशन एक्सट्रैक्शन)। इन्हीं दो आयामों के आधार पर हम अपने योगदानों को इस शोध प्रबंध के दो हिस्सों में व्यवस्थित करते हैं।

पहले हिस्से में, हम सीक्वेंस लेबलिंग टास्क – नेम्ड एंटिटी रिकग्निशन (NER) और पीओएस (POS) टैगिंग – के लिए CLT का अध्ययन करते हैं। मौजूदा फाइन–ट्यूनिंग तरीके आम तौर पर एक ही सोर्स भाषा पर निर्भर रहते हैं और अडैप्टेशन के लिए टारगेट भाषा का अनलेबल्ड डेटा माँगते हैं, जिससे वास्तव में अनदेखी भाषाओं पर उनका उपयोग सीमित हो जाता है। हम ZGUL प्रस्तुत करते हैं – एक मल्टी–सोर्स लैंग्वेज अडैप्टर एन्सेम्बलिंग फ्रेमवर्क, जो लोकल (टोकन–लेवल) और ग्लोबल (लैंग्वेज–लेवल) दोनों अटेंशन मैकेनिज्म का उपयोग करते हुए कई सोर्स भाषाओं के ज्ञान को गतिशील रूप से जोड़ता है और इस तरह अनदेखी भाषाओं पर जीरो–शॉट जनरलाइज़ेशन संभव बनाता है। इन्फरेंस के समय ZGUL टारगेट भाषा के लिए एन्सेम्बल वेट्स को अडैप्ट करने के लिए एन्ट्रॉपी मिनिमाइजेशन लगाता है। फाइन–ट्यूनिंग से आगे बढ़ते हुए हम SSP प्रस्तुत करते हैं – एक दो–स्टेज वाला सेल्फ–सुपरवाइज्ड प्रॉम्प्टिंग फ्रेमवर्क, जो पहले एक उपलब्ध CLT मॉडल से नॉइज़ी स्यूडो–लेबल बनाता है और फिर इन स्यूडो–लेबलों का उपयोग करते हुए मोनोलिंगुअल इन–कॉन्टेक्स्ट लर्निंग (ICL) के लिए इंटीजर लीनियर प्रोग्रामिंग (ILP) आधारित एग्ज़ेम्प्लर सिलेक्शन रणनीति लगाता है। इस तरह यह फाइन–ट्यून्ड CLT मॉडलों को LLMs से जोड़ता है, जिससे टारगेट भाषा के किसी भी एनोटेशन के बिना असरदार प्रॉम्प्टिंग संभव हो जाती है।

दूसरे हिस्से में, हम डिस्टेंटली सुपरवाइज्ड रिलेशन एक्सट्रैक्शन (DSRE) की ओर बढ़ते हैं, जहाँ सोर्स भाषा का सुपरविज़न भी अपने आप में नॉइज़ी होता है। हम PARE प्रस्तुत करते हैं, जो किसी एंटिटी पेयर के सभी वाक्यों को एक ही पैसेज के रूप में मिलाकर एनकोड करता है और वाक्यों को अलग–अलग एनकोड करने वाले पहले के बैग–लेवल एग्रीगेशन तरीकों से बेहतर प्रदर्शन करता है। लेकिन PARE का बैग–लेवल ट्रेनिंग ऑब्जेक्टिव वाक्य–लेवल इन्फरेंस में उसकी प्रिसीज़न और क्रॉस–लिंगुअल ट्रांसफर के अंतर्गत उसकी रोबस्टनेस को सीमित कर देता है। इन कमियों को दूर करने के लिए हम HYDRE प्रस्तुत करते हैं – एक हाइब्रिड इन्फरेंस–टाइम फ्रेमवर्क, जो नॉइज़ी ट्रेनिंग डेटा से रिलेशन–अवेयर एग्ज़ेम्प्लर रिट्रीवल की एक नई रणनीति द्वारा निर्देशित LLM–ऐज़–सिलेक्टर मैकेनिज्म के ज़रिये PARE को और मज़बूत बनाता है। HYDRE अंग्रेज़ी और क्रॉस–लिंगुअल, दोनों RE में काफ़ी सुधार देता है। सख़्त क्रॉस–लिंगुअल मूल्यांकन के लिए हम INDICRE4 भी तैयार करते हैं – चार टाइपोलॉजिकली विविध, लो–रिसोर्स भारतीय भाषाओं को कवर करने वाला एक नया, मानव–सत्यापित DSRE बेंचमार्क।

अंत में, हम अपने योगदानों का सार प्रस्तुत करते हैं और इस शोध प्रबंध में दिए गए तरीकों पर आधारित भावी काम की दिशाएँ सुझाते हैं।

Contents

Acronyms xvii
Notation and Symbols xviii

1 Introduction **1**
1.1 Problem Definition and Scope 2
1.2 Challenges in Current Cross-Lingual Transfer (CLT) 4
1.3 Research Gaps and Contributions 6
1.4 Thesis Organization 7

2 Related Work and Background **9**
2.1 Early Approaches in Cross-Lingual Transfer (CLT) 10
2.1.1 Data-based Transfer 10
2.1.2 Model-based Transfer 12
2.2 Model Adaptation for Cross-Lingual Transfer 13
2.2.1 Supervised Fine-Tuning (SFT) 13
2.2.2 Parameter-Efficient Fine-Tuning (PEFT) Methods 14
2.2.2.1 Adapter-based Methods 14
2.2.2.2 Hypernetwork-based Parameter Generation 15
2.2.2.3 Lexicon- and Embedding-based Adaptation 16
2.3 Cross-lingual Prompting with Large Language Models (LLMs) 18
2.4 Hybrid Models: Combining Task-specific Models with LLM Reasoning 19
2.5 Evaluation Tasks and Metrics 20

3 Cross-lingual Generalization using Multi-source Language Adapter Ensembling **25**
3.1 Introduction 25
3.2 Preliminaries & Related Work 26
3.3 ZGUL: Zero-Shot Generalization via Multi-Source Adapter Ensembling 27
3.4 Experimental Setup 31
3.5 Results and Analysis 33
3.5.1 0-CLT Results 34
3.5.2 Robustness in U-CLT Setting 35

3.5.3 Qualitative and Error Analysis . . . 37
3.5.3.1 Global Fusion: Linguistic Interpretability . . . 37
3.5.3.2 EM: Effect on Predictions and Confidence . . . 38
3.5.4 Precision and Calibration Analysis . . . 40
3.6 Conclusion . . . 41

4 Self-Supervised Prompting for Cross-Lingual Transfer using Large Language Models 43
4.1 Introduction . . . 43
4.2 Background and Related Work . . . 45
4.3 Self-Supervised Prompting (SSP) . . . 46
4.3.1 A Thought Experiment . . . 46
4.3.2 Stage 1: Cross-Lingual Labeling . . . 47
4.3.3 Stage 2: Monolingual Prompting with Noisy Exemplars . . . 47
4.4 Experimental Setup . . . 48
4.5 Results and Analysis . . . 51
4.5.1 Key Results (0-CLT) . . . 51
4.5.2 Ablation Analysis . . . 53
4.5.3 Robustness under T-CLT and U-CLT Settings . . . 54
4.5.4 Exemplar Quality and Selection Behavior . . . 54
4.5.5 Scalability of SSP with Candidate Pool Size . . . 56
4.5.6 Scalability with Exemplar Count and Model Size . . . 58
4.6 Conclusion . . . 59

5 A Simple and Robust Baseline for Monolingual and Cross-Lingual DSRE 61
5.1 Introduction . . . 61
5.2 Related Work . . . 62
5.3 PARE: A Simple and Strong Supervised DSRE Method . . . 62
5.4 Indic Dataset Curation . . . 64
5.5 Experimental Setup . . . 66
5.6 Results . . . 68
5.6.1 Bag-level Evaluation . . . 68
5.6.2 Sentence-level Evaluation . . . 69
5.6.3 CLT Evaluation . . . 70
5.6.4 Qualitative & Error Analysis . . . 71
5.6.5 Analysis: Efficiency, Ablation, and Attention Behaviour . . . 73
5.7 Discussion and Conclusion . . . 74

6 Hybrid DSRE Approaches using Large Language Models **76**
6.1 Introduction . . . 76
6.2 Related Work . . . 77
6.3 HYDRE: The Hybrid DSRE Framework . . . 78
6.4 Experiment Design . . . 81
6.5 Results & Analysis . . . 82
6.5.1 Ablations . . . 85
6.5.2 Qualitative & Error Analysis . . . 86
6.5.3 Robustness to Bag-level Evaluation . . . 88
6.5.4 HYDRE Sensitivity to K . . . 88
6.5.5 Additional Results on IndoRE (Supervised RE Setting) . . . 90
6.5.6 Additional Results on Wiki-20m . . . 91
6.6 Discussion and Conclusion . . . 92

7 Conclusion and Future Work **94**
7.1 Conclusion and Summary . . . 94
7.2 Future Work . . . 95

A Appendix to Chapter 3 **98**
A.1 Implementation Details . . . 98
A.1.1 Hyperparameters . . . 98
A.1.2 Data Statistics . . . 98
A.1.3 Standard Deviations . . . 99
A.1.4 EM Algorithm . . . 99
A.2 Robustness with XLM-Roberta . . . 100
A.3 Few-Shot (FS-CLT) Performance . . . 101
A.3.1 Language-wise Few Shot Performance . . . 102

B Appendix to Chapter 4 **103**
B.1 Implementation Details . . . 103
B.1.1 Estimating Confidence $\hat{p}_i^{(l)}$ . . . 103
B.1.2 Prompt Details . . . 103
B.1.3 Dataset Details . . . 105
B.2 Results on Natural language Inference (NLI) . . . 105

C Appendix to Chapter 5 **109**
C.1 Data Annotation and Quality Assessment . . . 109
C.2 Additional Qualitative Examples . . . 110

D Appendix to Chapter 6 **112**
D.1 Prompt Details 112
D.2 Additional Results 113
D.2.1 Detailed Language-wise Results 113
D.2.2 Ablations to Assess Stage 3 Design 113
D.2.3 Additional Qualitative Analysis 114

Biography **151**

List of Figures

1.1 **Illustrative cross-lingual transfer failures (English → Hindi). (a)** Because English follows SVO order, an English-trained model expects a verb immediately after the subject. In Hindi (SOV), the object noun किताब (*book*) occupies this position and is misclassified as `VERB`, cascading errors to the actual verb पढ़ता (*reads*). **(b)** लखनऊ पीठ (*Lucknow Bench*) refers to a High Court branch (`ORG`), but the model tags लखनऊ as `B-LOC` based on English lexical priors and ignores the continuation पीठ (*Bench*) entirely. 3

1.2 **Thesis contributions** organized along two axes: *task* (sequence labeling vs. DSRE) and *modeling paradigm* (task-specific fine-tuning vs. hybrid). 8

2.1 **DSRE bag illustration** for head entity *Barack Obama* and tail entity *USA* illustrating the presence of both informative and noisy sentences. 22

3.1 **Dual-level fusion in ZGUL.** *Local* (token-level, left) and *Global* (language-vector, right) fusion outputs are combined and processed by the Task Adapter to produce the final representation at layer ℓ. 30

3.2 **Effect of unlabeled data scale**: average F1 over target languages vs. percentage of Wikipedia data used. *0%* denotes the 0-CLT setting. 37

3.3 **Interpretability of Global Fusion.** Learned attention weights (left) align with typological relatedness (right), confirming that ZGUL assigns higher weight to linguistically closer source languages. 38

4.1 **SSP pipeline.** *Stage 1*: Fine-tune a CLT model (e.g., ZGUL) or perform cross-lingual prompting using MRL data to generate pseudo-labels. *Stage 2*: Select exemplars from the pseudo-labeled pool via ILP and perform ICL using the target-language exemplars. 44

4.2 **Sensitivity to label noise**: GPT-4 POS tagging F1 on 3 Germanic LRLs with varying fractions of corrupted target-language exemplar labels. Dashed lines: performance with clean source MRL exemplars. 46

4.3 **Label diversity and correctness in SSP's exemplar selection** (Gothic POS). *TopK* retrieves redundant exemplars starting with "ƅa" (*What* in English) – sentences containing only 2 POS tags, collapsing the LLM's prediction space. SSP's ILP enforces label coverage across 7 tags, enabling correct predictions for PRON and VERB – labels that were absent from *TopK*'s demonstrations. Green: Correct label; Red: Incorrect label. 55
4.4 **Effect of candidate pool size** on GPT-4 F1 for *African NER* and *Germanic POS* under different retrieval strategies. 56
4.5 **Effect of K on African NER** with LLaMA-3.1-8B-Instruct. SSP outperforms *MMR*, *TopK*, and Random at every K. 58
4.6 **Effect of K on Germanic POS** with LLaMA-3.1-8B-Instruct. SSP outperforms *MMR*, *TopK*, and Random at every K. 59

5.1 **Example sentence-level DSRE query** in English and its Bhojpuri translation for the entity pair (**Joe_Biden**, **USA**). 61
5.2 **PARE Schematic Architecture**. The model jointly encodes entire passage of bag sentences via a shared PLM encoder, with a token-level attention mechanism pooling evidence across sentences to produce the final bag representation. Entity Markers omitted for brevity. 63
5.3 **Entity misalignment under erroneous projection**. The tail marker attaches to वाशिंगटन डी.सी. (Washington D.C.) instead of अमेरिका (USA), invalidating the relation judgment. 65
5.4 **Cross-sentence reasoning in PARE.** The three gold relations for (**Turkey**, **Ankara**) require evidence distributed across different sentences. PARE's passage-level encoding recovers all three, while CIL captures only the dominant *contains* pattern. 70
5.5 **PARE's bag size robustness**: AUC on NYT-10m across bins of untruncated bag length. PARE maintains a consistent advantage over CIL except for single-sentence bags. 74

6.1 **HYDRE pipeline** ($K = 3$). Given a query, PARE produces top-K candidate relations (r_1, r_2, r_3). For each candidate, the most relevant bag (B_1, B_2, B_3) is retrieved from the DSRE corpus. The most representative sentence from each bag is selected as an ICL exemplar and passed to the LLM alongside the query. 78
6.2 **Exemplar selection in HYDRE** (illustrated for the topmost candidate `place_of_birth`). *Left*: from bag B_1, the sentence with the highest aggregate confidence over both bag relations is selected (green tick). *Right*: the selected sentence is formatted as an in-context exemplar in the LLM prompt alongside the query. 79
6.4 **Bag-level prompting strategies** for GPT-4o on English NYT-10m: *zero-shot* vs. *full-bag* ICL vs. *reduced-bag* ICL. 88

6.3 **Qualitative examples** showing HYDRE's relation disambiguation on English (left) and Manipuri (right). Top boxes: ICL exemplars selected by HYDRE's pipeline. Bottom boxes: query and predictions from HYDRE, zero-shot, and PARE. 89
6.5 **Effect of candidate set size** (K) on PARE's Recall@K and downstream HYDRE performance (English, NYT-10m dev set). 90
6.6 HYDRE(GPT-4o) vs. PARE micro-F1 as a function of training-set percentage on IndoRE. 91

A.1 Few-shot F1 averaged over languages in a group for various few-shot bins. Top row: Germanic, Slavic. Bottom row: African, Indo-Aryan. 101
A.2 Language-wise few-shot performance results 102

B.1 **SSP's robustness to translation noise** (shown for Guarani language in NLI task). Stage 1: translate-test degrades due to NLLB translation errors. Stage 2: GPT-4-turbo correctly predicts the label despite a noisy exemplar, demonstrating SSP's ability to mitigate Stage-1 translation noise. 106
B.2 **Label coverage effect**: SSP with label constraint (left) correctly samples a `neutral` exemplar; without this constraint (right), no `neutral` is selected, causing a misprediction. 107
B.3 **English translations** for Exemplars shown in Fig. B.2 107

C.1 Additional error examples in English and Manipuri 111

D.1 **Confusion matrices for HYDRE vs. the baselines (part 1 of 2)**. 118
D.2 **Confusion matrices for HYDRE vs. the baselines (part 2 of 2)**. 119

List of Tables

1 **Acronyms and their Full Forms**. xvii

2 **Summary of Notations and Symbols** used in various contexts across the thesis. xviii

2.1 **Sequence labeling example** for a sentence in Bhojpuri. English Translation - "Former U.S. President Joe Biden has come on a visit to India ." 22

3.1 **Comparison of CLT paradigms** across four dimensions. ZGUL is the only method satisfying all four: *modularity*, *multi-source support*, *strict 0-CLT compatibility*, and *computational efficiency*. 28

3.2 **Language groups, source–target splits, and tasks.** *English is included as a common source in each language group. 31

3.3 **0-CLT POS tagging results** (micro-F1) on *Germanic* and *Slavic* target languages. Best results in **bold**; * denotes statistical significance (McNemar's $p < 0.005$) on aggregated group-level predictions. 35

3.4 **0-CLT NER results** (micro-F1) on *African* and *Indo-Aryan* target languages. Best results in **bold**; * denotes statistical significance (McNemar's $p < 0.005$) on aggregated group-level predictions. 36

3.5 **U-CLT results** (micro-F1) after incorporating unlabeled target-language data. Bottom row: size of unlabeled corpus $\mathcal{U}_t$ (in sentences). 36

3.6 **Spearman Correlation between attention weights and source-target linguistic similarity** (syntactic-genetic averaged). *Global*/*Local*: ZGUL's fusion networks; M^{multi}-*EM*: MADX^{multi}-EMEA baseline. 37

3.7 **POS (Germanic): True/Error counts and ECE**, with and without EM. 39

3.8 **NER (African): True/Error counts and ECE**, with and without EM. 39

3.9 **Efficiency comparison**: trainable parameters and per-epoch training time (minutes) across language groups. 40

3.10 **Expected Calibration Error (ECE)** (rounded to two decimal places) across target languages for Germanic POS and African NER. 40

3.11 **Class-wise performance on African NER** (averaged across all target languages) (P: Precision, R: Recall). 41

3.12 **Class-wise performance on Germanic & Slavic POS tags** (averaged across all target languages). 41

4.1 **SSP results**: micro-F1 for *African NER* (left) and *Germanic POS* (right) across CLT settings. **Bold**: best 0-CLT result; underline: overall best. Translate-train omitted for POS (due to unreliable projection); translate-test uses GPT-4 (Gothic unsupported by NLLB-200). Bold Avg. column p-values: 0.008 (NER), 0.0004 (POS). 52
4.2 **Ablation of ILP components** in SSP(ZGUL)-GPT-4x on *African NER* (left) and *Germanic POS* (right). **Bold**: Best per-language results. 53
4.3 **Exemplar quality** of SSP vs. the baselines across four metrics. All scores reported in the format African NER / Germanic POS. 56
4.4 **Selection efficiency**: average runtime (ms/query) across retrieval strategies. . . 57

5.1 **Indic test language characteristics**: typological diversity, resource availability, and official status in India. 65
5.2 **Distantly supervised bag-level evaluation** on the English NYT-10d dataset (left) and the multilingual DiS-ReX dataset (right). Best results in **bold**, and the second best ones underlined. * McNemar's p value $< 10^{-5}$. 69
5.3 **English bag-level and sentence-level evaluation** (micro/macro-F1) on NYT-10m, Wiki-20m and IndoRE. Bag-level IndoRE column is omitted as IndoRE is a sentence-level-only benchmark without a multi-instance bag structure. * McNemar's p value $< 10^{-5}$. † Could not run other baselines on IndoRE due to time constraints. 69
5.4 **Sentence-level cross-lingual evaluation on** INDICRE4 evaluation sets and English evaluation on the corresponding stratified NYT-10m evaluation set (micro/macro-F1). Language-wise results in §D.2.1. * McNemar's p value $< 10^{-5}$. 70
5.5 **Sentence-level cross-lingual evaluation on IndoRE** for Hindi, Bengali and Telugu. * McNemar's p value $< 10^{-5}$. 71
5.6 **PARE's frequent relation confusions** on NYT-10m (English) and INDICRE4 (Manipuri and Santali) under translate-train. Below dashed line: false abstention (non-NA → NA), the dominant error mode. 72
5.7 **PARE's representative error cases** on English and Manipuri (translate-train). Underlined tokens indicate the key semantic trigger. In Manipuri, the locative postposition ꯗꯤ ("at/in") suffixed to **Greenwich** mimics a residence pattern (row 4), and the explicit cue ꯏꯟꯖꯤꯜ ("capital") is ignored in favor of the high-frequency `contains` relation (row 5). Colors: **Head entity**, **Tail entity**, **Correct**, **Wrong**, **Partially Correct**. 73
5.8 **Comparison of trainable parameters** between PARE and other baselines . . . 73
5.9 **PARE ablation**: change in AUC (Δ AUC) with different components of PARE . 74
5.10 **PARE's semantic trigger identification**: attention analysis on correctly predicted relations on NYT-10m. **Bold** tokens: highest-attended token per instance. 75

6.1 **Recall@5** for PARE and CIL on English (NYT-10m) and Indic languages (INDICRE4, translate-train). PARE's higher recall motivates its use as the candidate filter in HYDRE. 78
6.2 **HYDRE configuration** used across various settings. 82
6.3 **Main results** (micro/macro-F1) on English (NYT-10m) and Indic languages (INDICRE4, averaged over 4 languages). Indic results span three CLT settings; per-language breakdowns in §D.2.1. **Bold**: best per-column; *: statistically significant (McNemar's $p < 10^{-5}$). 84
6.4 **Ablation study for HYDRE** on English NYT-10m and INDICRE4 (translate-train), reporting micro/macro-F1. **L3.1**: LLaMA-3.1-8B; **L3.1-FT**: LLaMA-3.1-8B-FT; **Qw3**: Qwen3-235B; **G4o**: GPT-4o. "–": not applicable (omitted for Indic languages due to excessive prompt length from high token fertility). . 85
6.5 **Confusion analysis for HYDRE vs. the baselines**, aggregated over all relations (English, NYT-10m). 87
6.6 **HYDRE vs. PARE confusion on gold non-NA instances** (English NYT-10m). HYDRE recovers **166** relations where PARE defaults to NA, while losing only 6. . 87
6.7 HYDRE(GPT-4o) ablations on IndoRE. Bold: best variant. 91
6.8 F1 scores (micro/macro) on Wiki-20m test set. 92

A.1 **ZGUL hyperparameter grid details** 98
A.2 **Training and development set sizes** (sentences) per language group. English included in all groups. 98
A.3 **Target-language test sets**: script, size (sentences), and token-level overlap (%, IOU) with source training data. 99
A.4 **Standard deviation for ZGUL vs. the baselines** over micro-F1 (rounded to 1 decimal). *Avg* column: std. dev. of the group-average F1, not the average of per-language std. devs. 99
A.5 **EM weight granularity ablation**: token-wise trainable weights (ZGUL) vs. layer-wise tied weights (used in original EMEA). 100
A.6 **XLM-R robustness of ZGUL**: Indo-Aryan NER (trained on En,Hi,Ur) and Germanic POS (trained on En, Is, De) using XLM-R-Base adapters. 100

B.1 **Dataset statistics**: source training size (sentences) and average target test size (labels per language). Es: Spanish, Aym - Aymara, Gn: Guarani, Nah: Nahuatl. 105
B.2 **NLI results for SSP** (micro-F1 for three indigenous languages in AmericasNLI dataset). **Bold**: best 0-CLT; underline: overall best. * Nahuatl language not supported in NLLB-200. McNemar's $p = 0.054$ for bold results. 108
B.3 **Labelwise Recall** for fine-tuned stage 1 model (mDeBERTa^{FT}) and SSP variants – w. and w/o Label coverage (GPT-4-Turbo). 108
B.4 **Effect of candidate pool size** on SSP's per-language F1. 108

C.1 **Relation ontology and INDICRE4 label distribution.** Label count (722) exceeds sentence count (538) due to multi-label instances. 109
C.2 **INDICRE4 quality assessment**: percentage of system translations requiring no human rectification, and character-level F1 similarity for rectified translations. 110
C.3 **INDICRE4 inter-annotator agreement** for 100 samples in each language . . . 110

D.1 **HYDRE's language-wise results (0-CLT)**: micro/macro-F1 on English and Indic languages. †Averages over non-English languages. *McNemar's $p < 10^{-5}$. 114
D.2 **HYDRE's language-wise results (translate-train)**: micro/macro-F1 on Indic languages. *McNemar's $p < 10^{-5}$. 115
D.3 **HYDRE's language-wise results (translate-test)**: micro/macro-F1 on Indic languages. *McNemar's $p < 10^{-5}$. 115
D.4 **HYDRE's stage 3 design ablation** (micro/macro-F1, English NYT-10m). . . . 116
D.5 **Multi-label coverage.** Comparison of recall metrics on multi-label queries, demonstrating HYDRE's ability to actively recover multiple relations. 116
D.6 **Impact of single-label prompting.** Performance breakdown (Micro-F1) of GPT-4o on single-label vs. multi-label queries when restricting exemplars to a single relation label. 117
D.7 **Impact of DSRE candidate recall.** Performance breakdown based on whether the gold relation is present in PARE's top-5 retrieved candidates. 117

Acronyms

Acronym	Full Form
CLT	cross-lingual transfer
0-CLT	zero-resource CLT
T-CLT	translation-based CLT
U-CLT	unlabeled-data-based CLT
HRL	high-resource language
LRL	low-resource language
MRL	medium-resource language
PLM	pre-trained language model
LLM	large language model
mPLM	multilingual PLM
mBERT	multilingual BERT
XLM-R	XLM-Roberta
NER	named entity recognition
POS	part-of-speech tagging
RE	relation extraction
DSRE	distantly supervised relation extraction
NLI	natural language inference
ZGUL	zero-resource generalization to unseen languages
SSP	self-supervised prompting
PARE	passage-attended relation extraction
HYDRE	hybrid DSRE system
INDICRE4	Indic relation extraction benchmark (4 languages)
MIML	multi-instance multi-label
SFT	supervised fine-tuning
PEFT	parameter-efficient fine-tuning
ICL	in-context learning
MAD-X	multilingual adapter for cross-lingual transfer
LA	language adapter
TA	task adapter
LoRA	low-rank adaptation
MLM	masked language modeling
CPG	conditional parameter generation
EM	entropy minimization
CIL	contrastive instance learning
MMR	maximal marginal relevance
ILP	integer linear programming
CODEC	constrained decoding for cross-lingual label projection

Table 1: **Acronyms and their Full Forms**.

Notation and Symbols

Symbol	Description
$\mathcal{T}$	A structured prediction task (e.g., NER, POS tagging)
$\mathcal{L}_s \in \mathcal{S}$	A source language $\mathcal{L}_s$ belonging to a set of source languages $\mathcal{S}$
$\mathcal{L}_t$	A target low-resource language
$\mathcal{D}_s$, ($\mathcal{D}_t$)	Labeled dataset for a source language $\mathcal{L}_s$ (or a target language $\mathcal{L}_t$)
Multi-source and Modular CLT (ZGUL)	
$\text{LA}_{\mathcal{L}_s}$	Language adapter for a language $\mathcal{L}_s$
$\text{TA}_{\mathcal{T}}$	Task adapter for a task $\mathcal{T}$
$v_{\mathcal{L}}$	Dense typological feature vector for a given language $\mathcal{L}$
α_s	Global (typological) fusion weight for source language $\mathcal{L}_s$
$\beta_{j,s}^{(\ell)}$	Local (token-level) attention weight for source $\mathcal{L}_s$ at token x_j in layer ℓ
$o_{j,s}^{(\ell)}$	Output of the s^{th} language adapter for token x_j in layer ℓ
$\tilde{o}_{j,\text{global}}^{(\ell)}$	Global typology-aware mixture output for token x_j in layer ℓ
$\tilde{o}_{j,\text{local}}^{(\ell)}$	Local token-level mixture output for token x_j in layer ℓ
$z_j^{(\ell)}$	Combined multi-source representation for token x_j in layer ℓ
Passage-Level Relation Extraction (PARE)	
$B(e_1, e_2)$	Bag of sentences mentioning an entity pair (e_1, e_2) (e_1: head entity, e_2: tail entity).
$\mathbf{z}_i$	Contextualized representation of token w_i generated by the PARE encoder
v_{r_j}	Learnable relation-specific query vector for relation r_j
α_i^j	Relation-specific attention weight for token w_i given relation r_j
$\mathbf{z}_{(e_1, r_j, e_2)}$	Relation-conditioned summary representation for the entity pair (e_1, e_2)
p_j	Predicted probability that relation r_j holds between entity pair (e_1, e_2)
Hybrid Prompting (SSP & HYDRE)	
$\mathcal{Y}$	Task-specific label set (ontology)
q	Target query instance
K	Number of in-context exemplars provided to the LLM
$\hat{y}_i^{(t)}$	Predicted pseudo-label sequence for the i^{th} candidate instance in language $\mathcal{L}_t$
$\hat{p}_i^{(l)}$	Aggregate confidence score for label l in the i^{th} candidate instance
s_i	Semantic similarity score between the i^{th} candidate and the query q
z_i	Binary decision variable (ILP) indicating if the i^{th} candidate is selected
τ_l	Confidence threshold for label l during ILP selection
$f_{\text{PARE}}(q, r)$	PARE's confidence score for relation r given query q
$\mathcal{R}'$	top-K candidate relations selected by PARE
SSP(M_1)-M_2	SSP performed with stage 1 using model M_1 and stage 2 using model M_2
HYDRE(M)	HYDRE performed using LLM M
LM-MMR	LLM prompting using MMR-based retrieval

Table 2: **Summary of Notations and Symbols** used in various contexts across the thesis.

Chapter 1

Introduction

Over the past decade, Natural Language Processing (NLP) has advanced rapidly, driven by large-scale pre-trained language models (PLMs) such as BERT, RoBERTa, and T5 (Devlin et al., 2019; Liu et al., 2019; Raffel et al., 2020), and more recently by large language models (LLMs) including GPT-3, GPT-4, LLaMA-3, and Qwen-3 (Brown et al., 2020; Achiam et al., 2023; Grattafiori et al., 2024; Yang et al., 2025a). Trained on massive multilingual corpora, these models achieve strong generalization through language modeling and instruction tuning (Devlin et al., 2019; Radford et al., 2019; Zoph et al., 2022; Ouyang et al., 2022). As a result, they have enabled state-of-the-art performance across a wide range of applications, including neural machine translation (NMT), dialogue systems, and knowledge base construction (KBC) (Lewis et al., 2020; Wu et al., 2020a; Zhou et al., 2022).

Despite this progress, advances in NLP remain highly *uneven* across languages (Joshi et al., 2020; Bender et al., 2021; Blasi et al., 2022). Only a small subset of high-resource languages (HRLs) – notably English, Chinese, French, German, and Spanish, which collectively represent ~0.1% of the world's ~7000 languages – benefits from abundant digital resources (Joshi et al., 2020), curated benchmarks (Conneau et al., 2018; Liang et al., 2020) and sustained research attention. Indeed, nearly 93% of NLP papers at top conferences target English alone (Joshi et al., 2020). In contrast, over 90% of the world's languages are classified as low-resource (Hedderich et al., 2021), and remain highly underrepresented, even in modern multilingual PLMs (mPLMs) and LLMs, which typically cover only 100–200 languages and rely on pretraining corpora heavily skewed toward HRLs (Conneau et al., 2020; Workshop et al., 2022; Muennighoff et al., 2023).

This imbalance has profound societal consequences: nearly 88% of the world's languages, spoken by over 1.2 billion people, receive little to no support from modern language technologies (Joshi et al., 2020). Many of these languages are spoken by large populations but lack annotated datasets, pretrained models, or basic digital infrastructure. For example, Quechua – a South American language spoken by millions – remains severely under-resourced for NLP despite prior efforts (Melgarejo et al., 2022). Similar gaps persist for languages such as Bhojpuri (~50 million speakers), Oriya, Santali, and several African languages, including Swahili and

Amharic (Adelani et al., 2021). The prolonged absence of NLP support for these demographics directly exacerbates digital inequality, substantially limiting global access to information, education, and technology (Bender et al., 2021).

Consequently, a central research focus in multilingual NLP has shifted (Ruder et al., 2019). Rather than emphasizing continuous performance gains achieved through ever-larger, web-scale data (Kaplan et al., 2020; Brown et al., 2020), current research increasingly asks how models can generalize reliably in settings where labeled data are scarce, noisy, or entirely unavailable (Hu et al., 2020b; Lauscher et al., 2020). Addressing this challenge requires methods that are data-efficient, adaptable to new languages, and capable of transferring subtle linguistic structures across typologically diverse low-resource languages.

1.1 Problem Definition and Scope

A principled approach to addressing linguistic inequity in NLP is *cross-lingual transfer (CLT)* (Ruder et al., 2019), which leverages supervision and representations learned from high-resource languages (HRLs) to enable effective task performance in low-resource languages (LRLs). Here, *low-resource* denotes a lack of sufficient unlabeled web data or task-specific labeled data in a language, regardless of its native speaker population (Joshi et al., 2020).

Formally, given labeled data $\mathcal{D}_s = \{(x_i^{(s)}, y_i^{(s)})\}$ for a task $\mathcal{T}$ in one or more source languages $\mathcal{L}_s \in \mathcal{S}$, and scarce or absent labeled data $\mathcal{D}_t$ in a target language $\mathcal{L}_t$ – ranging from zero examples (0-CLT) to a small handful (FS-CLT) – the objective of CLT is to learn a model f_θ that generalizes effectively to $\mathcal{L}_t$. An ideal CLT system should accommodate varying levels of available supervision while operating robustly across diverse linguistic landscapes.

Language Scope. This thesis specifically focuses on target languages that are not only low-resource but also *typologically distant* from commonly studied HRLs – differing substantially in script, morphology, and syntax (Littell et al., 2017). As discussed in §1.2, bridging this structural gap between languages remains one of the most significant challenges for robust CLT.

Task Scope. While CLT achieves competitive results on sentence-level classification tasks (Phan et al., 2021; De Raedt et al., 2023), this thesis focuses on *structured prediction* – specifically *sequence labeling* (POS tagging and NER) and *distantly supervised relation extraction* (DSRE). POS tagging and NER require token-level labeling with interdependent output labels, while DSRE extends this to relation prediction under noisy, bag-level supervision (Mintz et al., 2009). Together, these tasks form the backbone of multilingual information extraction pipelines such as KBC. Unlike sentence-level classification, structured prediction is highly sensitive to cross-lingual divergence in morphology and syntax (as illustrated in Figure 1.1), making it a rigorous testbed for evaluating CLT robustness (Bareket and Tsarfaty, 2021; Martínez-García

et al., 2021; Parekh et al., 2024). Formal task definitions, label sets, and evaluation metrics are provided in §2.5 (Chapter 2).

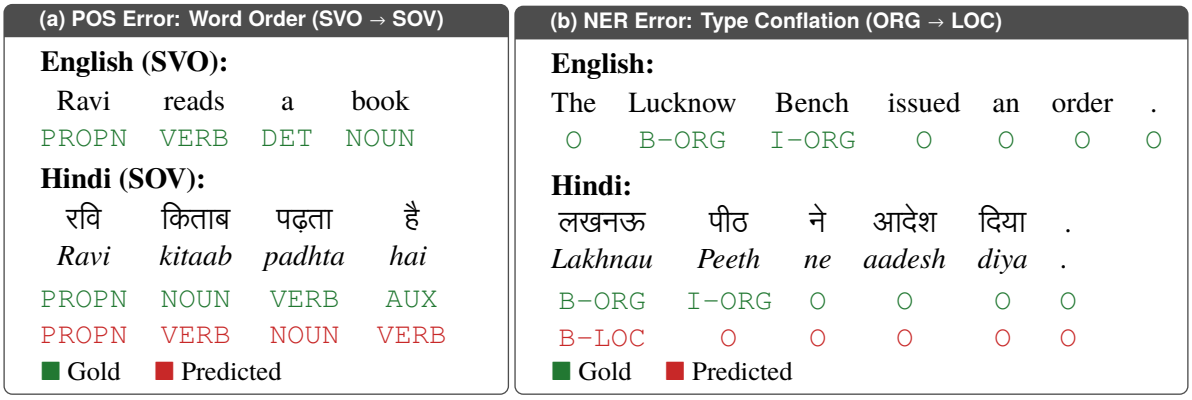


Figure 1.1: **Illustrative cross-lingual transfer failures (English → Hindi). (a)** Because English follows SVO order, an English-trained model expects a verb immediately after the subject. In Hindi (SOV), the object noun किताब (*book*) occupies this position and is misclassified as VERB, cascading errors to the actual verb पढ़ता (*reads*). **(b)** लखनऊ पीठ (*Lucknow Bench*) refers to a High Court branch (ORG), but the model tags लखनऊ as B-LOC based on English lexical priors and ignores the continuation पीठ (*Bench*) entirely.

Supervision Scope. We primarily study *multi-source cross-lingual transfer*, where labeled data from multiple HRLs are jointly leveraged to capture more universal linguistic regularities, reduce over-reliance on English, and mitigate language-specific biases (Üstün et al., 2020; Alabi et al., 2022). Within this setting, we categorize CLT into four target-language supervision regimes:

- **0-CLT (Zero-resource transfer):** No target-language data are available; only HRL data is utilized.
- **T-CLT (Translation-based transfer):** Machine translation is used at training (*Translate-train*) or inference (*Translate-test*) time.
- **U-CLT (Unlabeled-data transfer):** External unlabeled target-language data are used for unsupervised adaptation, e.g., via continued pretraining or language-adaptive fine-tuning (Pfeiffer et al., 2020; Chau et al., 2020).
- **FS-CLT (Few-shot transfer):** A small number of labeled target-language instances are available, but the data is orders of magnitude smaller compared to the task-specific labeled dataset of the source language.

While FS-CLT represents a practically important and well-studied setting (Schmidt et al., 2022; Xu and Murray, 2022), this thesis focuses on the 0-CLT, T-CLT, and U-CLT regimes, which do not assume any *labeled* target-language data.

Modeling Paradigm. Modern CLT systems typically rely on multilingual pretrained language models (mPLMs), such as mBERT and XLM-R (Devlin et al., 2019; Conneau et al., 2020), which are fine-tuned on task-specific HRL data and applied directly to LRLs in a zero-shot manner (Muller et al., 2021). While this paradigm achieves competitive performance on simple classification tasks (Phan et al., 2021), it remains fragile for structured prediction tasks defined above, as cross-lingual divergence induces cascading token-level errors. This thesis addresses these limitations by exploring *hybrid modeling paradigms* that combine the structured precision of task-specific supervised models with the broad semantic reasoning of large language models (LLMs).

While our empirical evaluations are grounded in standard benchmarks for the aforementioned tasks, the broader objective is to develop cross-lingual adaptation mechanisms that generalize across unseen languages, tasks, and supervision regimes. Ultimately, this thesis investigates the interplay between linguistic diversity, task complexity, and supervision quality to design methods that remain robust under acute resource constraints.

1.2 Challenges in Current Cross-Lingual Transfer (CLT)

Despite substantial progress with mPLMs, reliable and scalable CLT remains challenging – particularly for low-resource and typologically diverse languages. Performance often degrades sharply, dropping by 20-40 F1 points in several structured prediction benchmarks (Hu et al., 2020b; Lauscher et al., 2020), when models trained on HRLs are applied directly to LRLs. These limitations stem from a combination of factors spanning linguistic diversity, representational imbalance, task complexity, and rigid adaptation strategies.

Typological and Script Diversity. Human languages vary widely in script, morphology, and syntax, yet mPLMs rely on shared subword vocabularies that are heavily biased toward high-resource, Latin-script languages (Conneau et al., 2020). This bias leads to fragmented tokenization and poor lexical coverage for many scripts, including Devanagari, Ol Chiki, or Amharic (Pfeiffer et al., 2021b; Nag et al., 2024). For example, characters in Santali (Ol Chiki) may be decomposed into multiple unknown or byte-level tokens, whereas English words are represented as meaningful subwords. These tokenization gaps degrade representations before any higher-level reasoning can even occur.

Representational Asymmetry. At its core, CLT's limited effectiveness for LRLs stems from their *inherently small pretraining corpora*: both mPLMs and LLMs are trained on web-scale data that disproportionately favours a handful of dominant HRLs (Conneau et al., 2020; Workshop et al., 2022), leaving LRL representations inherently weak regardless of model scale or architecture. This manifests as *representation-space imbalance* – LRL representations are weakly aligned and occupy poorly calibrated regions of the shared embedding space (Blevins et al.,

2022). This statistical asymmetry persists even when tokenization issues are not severe, significantly limiting CLT robustness when corrective fine-tuning cannot be applied.

Structural Fragility under Weak Supervision. Structured prediction tasks further expose the fragility of current CLT systems. Because these tasks require explicit modeling of token boundaries and interdependent output labels, they lack the error tolerance of simple classification tasks (Parekh et al., 2024). Consequently, most CLT approaches struggle to preserve structural consistency across languages (Subburathinam et al., 2019). In weakly or distantly supervised settings, translation noise, segmentation errors, and entity misalignment easily propagate through entire prediction sequences (Chen et al., 2025), especially in morphologically rich or free-word-order languages.

These gaps manifest as cascading failures (Figure 1.1) in practice: an English-trained POS tagger misclassifies Hindi tokens due to SVO → SOV word order divergence, while an NER model conflates organization and location entities by over-relying on source-language lexical priors.

Rigidity and Inefficiency of Adaptation. Furthermore, existing adaptation strategies are often computationally inefficient and structurally rigid. Fully fine-tuning large mPLMs for every language-task pair is not only impractical at scale but also entangles language- and task-specific representations (Pfeiffer et al., 2020). This rigidity is particularly detrimental in typologically distant transfer scenarios – for instance, from an SVO language like English to an SOV language like Bhojpuri – where direct transfer performs poorly. In such cases, a rigid pipeline fails to effectively leverage *modular* supervision from typologically related *pivot* languages (e.g., Hindi for Bhojpuri), which could otherwise provide a strong structural anchor for transfer.

Can LLMs Bridge the Gap? A natural question is whether LLMs – trained on web-scale multilingual corpora – can circumvent these limitations through in-context learning (ICL) (Brown et al., 2020) capabilities. While LLMs demonstrate strong zero- and *few-shot* performance in HRLs (where few-shot refers to in-context exemplars rather than labeled target-language data), their effectiveness in LRLs remains limited by the same root cause – small pretraining corpora for LRLs – that constrains mPLMs (Ahuja et al., 2023; Conneau et al., 2020). This is compounded by their difficulty with structured prediction tasks requiring precise, token-level reasoning (Blevins et al., 2023; Parekh et al., 2024). Moreover, effective ICL relies on *clean*, *task-aligned* exemplars, which are inherently unavailable under zero-resource (0-CLT) conditions. Thus, while LLMs offer powerful complementary reasoning capabilities, they do not obviate the need for principled cross-lingual transfer mechanisms.

1.3 Research Gaps and Contributions

Building on the representational and structural limitations discussed in §1.2, we identify three concrete research gaps that limit the reliability and scalability of current CLT systems for structured prediction. These gaps motivate a coherent trajectory of contributions – spanning modular fine-tuning (ZGUL (Rathore et al., 2023)), self-supervised prompting (SSP (Rathore et al., 2024)), and hybrid reasoning under distant supervision (PARE (Rathore et al., 2022) and HYDRE (Rathore et al., 2025)) – to achieve efficient and robust CLT.

1. **Gap 1: Absence of Modular, Multi-source Systems for Strict 0-CLT.**
   Robust cross-lingual transfer under strict zero-resource (0-CLT) conditions requires systems that can reuse and combine task knowledge across multiple source HRLs. However, most existing approaches violate strict 0-CLT assumptions by implicitly relying on target-language text for adaptation or vocabulary alignment (Pfeiffer et al., 2020). When restricted to true 0-CLT settings, systems typically default to single-source transfer (usually English), which fails to capture relevant linguistic structures for typologically distant LRLs (He et al., 2021b). Crucially, current frameworks lack a unified, modular mechanism to dynamically fuse task knowledge from multiple HRLs. Note that while few-shot transfer (FS-CLT) is a related setting, it introduces distinct considerations and is deferred to future work (§7).
   Objective: Develop a modular, data-efficient CLT framework that aggregates supervision from multiple typologically related source languages without requiring any target-language data.
   Contribution – ZGUL: Chapter 3 introduces ZGUL, which *fuses* information from multiple source HRLs via a dual-level ensembling mechanism, enabling highly effective zero-resource transfer for sequence labeling tasks.

2. **Gap 2: Lack of Task-Aware, Monolingual Prompting Mechanisms for LRLs.**
   In 0-CLT settings, LLM prompting suffers from the absence of target-language exemplars. Existing prompting strategies try to mitigate this by retrieving translated or source-language (English) exemplars. However, this introduces *semantic drift* and alignment errors that weaken the LLM's contextual reasoning (Ahuja et al., 2023; Kim et al., 2024). Furthermore, standard exemplar retrieval is *static* and *task-agnostic* (e.g., relying solely on semantic similarity). This is markedly suboptimal for structured prediction tasks, which require nuanced, label-aware demonstrations to prevent the model from collapsing to majority classes (Blevins et al., 2023).
   Objective: Develop a self-adaptive prompting framework that autonomously synthesizes *high-quality*, task-specific target-language exemplars without requiring any target-language annotations.
   Contribution – SSP: Chapter 4 introduces *Self-Supervised Prompting (*SSP*)*, which for-

mulates exemplar selection as an Integer Linear Programming (ILP) problem. By leveraging pseudo-labels generated by existing CLT models, SSP enforces label diversity and model confidence to construct optimal monolingual (target-language) exemplars. This yields consistent gains across sequence labeling benchmarks (POS, NER), as well as sentence-level tasks such as Natural Language Inference (NLI).

3. **Gap 3: Fragility of Cross-Lingual Structured Prediction under Distant Supervision.** Most CLT methods presume clean, manually annotated source data in HRLs. This assumption fails in distantly supervised relation extraction, where HRL supervision is generated heuristically and is inherently noisy (Riedel et al., 2010). In cross-lingual scenarios, translation artifacts, syntactic divergence, and entity misalignment severely amplify the distant supervision noise (Chen et al., 2025). Existing frameworks lack principled mechanisms to disambiguate relations under such compounded noise, keeping DSRE largely confined to English or European languages (Bhartiya et al., 2022) having Roman scripts.
   Objective: Develop a cross-lingual DSRE framework capable of robust inference-time disambiguation by mitigating both inherent distant supervision noise and cross-lingual transfer artifacts.
   Intermediate Contribution – PARE: Chapter 5 introduces PARE (Passage-Attended Relation Extraction). By encoding all sentences in a DSRE bag as a unified passage with token-level attention, PARE significantly improves intra-bag reasoning and provides a high-recall candidate filter.
   Contribution – HYDRE: Chapter 6 introduces *Hybrid Distantly Supervised Relation Extraction* (HYDRE), a hybrid inference-time framework that integrates PARE with LLMs. HYDRE leverages PARE's high-recall filtering to construct efficient, task-aware exemplars from noisy bags. By combining structured retrieval with LLM reasoning, HYDRE enables robust relation disambiguation under noisy cross-lingual supervision.

These four contributions organize naturally along two axes – *task* (sequence labeling vs. DSRE) and *modeling paradigm* (fine-tuning vs. hybrid) – forming a coherent 2×2 framework for scalable multilingual structured prediction (Figure 1.2).

1.4 Thesis Organization

The remainder of this thesis is organized as follows:

- **Chapter 2: Related Work** surveys multilingual representation learning, parameter-efficient adaptation, cross-lingual structured prediction, and LLM-based prompting, situating each contribution within the broader literature.
- **Chapter 3: ZGUL** addresses Gap 1 by introducing a modular, multi-source adapter ensembling framework for zero-resource cross-lingual transfer in sequence labeling.

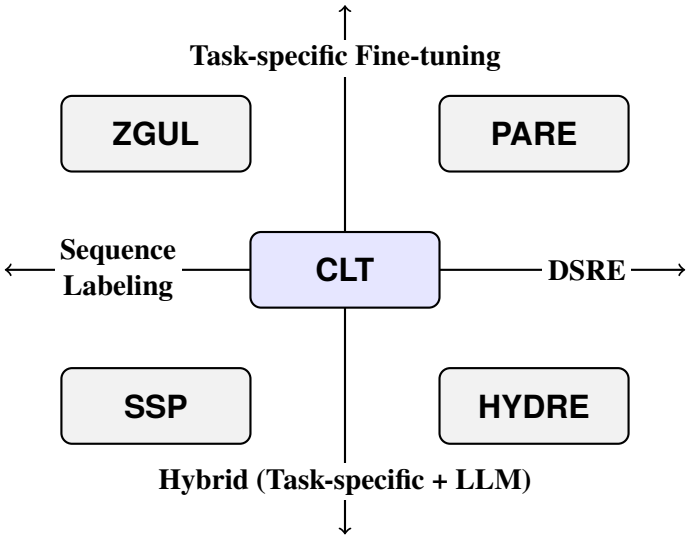


Figure 1.2: **Thesis contributions** organized along two axes: *task* (sequence labeling vs. DSRE) and *modeling paradigm* (task-specific fine-tuning vs. hybrid).

- **Chapter 4: SSP** addresses Gap 2 by developing a self-supervised prompting framework that constructs task-aware exemplars from pseudo-labels to enable effective cross-lingual LLM adaptation.
- **Chapter 5: PARE** addresses the first half of Gap 3 by proposing passage-level encoding for distantly supervised relation extraction, establishing a high-recall candidate retrieval backbone.
- **Chapter 6: HYDRE** addresses the second half of Gap 3 by presenting a hybrid inference-time framework that combines PARE's candidate filtering with LLM-based relation disambiguation.
- **Chapter 7: Conclusion and Future Directions** summarizes the core contributions of this thesis and outlines open research directions for scalable cross-lingual transfer.

Chapter 2

Related Work and Background

Chapter 1 established the cross-lingual transfer (CLT) paradigm, briefly discussed the structured prediction tasks studied in this thesis, and identified the three concrete research gaps motivating our contributions. This chapter surveys the prior work underlying those gaps.

Research in CLT has evolved through several distinct methodological paradigms. Early approaches leveraged multilingual pretrained language models (mPLMs) to establish shared representations for transfer with minimal target-language supervision (Wu and Dredze, 2019; Conneau et al., 2020). To scale this adaptation, subsequent research introduced parameter-efficient fine-tuning (PEFT) methods (Pfeiffer et al., 2020). More recently, in-context learning (ICL) has emerged as an alternative inference-time adaptation paradigm for low-resource languages (Brown et al., 2020; Ahuja et al., 2023). However, in isolation, these paradigms often fall short on complex extraction tasks that demand subtle linguistic or relational reasoning. This has motivated hybrid frameworks that combine task-specific models with Large Language Models (LLMs). Throughout this evolution, structured prediction tasks – namely sequence labeling and distantly supervised relation extraction (DSRE) – have acted as rigorous diagnostic testbeds to evaluate how well these systems generalize across diverse linguistic landscapes.

To systematically synthesize this progression, we map how *modeling paradigms*, *supervision granularity* and *task structure* intersect across five concrete themes:

1. **Early Approaches:** Data- and model-centric strategies for CLT and their limitations in low-resource settings.

2. **Parameter-Efficient Fine-Tuning (PEFT):** Modular and efficient adaptation mechanisms for mPLMs.

3. **Cross-Lingual Prompting:** In-context learning paradigms for low-resource languages and their limitations for structured prediction.

4. **Hybrid Reasoning Frameworks:** Architectures that integrate task-specific models with LLM reasoning for complex tasks.

5. **Structured Prediction as a Diagnostic Testbed:** Formal definitions of sequence labeling and DSRE, along with evaluation settings and metrics used throughout this thesis.

2.1 Early Approaches in Cross-Lingual Transfer (CLT)

Early research in CLT is typically divided into two key paradigms: *data-based* transfer and *model-based* transfer (Ruder et al., 2019).

Data-based methods transfer supervision across languages via explicit data transformations – most commonly *translation* or *annotation projection* – whereas model-based approaches rely on shared multilingual representations learned by mPLMs. We first discuss data-based transfer, which dominated CLT research before the widespread adoption of large-scale mPLMs.

2.1.1 Data-based Transfer

Data-based approaches enable cross-lingual transfer by *explicitly mapping annotations* from a high-resource language (HRL) to a target low-resource language (LRL). These methods can be further grouped based on the type of cross-lingual mapping they employ.

Translation-based Projection. Early work relied on parallel corpora and statistical word alignment to project labels across aligned word spans, enabling structured prediction without target-language annotations (Ehrmann et al., 2011; Ni et al., 2017; Agerri et al., 2018). With the advent of neural machine translation (NMT), later approaches replaced statistical translation pipelines with neural translators, often paired with alignment tools such as *FastAlign* (Dyer et al., 2013), improving annotation quality for sequence labeling tasks (Jain et al., 2019; Fei et al., 2020). Extending this to more complex structures, Kolluru et al. (2022) address multilingual Open Information Extraction (OpenIE) by jointly modeling consistent translation and cross-lingual label projection to transfer *subject-relation-object* spans across five diverse languages.

More recent work has further tightened the coupling between translation and labeling. Techniques such as span substitution and label projection aim to reduce label misalignment during translation (Chen et al., 2023), while transformer-based alignment models (e.g., *AWESOME*) leverage contextual multilingual representations to improve word- and span-level alignment (García-Ferrero et al., 2022). The *T-Projection* framework (García-Ferrero et al., 2023) exemplifies this direction by combining multilingual text-to-text models for candidate generation with translation probabilities from large NMT systems (e.g., M2M100, NLLB200) for candidate selection. More recently, CODEC (Le et al., 2024) further improves robustness by enforcing output-format constraints during translation via constrained decoding. Since these approaches synthesize labeled training data in the target language, they are commonly referred to as *translate-train* methods.

A complementary line of work revisits the relative strength of translate-train for sentence-level classification tasks. Artetxe et al. (2023) demonstrate that translate-train with strong modern NMT systems can match or even surpass mPLM-based zero-shot transfer on several classification benchmarks, challenging the prevailing assumption that model-based transfer has superseded translation-based approaches. Their analysis suggests that the quality of the underlying NMT system, rather than the transfer paradigm per se, is often the decisive variable – a finding that underscores the importance of carefully controlling NMT quality in translation-based comparisons.

A recent systematic study by Ebing and Glavaš (2025) provides an important empirical perspective on the relative merits of word-alignment-based projection versus tag-and-translate approaches such as CODEC for token classification tasks. Their findings reveal that word alignment quality is the decisive factor in projection-based methods: contextual alignment models (e.g., *AWESOME*) substantially outperform statistical aligners (e.g., *FastAlign*) when paired with high-quality NMT, narrowing – though not fully closing – the gap with constrained tag-and-translate methods like CODEC. Crucially, they find that tag-and-translate approaches retain an advantage for morphologically complex and typologically distant languages, where back-projection errors in alignment-based methods compound more severely. This finding motivates our focus in Chapter 3 on zero-shot transfer without any translation resources, bypassing these translation-induced brittleness issues entirely.

Translate-test Approaches. An alternative paradigm is *translate-test*, where target-language inputs are translated into an HRL at inference time, processed using a source-trained model, and its labels are back-projected to the target language (Bel et al., 2003; Shah et al., 2010). While attractive due to its ability to reuse strong monolingual HRL models without target-language supervision, translate-test is brittle for typologically distant languages. For token-level tasks such as NER and POS tagging, this pipeline requires three sequential steps: (i) translating the target-language query into an HRL (e.g., English), (ii) obtaining token-level predictions on the translated text, and (iii) back-projecting these labels to the original target-language tokens via word alignment. Each step introduces independent error sources – translation errors obscure task-relevant cues, while back-projection relies on word alignment tools between English and low-resource languages that are notoriously imprecise and brittle (Dyer et al., 2013; Ebing and Glavaš, 2025), particularly for morphologically rich or free-word-order languages where one-to-one token correspondences rarely hold. Moreover, translate-test incurs additional inference latency and depends on external translation systems, limiting its practicality in low-resource settings.

A systematic investigation by Ebing and Glavaš (2024) further clarifies the conditions under which translation-based paradigms – both translate-train and translate-test – are preferable to direct zero-shot model-based transfer for low-resource languages. Their study finds that translate-train consistently outperforms translate-test, particularly for structured prediction tasks, owing

to the compounding errors introduced by inference-time back-projection. Crucially, neither paradigm dominates across all settings: translation-based transfer tends to be more competitive when target-language pretraining data is extremely scarce, whereas model-based zero-shot transfer is preferable when reasonable multilingual coverage exists. These findings directly motivate this thesis's focus on hybrid approaches that can adapt to varying resource levels rather than committing to a single transfer paradigm.

Dictionary- and Transliteration-based Projection. To reduce reliance on full-fledged translation systems, some work explores more lightweight lexical mapping strategies. Guo and Roth (2021) propose bilingual-dictionary-based word-by-word translation followed by constrained reordering using pretrained language models. Related approaches extend this idea beyond word-level mappings by maximizing lexical overlap at the subword level (§2.2.2.3) or normalizing surface forms across scripts using phonemic- or transliteration-based representations. For instance, Sohn et al. (2024) encode entity names using the *International Phonetic Alphabet (IPA)* (e.g., "China" is transcribed as /ʧaɪ.nə/ in English and /ʧiː.nə/ in Sinhala), improving cross-lingual NER by exploiting phonetic similarity across typologically diverse languages.

Overall, while data-based methods provide effective cross-lingual supervision, they depend heavily on external linguistic resources – parallel corpora, translation models, dictionaries, or transliteration models – which are often unavailable or unreliable for low-resource languages.

2.1.2 Model-based Transfer

Model-based approaches aim to internalize *cross-lingual alignment* within the model itself, without explicit annotation projection. These methods leverage mPLMs, such as mBERT and XLM-R, which learn shared representations across languages using joint subword vocabularies and self-supervised objectives.

In this paradigm, a model fine-tuned on HRL task data often transfers effectively to unseen target languages in a *zero-shot* setting. However, zero-shot performance varies substantially across tasks and language families, particularly for structured prediction problems.

To improve cross-lingual generalization, several alignment-focused extensions have been proposed. Teacher-student frameworks (Wu et al., 2020b) distill task knowledge from source-trained teachers into target-language students using unlabeled data. Other work exploits pseudo-parallel corpora or multilingual alignment objectives to improve representational coherence across languages (Ouyang et al., 2021). Self-training and consistency regularization methods (Bari et al., 2021) further refine target-language representations through iterative pseudo-labeling.

Beyond standard fine-tuning, recent research explores continual pretraining and modular adaptation mechanisms that enhance multilingual representations while limiting parameter up-

dates. We defer a detailed discussion of these techniques to §2.2.2, where they are examined in the context of modular and efficient CLT.

More recent work, such as García-Ferrero (2025), explores grammar-constrained decoding for text-to-text models, reflecting a broader trend toward incorporating explicit structural constraints into model-based cross-lingual transfer.

2.2 Model Adaptation for Cross-Lingual Transfer

2.2.1 Supervised Fine-Tuning (SFT)

Building on the foundation of shared multilingual representations established by mPLMs (Devlin et al., 2019; Conneau et al., 2020), *supervised fine-tuning* (SFT) has become the standard adaptation paradigm in cross-lingual NLP (Pires et al., 2019; Wu and Dredze, 2019). SFT typically involves updating all model parameters on HRL task data before *zero-shot* application to target languages.

Modern SFT is often studied through the lens of *multi-source supervision*, where labeled data from multiple HRLs are jointly used to encourage the learning of language-agnostic task representations. Early work primarily focused on fine-tuning with a single source language (Pires et al., 2019), while subsequent studies showed that multi-source fine-tuning leads to more robust generalization, particularly for low-resource and typologically diverse target languages (Ogueji et al., 2021; Alabi et al., 2022).

Within this paradigm, several training strategies have been explored. Some approaches adopt *sequential* curriculum-style fine-tuning over source languages (Badola et al., 2023), while others perform joint multilingual fine-tuning using mixed-language batching (Alabi et al., 2022). More recently, *instruction tuning* has been proposed as an alternative adaptation paradigm, where models are finetuned using natural-language task descriptions – typically in English – to improve cross-lingual generalization (Shaham et al., 2024; Wu et al., 2025). However, as discussed in Chapter 1, these gains are more prominent for sentence-level classification, with more limited effectiveness for the structured prediction tasks that we study in this thesis.

Despite its prevalence, SFT faces several challenges. Updating all model parameters often leads to *catastrophic forgetting* of the underlying multilingual knowledge, particularly when dominated by specific HRLs (Badola et al., 2023; Khelli et al., 2025). Moreover, SFT lacks modularity and scalability: each task–language pair configuration typically requires maintaining a separate model checkpoint, resulting in substantial computational and storage overhead in practical deployment scenarios. These limitations provide the primary motivation for the *parameter-efficient fine-tuning* (PEFT) frameworks discussed in the following sections.

2.2.2 Parameter-Efficient Fine-Tuning (PEFT) Methods

To address the scalability and modularity limitations of SFT methods, a broad class of approaches has emerged under the PEFT paradigm. These methods adapt mPLMs by training only a small subset of parameters, enabling efficient training and reusable components across languages and tasks.

2.2.2.1 Adapter-based Methods

Adapters (Houlsby et al., 2019) comprise compact neural modules inserted into the Transformer layers to enable parameter-efficient adaptation. Instead of updating the full model parameters Θ, adapters introduce a small set of task- or language-specific parameters Φ while keeping the pretrained backbone frozen.

Formally, let $\mathscr{M}_\Theta$ denote an N-layer multilingual PLM. At layer ℓ, given the hidden state $h^{(\ell)} \in \mathbb{R}^d$, the adapter applies a bottleneck transformation followed by a residual connection:

$$\mathbf{o}^{(\ell)} = \mathbf{h}^{(\ell)} + f_{\text{adapter}}(\mathbf{h}^{(\ell)}), \tag{2.1}$$

where

$$f_{\text{adapter}}(\mathbf{x}) = \mathbf{W}_{\text{up}}^{(\ell)}\, \sigma(\mathbf{W}_{\text{down}}^{(\ell)}\mathbf{x}) \tag{2.2}$$

with $\mathbf{W}_{\text{down}}^{(\ell)} \in \mathbb{R}^{m\times d}$, $\mathbf{W}_{\text{up}}^{(\ell)} \in \mathbb{R}^{d\times m}$, and bottleneck dimension $m \ll d$. This design keeps the number of trainable parameters small (typically under 1% of the full model), reducing overfitting while preserving the general knowledge encoded in the pretrained backbone.

The MAD-X Framework. A widely adopted adapter-based CLT framework is *MAD-X* (Pfeiffer et al., 2020), which explicitly separates adaptation into two modular components:

1. **Language Adapters (LAs):** For each language $\mathscr{L}_s$, a language adapter $\text{LA}_{\mathscr{L}_s}$ is trained on unlabeled text $\mathscr{U}_s$ using masked language modeling:

   $$\phi_{\mathscr{L}_s} \leftarrow \arg\min_{\phi} \mathscr{L}_{\text{MLM}}(\mathscr{M}_\Theta \circ \text{LA}_{\mathscr{L}_s}; \mathscr{U}_s),$$

   with Θ frozen. Each language adapter captures language-specific lexical, syntactic, and morphological patterns.

2. **Task Adapters (TAs):** Given labeled task data $\mathscr{D}_s$ for task $\mathscr{T}$, a task adapter $\text{TA}_{\mathscr{T}}$ is trained on top of the base model and the corresponding language adapter:

   $$\psi_{\mathscr{T}} \leftarrow \arg\min_{\psi} \mathscr{L}_{\mathscr{T}}(\mathscr{M}_\Theta \circ \text{LA}_{\mathscr{L}_s} \circ \text{TA}_{\mathscr{T}}; \mathscr{D}_s),$$

   while keeping both Θ and $\phi_{\mathscr{L}_s}$ frozen.

This separation enables a *plug-and-play* paradigm: at inference time, the task adapter can be reused while swapping the language adapter for a new language.

MAD-X performs well when source and target languages share scripts or vocabulary overlap. MAD-X 2.0 (Pfeiffer et al., 2021b) improves robustness to unseen scripts by extending adaptation to the embedding layer, factorizing embeddings into shared multilingual components and lightweight script-specific parameters.

Despite these advances, adapter-based CLT still faces two key limitations:

1. **Single-source dependence:** Transfer is typically mediated through one source language, making performance highly sensitive to lexical overlap or typological similarity between source and target languages.

2. **Reliance on target-language data:** Adapting to a new language usually requires unlabeled target-language text to train additional adapters, preventing deployment in strict zero-resource (0-CLT) settings.

These limitations motivate the need for *multi-source* and *strict 0-CLT* transfer mechanisms that can integrate knowledge from multiple languages without relying on target-language supervision.

2.2.2.2 Hypernetwork-based Parameter Generation

A complementary line of work addresses multi-source adaptation through *hypernetworks* (Ha et al., 2016), which generate adapter parameters ϕ conditioned on auxiliary inputs. In the cross-lingual setting, *Conditional Parameter Generation (CPG)* (Üstün et al., 2020) represents a canonical instantiation of this idea.

Formally, CPG employs a hypernetwork H that maps a language embedding $v_{\mathscr{L}}$ – encoding typological or linguistic features of a language $\mathscr{L}$ (Malaviya et al., 2017) – to synthesize its adapter parameters:

$$\phi_{\mathscr{L}} = H_{\Omega}(v_{\mathscr{L}}),$$

where Ω represents the learnable parameters of the hypernetwork. This formulation enables the model to generate adapter parameters for an unseen target language based on its typological proximity to source languages, facilitating limited zero-shot generalization.

Language embeddings $v_{\mathscr{L}}$ typically encode typological properties such as word order (e.g., SVO vs. SOV), morphological complexity, or phonological traits. A detailed discussion of typological representations is provided in §3.2 (Chapter 3).

A prominent extension is MAD-G (Ansell et al., 2021), which applies the same hypernetwork formulation to *continual pretraining*, synthesizing language adapters across up to 95 languages without storing separate per-language parameters. Unlike CPG, which targets task-specific multi-source transfer, MAD-G focuses on language-level adaptation, demonstrating that a single hypernetwork can scale across diverse typological families. We discuss the implications of this design choice relative to ZGUL in the Limitations paragraph below.

Limitations and Positioning. Despite its effectiveness in multi-source CLT, CPG presents several practical limitations that motivate alternative design choices. First, adapter parameters are synthesized through a shared hypernetwork, entangling language-specific knowledge within a single parameter space and limiting modular reuse. For example, a hypernetwork trained on English, Hindi, and German cannot selectively attenuate German's contribution when transferring to a typologically distant language such as Bhojpuri. Second, jointly training a high-capacity hypernetwork across multiple source languages is computationally expensive and requires substantial task-specific multilingual supervision. Third, CPG provides no explicit mechanism to quantify or reweight individual source-language contributions at inference time, unlike ensemble-based approaches that maintain per-language modules with interpretable ensemble weights. Collectively, these limitations motivate the design of ZGUL (Chapter 3), which preserves independently trained language adapters and enables dynamic, target-aware, and interpretable ensembling at test time.

Prompt-, Prefix-, and Low-rank Tuning Methods. Beyond adapters and hypernetworks, several PEFT approaches adapt mPLMs by optimizing small continuous modules while keeping the backbone frozen. *Prompt-tuning* (Lester et al., 2021) prepends learnable, task-specific vectors at the input embedding layer, whereas *prefix-tuning* (Li and Liang, 2021) injects these trainable vectors into the intermediate activations across transformer layers, offering greater expressiveness. Both approaches update only a tiny fraction of parameters, enabling highly efficient adaptation.

Low-rank adaptation (LoRA) (Hu et al., 2022) adds low-rank decomposition matrices into the weights of transformer layers and has been adapted for multilingual and cross-lingual applications. These include summarization tasks (Whitehouse et al., 2024), natural language understanding (NLU) tasks via fused LoRA modules (Borchert et al., 2025), and machine translation using language-specific or mixture-of-expert LoRA configurations (Zhao et al., 2025).

While these methods are highly parameter-efficient, they are naturally less suited to structured cross-lingual transfer. In particular, they do not explicitly separate language-specific and task-specific knowledge, nor do they provide mechanisms for controlled multi-source fusion or interpretability of weights. As a result, they offer limited support for the modular and multi-source CLT settings studied in this thesis.

2.2.2.3 Lexicon- and Embedding-based Adaptation

Another complementary set of approaches focuses on adapting an mPLM's embeddings and subword vocabularies to better align with a new language. We categorize these methods as follows:

Continual Pretraining for Language Adaptation. A complementary alternative to supervised fine-tuning (SFT) is *continual pretraining* (CPT), where an mPLM is further trained on

unlabeled monolingual corpora to align its representations with a target language. This is typically followed by SFT on HRL task data. For example, Nag et al. (2025) demonstrate that continued masked language modeling (MLM) on target-language corpora substantially boosts performance on downstream tasks such as generation (Singh et al., 2024), even without labeled data in the target language. Similarly, language- or domain-balanced pretraining objectives mitigate catastrophic forgetting and improve coverage for low-resource languages (Khanuja et al., 2021; Xue et al., 2021).

Many LRLs exhibit rich morphology with complex inflectional and derivational patterns. When processed using tokenizers optimized for HRLs, these languages frequently suffer from *excessive subword fragmentation*, leading to inflated sequence lengths and weakened lexical alignment (Gerz et al., 2018; Park et al., 2021). This degradation is particularly harmful for cross-lingual transfer in structured prediction tasks that rely on precise token-level alignment. To mitigate these issues, prior research has focused on incorporating morphological and subword structure directly into the representation learning process:

Morphological and Subword Adaptation.

- **Morphological Segmentation**: Using tools like *Morfessor* (Kunchukuttan et al., 2020) to derive morpheme-aware units captures inflectional regularities more effectively than standard subword algorithms. Similarly, Nzeyimana and Niyongabo Rubungo (2022) demonstrate that morphology-aware segmentation substantially improves token representations for Bantu languages.

- **Vocabulary Alignment**: Other approaches improve cross-lingual alignment by directly adapting subword vocabularies. Early methods expand or interpolate vocabularies to increase multilingual coverage without retraining large PLMs (Artetxe et al., 2020; Rust et al., 2021). Patil et al. (2022) maximize subword sharing across typologically related languages, yielding stronger transfer for Indic languages. Unlike dictionary-based projection, these approaches operate purely at the subword level and require no parallel data.

- **Byte-Level Modeling**: Architectures such as ByT5 and CANINE eliminate fixed vocabularies entirely by operating directly on raw bytes or characters (Xue et al., 2022; Clark et al., 2022). While highly robust to script variation, their computational cost and efficacy for structured prediction under 0-CLT remain open challenges.

Embedding-Level Adaptation. Complementary to tokenization-aware methods, embedding-level adaptation explicitly modifies the input embedding space to address script and typological mismatches:

- **Matrix Factorization**: Pfeiffer et al. (2021b) factorize the pretrained embedding matrix into low-dimensional latent components that capture transferable multilingual structure, enabling lightweight adaptation to unseen scripts.
- **Relational Alignment**: Frameworks like *RelateLM* (Khemchandani et al., 2021) exploit language relatedness by aligning target-language embeddings with those of typologically related HRLs via bilingual dictionaries and continued MLM.

Together, morphology-aware segmentation, vocabulary adaptation, and embedding alignment form a robust toolkit for improving lexical coverage in mPLMs. Because these methods operate primarily at the input layer, they are fully compatible with PEFT techniques, supporting modular and scalable cross-lingual transfer.

2.3 Cross-lingual Prompting with Large Language Models (LLMs)

Recent multilingual large language models (LLMs), such as GPT-4, Llama-3.1, and Qwen 3, have substantially improved zero- and few-shot transfer across languages (Zhu et al., 2024). While these models exhibit emergent abilities to perform tasks in unseen languages without task-specific supervision, their performance remains uneven, often degrading sharply for typologically distant or low-resource languages (Ahuja et al., 2023; Vatsal et al., 2025).

Most existing cross-lingual prompting paradigms rely on transferring supervision from high-resource languages, typically English. Common strategies include zero-shot prompting with translated instructions, few-shot prompting with English exemplars, or mixed-language prompting (Winata et al., 2021; Lin et al., 2022; Qin et al., 2023). While translation-based approaches partially bridge linguistic gaps, they frequently introduce semantic drift and artifacts that degrade performance, particularly in structured prediction tasks (Ahuja et al., 2023; Kim et al., 2024).

A recurring limitation across these paradigms is that exemplar selection is largely *static* and *task-agnostic*. Exemplars are typically manually curated or heuristically sampled, lacking mechanisms to ensure alignment with task structure, label distribution, or model uncertainty (Winata et al., 2021; Asai et al., 2024a). This shortcoming is especially detrimental to structured prediction, where precise task grounding is critical (Blevins et al., 2023).

Exemplar retrieval strategies. To improve prompting reliability, recent work has explored automated retrieval methods for selecting in-context exemplars.

- **Similarity-based retrieval:** A standard approach retrieves exemplars that are semantically similar to the input query. Because sparse lexical methods (e.g., TF-IDF) fail across languages due to script and morphological differences (Litschko et al., 2022), retrieval is typically

performed using dense multilingual encoders such as LaBSE (Feng et al., 2020), mSimCSE (Huang et al., 2023), or BGE-M3 (Chen et al., 2024a). Retrieved demonstrations are then ordered by relevance within the prompt to maximize priming effects (Liu et al., 2024).

- **Diversity-aware retrieval:** While similarity improves relevance, it can lead to redundant or narrow exemplar sets. To mitigate this, diversity-aware strategies aim to cover multiple semantic regions or label types. Approaches include Maximal Marginal Relevance (MMR), coverage-based subset selection, Determinantal Point Processes (DPPs), and reinforcement-learning-based selection policies (Kapuriya et al., 2025; Gupta et al., 2023; Wang et al., 2024a, 2025).

However, these methods are largely evaluated in English-only settings and assume access to gold-labeled exemplar pools.

Limitations of Current Cross-Lingual Prompting. Despite advances in retrieval, cross-lingual prompting pipelines face persistent challenges:

- **Source-Language Bias:** Reliance on English exemplars fails to capture target-language-specific linguistic phenomena.
- **Static Retrieval:** Retrieval optimizes semantic similarity without explicitly modeling task structure, prediction confidence, or label distribution.
- **LRL Instability:** LLMs remain biased toward high-resource languages, and translation noise compounds retrieval errors in low-resource settings (Ahuja et al., 2023).

These limitations underscore the need for prompting frameworks capable of constructing task-aware, linguistically aligned exemplar sets, even when gold target-language annotations are unavailable.

2.4 Hybrid Models: Combining Task-specific Models with LLM Reasoning

Hybrid approaches integrate the precision and efficiency of supervised, task-specific models with the contextual flexibility of LLMs, offering a robust alternative to purely self-adaptive prompting. In the domain of question answering and dialogue, FuSIC-KBQA (Patidar et al., 2024) exemplifies this paradigm through a *retrieve–then–generate* pipeline for few-shot KBQA, where symbolic retrievers provide grounded candidate interpretations that an LLM subsequently refines. SyncTOD (Saley et al., 2024) follows a similar pattern in task-oriented dialogue: light-weight supervised modules generate task hints and exemplar demonstrations, which the LLM then uses to improve goal adherence, consistency, and brevity under low-data constraints. These

systems demonstrate that providing structured guidance upstream can substantially enhance LLM sample efficiency and reliability.

Beyond task-specific designs, recent frameworks incorporate reflection, retrieval, or tool invocation to guide LLM reasoning. For instance, *Reflexion* (Shinn et al., 2023) introduces iterative self-reflection loops for error correction; *Self-RAG* (Asai et al., 2024b) integrates retrieval into the decoding process with self-evaluation to ensure factual grounding; *SPARKLE* (Yu et al., 2024) employs tool-augmented reasoning to improve precision in multi-hop QA; and *Toolformer* (Schick et al., 2023) autonomously learns when to call external APIs during inference. Together, these approaches highlight how external guidance from symbolic sources or tools improves control, factuality, and robustness.

Another crucial dimension of hybrid reasoning involves ensembling or dynamically routing between heterogeneous language models to balance capability and computational cost. Frameworks such as *AutoMix* (Aggarwal et al., 2024) optimize this trade-off by strategically routing queries to a larger, more capable LLM only when a smaller model's few-shot self-verification mechanism determines its own output is unreliable. Similar cascading and blending architectures, such as FrugalGPT (Chen et al., 2024b) or LLM-Blender (Jiang et al., 2023b), integrate the strengths of multiple disparate models to achieve superior reasoning performance compared to a single monolithic system.

While these *Agentic Workflows* have matured in monolingual settings, their cross-lingual applications remain limited. Current multilingual systems typically rely on pipelines combining language identification, machine translation, and multilingual retrieval for cross-lingual QA (Palma Gomez et al., 2024), or multi-agent setups for code translation (Yang et al., 2025b). However, there is little exploration of hybrid reasoning explicitly tailored for cross-lingual transfer in structured prediction tasks such as NER, POS tagging, or NLI.

This thesis addresses these limitations through two contributions. *SSP* (Chapter 4) introduces a hybrid prompting framework for 0-CLT that leverages pseudo-labels from supervised CLT models to construct task-aware target-language exemplars via structured optimization for prompting the LLM, removing the need for target-language annotations. *HYDRE* (Chapter 6) extends hybrid reasoning to cross-lingual DSRE, integrating candidate filtering from a supervised DSRE model with LLM-based relation disambiguation under distant supervision.

2.5 Evaluation Tasks and Metrics

To evaluate cross-lingual transfer in structured prediction settings, we focus on two complementary task families: (i) *sequence labeling*, which probes token-level morpho-syntactic and semantic modeling, and (ii) *Distantly Supervised Relation Extraction (DSRE)*, which evaluates semantic relational reasoning under noisy supervision.

Sequence Labeling. This task requires assigning a categorical label y_i to each token x_i in a sentence $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. We use two canonical tasks that test different aspects of linguistic transfer.

- **Part-of-Speech (POS) Tagging:** POS tagging assigns each token a morpho-syntactic category (e.g., noun, verb, adposition). It serves as a fine-grained diagnostic of a model's understanding of local syntax and grammatical structure. Since every token must be labeled, POS tagging requires dense token-level classification grounded in both morphology and immediate grammatical context (Nivre et al., 2020). The label set for POS comprises 17 tags: `ADJ` (adjective), `ADP` (adposition), `ADV` (adverb), `AUX` (auxiliary), `CCONJ` (coordinating conjunction), `DET` (determiner), `INTJ` (interjection), `NOUN` (noun), `NUM` (numeral), `PART` (particle), `PRON` (pronoun), `PROPN` (proper noun), `PUNCT` (punctuation), `SCONJ` (subordinating conjunction), `SYM` (symbol), `VERB` (verb), and `X` (other).

- **Named Entity Recognition (NER):** NER detects and categorizes textual spans into predefined semantic types such as `Person`, `Organization`, or `Location`. Unlike POS tagging, NER is a sparse labeling task targeting only meaningful entity spans rather than each token. It involves two sub-problems: (i) *boundary detection*, which determines the span of the entity, and (ii) *type classification*. Compared to POS tagging, NER relies more on lexical cues and broader contextual signals than strict morpho-syntactic rules (Pan et al., 2017). The label set for NER follows the BIO scheme over four entity types: `PER` (person), `LOC` (location), `ORG` (organization), and `DATE` (date), along with `O` (outside any entity).

Table 2.1 illustrates both tasks using a Bhojpuri sentence, with transliteration and word-level glosses provided to aid non-native readers.

Distantly Supervised Relation Extraction (DSRE). Relation Extraction (RE) extends understanding from token-level semantics to *relational reasoning* between entity pairs (Zelenko et al., 2003; Zhou et al., 2005). Given a sentence with an entity pair (e_1, e_2), the goal is to predict their semantic relation r from an ontology $\mathcal{R} \cup \{\texttt{NA}\}$, where `NA` indicates that none of the relations in $\mathcal{R}$ are expressed.

To reduce the cost of manual annotation – especially in low-resource settings – this thesis focuses on distantly supervised relation extraction, where supervision is derived automatically from external knowledge bases (KBs).

Data Curation and the MIML Paradigm. DSRE datasets are constructed by aligning text with KBs, such as Freebase, to generate training data at scale (Mintz et al., 2009; Riedel et al., 2010). This process follows the Multi-Instance Multi-Label (MIML) paradigm (Hoffmann et al., 2011):

- **Bag Construction:** All sentences mentioning an entity pair (e_1, e_2) are grouped into a *bag* $B = \{s_1, s_2, \ldots, s_m\}$.

Table 2.1: **Sequence labeling example** for a sentence in Bhojpuri. English Translation - "Former U.S. President Joe Biden has come on a visit to India ."

Token	Transliteration	English Meaning	POS	NER
अमेरिका	*Amerika*	America	PROPN	B-LOC
के	*ke*	of	ADP	O
राष्ट्रपति	*rashtrapati*	President	NOUN	O
रहल	*rahal*	was	AUX	O
जो	*Joe*	Joe	PROPN	B-PER
बाइडेन	*Biden*	Biden	PROPN	I-PER
भारत	*Bharat*	India	PROPN	B-LOC
दौरा	*daura*	visit / tour	NOUN	O
पर	*par*	on	ADP	O
आइल	*aail*	came / come	VERB	O
बाड़न	*baadan*	has (honorific)	AUX	O
।	.	.	PUNCT	O

Example of a DSRE bag:

Entity Pair: Head: **Barack Obama** (Q76), Tail: **USA** (Q30)
KB Relations: `President_of, Born_in`
Bag:
s_1 : "**Obama** was born in Hawaii, **USA**." (expresses `Born_in`)
s_2 : "**Obama** served as the 44th president of the **USA**." (expresses `President_of`)
s_3 : "**Obama** returned to the **USA** after his 3-day European trip." (**Noise**: no relation expressed)

Figure 2.1: **DSRE bag illustration** for head entity *Barack Obama* and tail entity *USA* illustrating the presence of both informative and noisy sentences.

- **Distant Labeling:** If the KB asserts a relation r between (e_1, e_2), the entire bag is labeled with r.

This formulation introduces the *at-least-one* assumption: at least one sentence within the bag conveys the target relation (Surdeanu et al., 2012). In practice, however, many sentences merely mention the entities without expressing the relation, introducing substantial label noise (Xu et al., 2013). Consequently, DSRE models must perform *multi-instance reasoning*, identifying which sentences truly support each relation while ignoring spurious evidence.

Example of Bag-Level Noise. Figure 2.1 illustrates a typical DSRE bag for the entity pair (**Barack Obama**, **USA**). Sentences s_1 and s_2 express the relations `Born_in` and `President_of`, respectively, whereas s_3 contains no relation despite mentioning both entities. A robust DSRE model must extract evidence from s_1 and s_2 while suppressing the noise in s_3.

Challenges in DSRE and its cross-lingual implications. When DSRE models are transferred across languages, the intrinsic difficulties of distant supervision are further amplified by cross-lingual variation. We highlight four key challenges:

1. **Label Noise Amplification:** The label noise inherent in distant supervision is compounded by translation artifacts and entity alignment errors when transferring from HRLs to LRLs (Hennig et al., 2023).

2. **Bag-to-Sentence Mismatch:** DSRE models are trained using bag-level supervision in the source language but are often deployed for sentence-level inference in the target language. This mismatch becomes more severe when relational signals – such as word order, morphology, or argument structure – vary across languages, leading to degraded generalization (Gao et al., 2021).

3. **Multi-Label Learnability:** DSRE is inherently a *multi-label* learning problem, as a single entity pair can participate in multiple relations. For example, as shown in Figure 2.1, the entity pair (*Barack Obama*, *USA*) is associated with both `President_of` and `Born_in`. This substantially increases learning complexity: multiple relations often share overlapping lexical or semantic cues, making fine-grained distinctions difficult.

4. **Lack of Benchmarks:** Most DSRE benchmarks are English-centric. Existing multilingual datasets (Bhartiya et al., 2022) primarily cover high-resource European languages and offer limited insight into performance on typologically diverse low-resource languages. For Indic and other LRLs, reliable cross-lingual DSRE evaluation resources remain scarce.

To avoid redundancy, detailed dataset descriptions, preprocessing pipelines, and language-specific statistics are deferred to the corresponding experimental chapters. Instead, the next paragraph focuses exclusively on the core evaluation *settings* and *metrics* for POS tagging, NER, and DSRE that are applied consistently throughout the thesis.

Evaluation Settings and Metrics. To rigorously assess the adaptation mechanisms discussed throughout this thesis, we employ a standardized evaluation framework across all target tasks.

CLT Supervision Regimes. All models are evaluated under the three settings defined in §1.1 (Chapter 1): **0-CLT** (zero-resource), **T-CLT** (translation-based), and **U-CLT** (unlabeled-data transfer).

Classification Metrics. For a label set $\mathcal{Y}$ with positive classes $\mathcal{Y}^+$ (excluding background labels: `O` for NER, `X` for POS, and `NA` for DSRE), let TP_c, FP_c, and FN_c denote the true positives, false positives, and false negatives for class $c \in \mathcal{Y}^+$. We report two primary metrics:

- **Micro-F1** aggregates counts across all positive classes before computing precision (P_μ)

and recall (R_μ):

$$P_\mu = \frac{\sum_{c \in \mathcal{Y}^+} \mathrm{TP}_c}{\sum_{c \in \mathcal{Y}^+} (\mathrm{TP}_c + \mathrm{FP}_c)}, \quad R_\mu = \frac{\sum_{c \in \mathcal{Y}^+} \mathrm{TP}_c}{\sum_{c \in \mathcal{Y}^+} (\mathrm{TP}_c + \mathrm{FN}_c)}, \quad \text{Micro-F1} = \frac{2 P_\mu R_\mu}{P_\mu + R_\mu}.$$

- **Macro-F1** computes per-class F1 scores and averages them over the $|\mathcal{Y}^+|$ classes:

$$\mathrm{F1}_c = \frac{2 P_c R_c}{P_c + R_c}, \qquad \text{Macro-F1} = \frac{1}{|\mathcal{Y}^+|} \sum_{c \in \mathcal{Y}^+} \mathrm{F1}_c.$$

Task-Specific Evaluation Protocols. For sequence labeling, NER evaluation is performed at the *entity-span level* (requiring exact boundary and type match), while POS is evaluated at the *word level*. All sequence labeling scores are computed using the `seqeval`[1] toolkit and averaged over three random seeds.

For DSRE models that produce calibrated confidence scores, we additionally report two ranking-based metrics:

- **Area Under the Curve (AUC):** All predictions across entity pairs are ranked by decreasing model confidence. The prediction threshold is varied, and precision and recall are computed at each threshold, yielding a Precision-Recall (PR) curve. AUC is the area under this curve, summarizing a model's ability to rank correct predictions above incorrect ones across varying confidence thresholds.

- **Precision@M (P@M):** Precision evaluated at rank M. This measures the fraction of correct predictions among the model's top-M most confident outputs.

LLM-based methods (e.g., GPT-4), which do not yield calibrated probabilities via standard APIs, are evaluated using F1 metrics exclusively. Finally, where applicable throughout the thesis, the statistical significance of performance gains is measured using McNemar's test (McNemar, 1947).

[1] https://github.com/chakki-works/seqeval

Chapter 3

Cross-lingual Generalization using Multi-source Language Adapter Ensembling

3.1 Introduction

Earlier chapters established that robust cross-lingual transfer requires *modular representations* to disentangle language-specific structure from task semantics (§2.2.2, Chapter 2). This property becomes crucial under strict zero-resource (0-CLT) conditions where target-language data is entirely absent (§1.3, Chapter 1). This chapter builds upon that foundation to address a critical remaining gap: how can we perform effective transfer when the target language is entirely unseen, precluding the training of a language-specific adapter?

Existing adapter-based frameworks, such as MAD-X (Pfeiffer et al., 2020), fundamentally assume the availability of a target-language adapter, which inherently requires unlabeled target-language corpora. Under strict 0-CLT constraints, this assumption breaks down. Instead, the model must infer how to represent an unseen language using only the knowledge already encoded within existing source-language components.

To address this challenge, we introduce ZGUL *(Zero-resource Generalization to Unseen Languages)*. ZGUL reframes cross-lingual transfer as a *dynamic routing problem*. Rather than relying on a single source language as a rigid proxy for the target, ZGUL synthesizes a target representation by combining multiple source-language adapters through a learned routing mechanism. By treating the source-language adapter bank as a mixture of linguistic experts, ZGUL leverages typological similarity and contextual cues to dynamically fuse their representations. Consequently, this framework enables robust structured prediction in previously unseen languages without requiring any target-language data.

3.2 Preliminaries & Related Work

We briefly summarize the modeling components used by ZGUL. Readers are assumed to be familiar with multilingual transformers and adapter-based parameter-efficient fine-tuning, introduced in Chapter 2. This section focuses only on notation and concepts required to describe the proposed method.

- **Transformer for Token-Level Prediction:** Let $\mathscr{M}_\Theta$ denote a multilingual pretrained transformer with N layers. For an input sentence $x = (x_1, \ldots, x_L)$, the encoder produces contextual token representations $h_j^{(N)} = \mathscr{M}_\Theta(x)_j$. A linear classifier maps these representations to label probabilities

$$p(y_j \mid x; \Theta) = \text{softmax}(W_o h_j^{(N)} + b_o). \tag{3.1}$$

  Following standard practice (Hu et al., 2020a), training applies supervision only to the first subword token of each word, with token-level cross-entropy loss computed over these positions.

- **Adapter Bank as Linguistic Experts:** We assume access to a set of S pretrained source-language adapters

$$\mathscr{A} = \{\text{LA}_{\mathscr{L}_1}, \ldots, \text{LA}_{\mathscr{L}_S}\},$$

  each encoding structural properties of a source language. A single task adapter $\text{TA}_{\mathscr{T}}$ captures task-specific semantics.

  Under standard adapter transfer, inference in a target language requires a dedicated adapter $\text{LA}_{\mathscr{L}_t}$. ZGUL instead assumes that the target representation can be approximated as a weighted combination of source adapters:

$$\text{LA}_{\mathscr{L}_t}^{\text{approx}} \approx \sum_{s=1}^{S} \alpha_s \text{LA}_{\mathscr{L}_s}, \tag{3.2}$$

  where weights α_s depend on linguistic similarity and contextual evidence.

- **Typological Language Features:** Typological features encode the structural properties of a language, including its grammar, word order, and sound system. These features serve as a *linguistic blueprint*, enabling models to understand an unseen language by comparing it to known ones – an especially valuable signal in low-resource settings where target language data is scarce.

  We obtain typological representations from the URIEL database (Littell et al., 2017), which aggregates syntactic and phonological information for over 1,000 languages. To operationalize these features, we use the `lang2vec` toolkit, which maps each language to a dense vector representation. In this work, we use the 289-dimensional `lang2vec` vectors that combine syntactic and phonological features.

- **Measuring Linguistic Similarity:** We measure similarity between languages using cosine similarity between their typological vectors. This score serves as a structural prior for identifying source languages most relevant to a target language and is used to guide adapter weighting in our model. Typology-based similarity has previously been used for source selection, model selection, and adapter synthesis in cross-lingual transfer (Eronen et al., 2023; Chen and Ritter, 2021; Üstün et al., 2020).

Conditional Parameter Generation (CPG). We adopt CPG (Üstün et al., 2020) as a representative multi-source baseline (formulation and limitations discussed in §2.2.2.2, Chapter 2). The key contrast with ZGUL is that CPG *generates* new adapter parameters via a jointly trained hypernetwork, whereas ZGUL *routes* over independently pretrained adapters – preserving modularity and avoiding the computational overhead of joint multi-language training that CPG requires (e.g., 95 languages in Ansell et al., 2021). This distinction allows us to empirically evaluate whether modular adapter ensembling can match or surpass hypernetwork-based synthesis while retaining interpretability.

Inference-Time Entropy Minimization. Entropy Minimization (EM) is a self-supervised adaptation technique that refines a model's predictions on unlabeled target-language inputs at inference time by minimizing predictive uncertainty. Given a query x of length L, the objective minimizes per-token Shannon entropy (Shannon, 1948):

$$\text{Loss}_{\text{EM}}(x;\Theta) = \sum_{j=1}^{L} H(p(y_j \mid x;\Theta)), \tag{3.3}$$

where entropy over label distribution p across the label set $\mathcal{Y}$ is computed as:

$$H(p) = -\sum_{y\in\mathcal{Y}} p(y)\log p(y) \tag{3.4}$$

In the context of cross-lingual transfer, EM allows for the adaptive reallocation of probability mass toward the most predictive source adapters for a given target input. This methodology builds upon the single-source training framework of MAD-X (Wang et al., 2021), in which the English adapter employed during training is replaced at inference time with a uniform ensemble of related source adapters. These ensemble weights are then refined via EM to minimize predictive uncertainty on the target query.

We summarize the limitations of the approaches discussed so far in Table 3.1.

3.3 ZGUL: Zero-Shot Generalization via Multi-Source Adapter Ensembling

ZGUL addresses the strict 0-CLT setting by eliminating the need for a target-language adapter

Method	Modular	Multi-Source	0-CLT Compatible	Computationally Efficient
SFT	✗	✓	✓	✗
MAD-X	✓	✗	✗	✓
CPG	✗	✓	✓	✗
ZGUL (ours)	✓	✓	✓	✓

Table 3.1: **Comparison of CLT paradigms** across four dimensions. ZGUL is the only method satisfying all four: *modularity*, *multi-source support*, *strict 0-CLT compatibility*, and *computational efficiency*.

$\text{LA}_{\mathscr{L}_t}$. Instead, it constructs a target representation by dynamically combining a bank of pre-trained source adapters $\mathscr{A} = \{\text{LA}_{\mathscr{L}_1}, \ldots, \text{LA}_{\mathscr{L}_S}\}$.

Unlike hypernetwork approaches, such as Conditional Parameter Generation (§2.2.2.2), which synthesize a single static adapter, ZGUL performs *adaptive routing* over existing adapters. This routing operates at two complementary levels: (i) a *global* typology-aware mixture that reflects language-level similarity, and (ii) a *token-level* mixture that captures context-specific variation.

Global Typology-Aware Fusion. We first derive a static prior over source languages using typological vectors. Each language $\mathscr{L}$ is associated with a raw Lang2Vec feature vector $v_{\mathscr{L}}^{\text{raw}}$ (introduced in §3.2), which we transform via a learnable MLP projection $g_{\boldsymbol{\phi}}(\cdot)$, trained jointly with the rest of ZGUL, to obtain

$$v_{\mathscr{L}} = g_{\boldsymbol{\phi}}\left(v_{\mathscr{L}}^{\text{raw}}\right).$$

This learned projection – rather than the raw Lang2Vec features directly – is what feeds Global Fusion (Figure 3.1), allowing the model to adapt the typological representation space to the downstream task rather than relying on the fixed, hand-engineered Lang2Vec geometry. For a target language $\mathscr{L}_t$, global mixture weights are then computed as:

$$\alpha_s = \frac{\exp\left(\text{sim}(v_{\mathscr{L}_s}, v_{\mathscr{L}_t})\right)}{\sum_{s'=1}^{S} \exp\left(\text{sim}(v_{\mathscr{L}_{s'}}, v_{\mathscr{L}_t})\right)}.$$

These weights depend only on language identity and remain fixed across tokens within a sentence. They encode a structural prior over which source languages are most informative for the target. For example, if the target language is Faroese, these weights will consistently favor Germanic or Icelandic adapters because of their inherent syntactic and phonological proximity, regardless of the individual words in a sentence.

Let $o_{j,s}^{(\ell)}$ denote the output of the s-th source adapter for token x_j at layer ℓ. The global mixture becomes

$$\tilde{o}_{j,\text{global}}^{(\ell)} = \sum_{s=1}^{S} \alpha_s \, o_{j,s}^{(\ell)}.$$

Token-level (Local) Fusion. While global fusion captures typological relatedness, it cannot account for word-specific nuances such as loanwords or code-switching. To address this, ZGUL employs a *local fusion* mechanism based on *AdapterFusion* (Pfeiffer et al., 2021a).

For each token x_j, we treat the pre-adapter hidden state as a Query (**q**) and the source adapter outputs as Keys (**K**) and Values (**V**).

Let $q_j^{(\ell)} = W_q X_j^{(\ell)}$ be the projected query vector ($X_j^{(\ell)}$ = output of feed-forward module in layer ℓ), and let the projected adapter outputs be defined as:

$$K_j^{(\ell)} = \{W_k o_{j,s_i}^{(\ell)}\}_{s_i \in \mathcal{S}}, \quad V_j^{(\ell)} = \{W_v o_{j,s_i}^{(\ell)}\}_{s_i \in \mathcal{S}}$$

The local weights $\beta_{j,s_i}^{(\ell)}$ are computed using a query-key-value attention mechanism:

$$\beta_j^{(\ell)} = \text{softmax}\left(\frac{q_j^{(\ell)\top} K_j^{(\ell)}}{\sqrt{d}}\right)$$

The resulting local fusion output is

$$\tilde{o}_{j,\text{local}}^{(\ell)} = \sum_{i=1}^{S} \beta_{j,s_i}^{(\ell)} V_{j,s_i}^{(\ell)}.$$

Local fusion enables context-specific routing. For instance, in a Hausa sentence, a specific token might be an English loanword. While the global prior favors African language adapters, local fusion can dynamically "route" the query to the English adapter to capture the token's specific semantics.

Combined Representation and Task Prediction. ZGUL integrates both signals by concatenating the global and local mixtures and projecting them back to the model dimension:

$$z_j^{(\ell)} = \tilde{o}_{j,\text{global}}^{(\ell)} \oplus \tilde{o}_{j,\text{local}}^{(\ell)}, \quad \hat{h}_j^{(\ell)} = W_f z_j^{(\ell)} + b_f.$$

The resulting intermediate representation is then passed to the task adapter to obtain output $h_j^{(\ell)}$ for the layer l:

$$h_j^{(\ell)} = \text{TA}_{\mathcal{T}}^{(\ell)}(\hat{h}_j^{(\ell)}).$$

This ensures predictions are grounded simultaneously in (i) global linguistic similarity and (ii) token-level contextual evidence.

Inference-Time Adaptation via EM. At test time, ZGUL refines the learned fusion weights (α, β) using the Entropy Minimization objective introduced in §3.2, Eq. 3.3. Applied to ZGUL's dual-level fusion, this becomes:

$$\text{Loss}_{\text{EM}}(x) = \sum_{j=1}^{L} H(p_\Theta(y_j \mid x; \alpha, \beta)).$$

In the zero-shot setting, the initial ensemble weights may be poorly calibrated for an unseen target language. By minimizing entropy, ZGUL reallocates weight toward the combination of

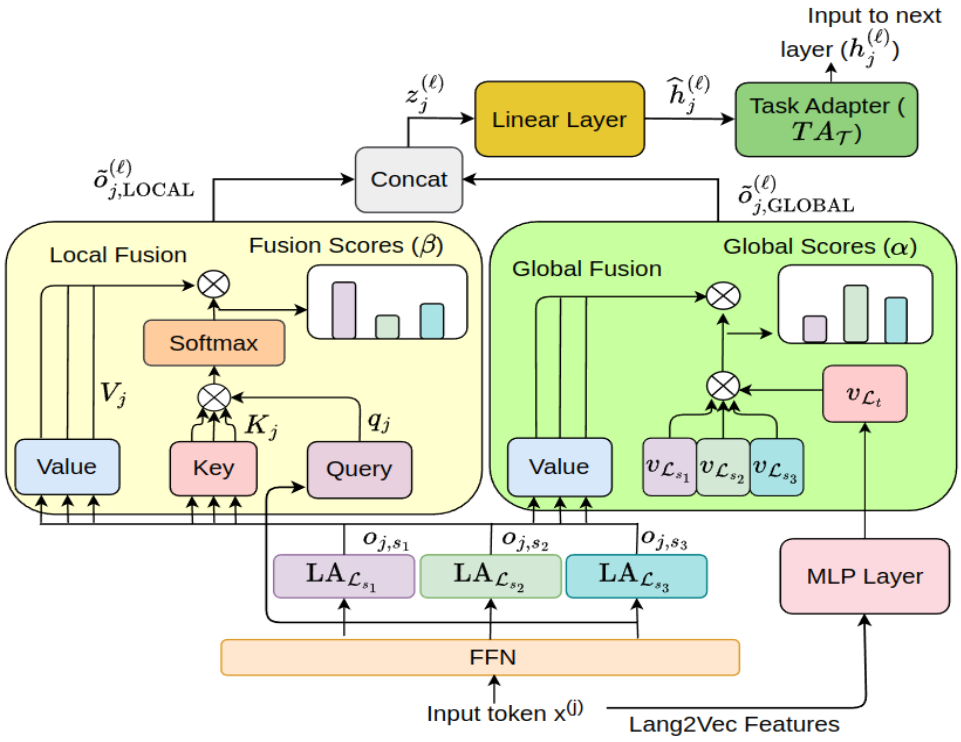


Figure 3.1: **Dual-level fusion in ZGUL.** *Local* (token-level, left) and *Global* (language-vector, right) fusion outputs are combined and processed by the Task Adapter to produce the final representation at layer ℓ.

source adapters that yields the most confident predictions, allowing the model to self-correct its routing strategy based on local (contextual) and global (linguistic) patterns observed at test time – without requiring any labeled data.

During this process, only the fusion parameters (α, β) are updated via gradient descent, while the base mPLM, all source language adapters, and the task adapter remain frozen. This ensures efficient and stable adaptation without altering the underlying representations. The optimization procedure is detailed in Algorithm 2, and hyperparameters (the number of gradient steps T and learning rate η) are reported in §A.1.4.

Connection to Mixture-of-Experts (MoE). The adapter ensembling framework in ZGUL can be naturally interpreted through the lens of *Mixture-of-Experts* (MoE) (Jacobs et al., 1991; Shazeer et al., 2017): each source-language adapter functions as an *expert*, while the dual-level fusion mechanism implements *two parallel MoE routing strategies* that are combined at each layer. Specifically, *Local Fusion* implements a standard *input-conditioned* MoE gate – fusion scores β are computed as a function of the input query token $x^{(j)}$, analogous to how MoE routers select experts based on the current input. *Global Fusion* implements a *language-conditioned* MoE gate – fusion scores α are computed from the query's language metadata via Lang2Vec typological vectors $v_{\mathcal{L}_t}$, disentangling language-level routing from token-level routing. The two gating mechanisms are complementary: Local Fusion captures fine-grained, token-level cross-lingual cues, while Global Fusion captures coarse-grained, typology-driven source selec-

Lang. Group	Train set*	Test set	Task
Germanic	German (De), Icelandic (Is)	Faroese (Fo), Gothic (Got), Swiss German (Gsw)	POS
Slavic	Russian (Ru), Czech (Cs)	Pomak (Qpm), Upper Sorbian (Hsb), Old East Slavic (Orv), Old Church Slavonic (Cu)	POS
African	Amharic (Amh), Swahili (Swa), Wolof (Wol)	Hausa (Hau), Igbo (Ibo), Luo (Luo), Luganda (Lug), Kinyarwanda (Kin)	NER
Indo-Aryan	Hindi (Hi), Bengali (Bn), Urdu (Ur)	Assamese (As), Bhojpuri (Bh)	NER

Table 3.2: **Language groups, source–target splits, and tasks.** *English is included as a common source in each language group.

tion. Their outputs are concatenated and projected at each layer, effectively combining two MoE representations per layer rather than selecting a single routing strategy. We view this dual-gate design as a form of *dual Mixture-of-Experts*, disentangling input-level and language-level routing into separate, interpretable gating mechanisms.

Prior work has explored MoE routing at the task or language-pair level within multilingual neural machine translation (Kudugunta et al., 2021; Nath and Gulzar, 2026), but not in the context of adapter-based or LLM-driven cross-lingual transfer; extending this dual-MoE formulation to LLM-based cross-lingual transfer remains an open direction, which we leave to future work.

3.4 Experimental Setup

This section describes the tasks, datasets, and evaluation protocols used to assess the effectiveness of ZGUL under cross-lingual transfer.

Tasks, Datasets, and Languages. We evaluate ZGUL on token-level sequence labeling across four diverse language families. We use Universal Dependencies (UD) (Nivre et al., 2020) for Part-of-Speech (POS) tagging in Germanic and Slavic languages. For Named Entity Recognition (NER), we use MasakhaNER (Adelani et al., 2021) for African languages and WikiAnn (Pan et al., 2017) for Indo-Aryan languages. We focus on these fine-grained, token-level tasks following earlier work on related PEFT approaches for multi-source fine-tuning (Üstün et al., 2020; Ansell et al., 2022, 2021) and language-adapter ensembling (Wang et al., 2021). We do not additionally evaluate sentence-classification tasks such as sentiment analysis here: while we consider NLI for Indigenous languages of the Americas in our LLM-based (SSP) evaluation in

Chapter 4, that benchmark includes only two high-resource languages (English and Spanish), fewer than the three required to study ZGUL's multi-source transfer setting. The evaluation spans 15 unseen target languages across the four language families, with detailed source–target splits shown in Table 3.2.

Source languages are selected exclusively from those included in mBERT's pretraining data, while target languages are deliberately chosen to be unseen during pretraining. Whenever available, two or more *phylogenetically related*[1] source languages (excluding English) are used, as in the Germanic, Slavic, and Indo-Aryan groups. For African languages, where closely related HRL sources are unavailable, source selection is guided by *demographic or geographic proximity*. English is included as a shared source HRL across all language families, yielding a consistent and linguistically grounded experimental setup.

Training and Evaluation Protocol. Our primary experimental framework uses mBERT-base (`bert-base-multilingual-cased`) (Devlin et al., 2019) as the underlying mPLM ($\mathscr{M}_\Theta$). Supplementary results using `XLM-R` (Conneau et al., 2020) are provided in §A.2. We leverage pretrained source-language adapters from AdapterHub[2] , which are kept frozen during training. The trainable components of ZGUL include the shared Task Adapter (TA) inserted at each transformer layer and the fusion parameters (global typological weights and local attention weights).

Evaluation Settings. The primary focus of ZGUL is **Strict 0-CLT**: no target-language data of any kind – labeled or unlabeled – is used during training or inference, representing the most challenging and practically motivated CLT setting.

To demonstrate robustness beyond this regime, we additionally evaluate ZGUL under the **U-CLT** setting, where small amounts of unlabeled target-language data are available for language adapter training. We note that while the evaluation languages do possess some unlabeled data, it is limited in quality and abundance relative to HRLs – they remain genuinely low-resource by any practical measure. Their inclusion is not a contradiction of the 0-CLT motivation; rather, it allows us to assess ZGUL's robustness across varying resource levels (§3.5.2).

Performance is reported as micro-F1 scores computed using the `seqeval` toolkit. All results are averaged over three random seeds, with standard deviations reported in §A.1.3.

Baselines. We compare ZGUL against strong supervised fine-tuning (SFT) and adapter-based baselines:

- **SFT-En / SFT-M (Full Fine-Tuning) (§2.2.1):** Standard fine-tuning of all model parameters Θ. We evaluate two variants: *SFT-En* (trained on English only) and *SFT-M* (trained on the concatenation of all available source languages).

[1] Phylogenetic information is obtained from WALS (`https://wals.info/languoid`)
[2] `https://adapterhub.ml/explore/text_lang/`

- **MADX (§2.2.2.1):** The modular adapter framework, in which we train a single task adapter stacked on English language adapter (LA). At inference, we evaluate two LA selection strategies: *(i) MADX-En*: Uses the English LA (standard 0-CLT transfer), or *(ii) MADX-Rel*: Uses the LA of the source language most typologically similar to the target.
- **CPG (§2.2.2.2):** The conditional parameter generation baseline.
- **Multi-source MADX ($\text{MADX}^{\text{multi}}$):** An extension of MADX where the model is trained on all source languages, activating the corresponding source LA for each batch. At inference, we compare four fusion strategies: *(i) En*: Reuses the English LA, *(ii) Rel*: Reuses the most typologically related source LA, *(iii) Uniform*: Averages the outputs of all source LAs with equal weight, or *(iv) EMEA*: Applies uniform averaging followed by inference-time EM (Eq. 3.3, §3.2).

Implementation Details. All experiments are performed on a single NVIDIA Tesla V100 GPU (32 GB). We implement all adapter-based models, including ZGUL, using the AdapterHub framework.[3] Table A.1 summarizes the hyperparameter search space for full fine-tuning (SFT) and adapter-based methods. For adapters, we tune the reduction factor in $\{3,4\}$ and consistently find 3 to perform best across datasets. Inference-time EM (for both ZGUL and MADX-EMEA) is tuned on the linguistically closest source language's dev set via grid search over the number of iterations $T \in \{1,5,10\}$ and the learning rate $lr \in \{0.05,0.1,0.5,1.0\}$.

3.5 Results and Analysis

To systematically evaluate the ZGUL framework, we investigate the following research questions:

1. **RQ1 (0-CLT Efficacy):** Does ZGUL outperform strong multi-source 0-CLT baselines, such as CPG and SFT, across diverse language families and tasks?
2. **RQ2 (Ablation Analysis):** What are the individual contributions of ZGUL's components – global fusion, local fusion, and inference-time EM – to overall performance?
3. **RQ3 (Robustness to U-CLT):** How robust is ZGUL under alternative supervision regimes, such as unlabeled cross-lingual transfer (U-CLT)?
4. **RQ4 (Interpretability):** Are ZGUL's learned ensemble weights linguistically interpretable, particularly in terms of their correlation with typological similarity between source and target languages?

[3] https://github.com/adapter-hub

3.5.1 0-CLT Results

Tables 3.3 and 3.4 summarize the performance for POS tagging and NER, respectively, across the Germanic, Slavic, African, and Indo-Aryan language groups. ZGUL achieves the best average performance among all evaluated methods in three out of four language groups and yields the highest micro-F1 on 10 of the 15 individual target languages.

Consistent 0-CLT gains. ZGUL outperforms full fine-tuning (SFT) by 3.2 and 0.9 points for Indic and African NER, respectively. It also surpasses CPG by 1.8 points in Germanic POS tagging. While CPG is a strong competitor for POS – likely due to its data-intensive training regime and the relatively larger availability of POS supervision (§A.1.2) – ZGUL remains the most consistent performer across both tasks and all four language families.

Superiority over MADX$^{\text{multi}}$. ZGUL shows significant gains over MADX$^{\text{multi}}$-Rel (up to 7.5 points in Indo-Aryan). This confirms that even when trained on multiple source languages, relying on a single "most related" language's LA at inference is sub-optimal. Furthermore, ZGUL outperforms MADX$^{\text{multi}}$-EMEA by an average of 4 points. This underscores that Entropy Minimization (EM) is highly sensitive to initialization; ZGUL's typological priors provide a grounded starting point that uniform initialization lacks.

Performance on low-scoring targets. ZGUL yields notable gains – as compared to the closest baselines (SFT-M/CPG) – on languages such as Gothic (Got) (20.2 vs. 16.8), Pomak (Qpm) (50.7 vs. 47.6) and Luo (40.2 vs. 36.0). This indicates effective aggregation of complementary signals from multiple source languages that benefits such low-scoring target languages.

We find that absolute performance and ZGUL's gain over baselines are governed by two distinct factors. *Absolute score* correlates strongly, within each task, with token overlap (Table A.3: $r = 0.88$ for POS, $r = 0.53$ for NER) and with maximum source-language similarity (Figure 3.3b: $r = 0.79$ POS, $r = 0.74$ NER) – higher-resource, more typologically similar targets simply achieve higher absolute scores under any method.

Mild gains for some targets. ZGUL's *gain* over the strongest baseline, in contrast, shows no relationship with token overlap ($r = 0.00$) but correlates negatively with typological proximity: both maximum source similarity ($r = -0.54$) and the gap between the best- and second-best-matching source ($r = -0.49$). Concretely, ZGUL's largest gains occur for languages with no strongly similar single source – Luo, Hau, and Got all have low maximum source similarity (≤ 0.35) and show gains of +4.2, +2.3, and +3.4 respectively – while languages with one clearly dominant source, such as Fo (Icelandic, similarity 0.64) and Lug (Swahili, 0.57), show ZGUL's smallest or negative gains (-0.4 and -1.8). This is consistent with ZGUL's design motivation: multi-source fusion adds the most value precisely when no single high-resource source language

is a strong typological match, while a well-matched single source already captures most of the achievable transfer on its own.

	Germanic				Slavic				
	Fo	Got	Gsw	**Avg**	Qpm	Hsb	Orv	Cu	**Avg**
SFT-En	72.4	13.2	55.1	46.9	40.0	64.8	56.5	28.0	47.3
MADX-En	71.6	15.6	57.1	48.1	41.8	63.2	54.3	29.5	47.2
MADX-Rel	72.8	15.2	57.0	48.3	39.3	60.2	57.0	32.2	47.2
MADX-EMEA	74.0	14.8	55.3	48.0	42.1	63.1	56.4	32.0	48.4
MADX$^{\text{multi}}$-Rel	75.8	14.7	60.4	50.3	47.2	74.8	62.9	**35.8**	55.2
MADX$^{\text{multi}}$-Uniform	75.6	8.3	56.4	46.8	46.4	74.6	62.1	34.3	54.4
MADX$^{\text{multi}}$-EMEA	76.2	12.5	62.1	50.3	48.0	75.0	63.2	34.7	55.2
SFT-M	**77.3**	16.8	61.8	52.0	47.6	75.4	63.7	34.7	55.4
CPG	77.0	16.0	63.6	52.2	46.9	76.7	**64.1**	35.6	55.8
ZGUL	76.9	**20.2***	**64.8***	**54.0***	**50.7***	**76.8**	63.0	34.4	**56.2**
ZGUL w/o EM	76.8	14.9	63.0	51.6	49.7	76.0	62.6	33.7	55.5
ZGUL w/o Global Fusion	76.9	17.8	61.3	52.0	49.8	76.0	62.8	33.3	55.5
ZGUL w/o Local Fusion	76.9	18.7	62.4	52.7	49.8	76.5	63.0	33.7	55.8

Table 3.3: **0-CLT POS tagging results** (micro-F1) on *Germanic* and *Slavic* target languages. Best results in **bold**; * denotes statistical significance (McNemar's $p < 0.005$) on aggregated group-level predictions.

Ablation Study. The bottom rows of Tables 3.3 and 3.4 demonstrate that all three ZGUL components consistently improve performance across both tasks and all language groups.

Global Fusion proves to be the most critical; ablating it causes severe average F1 drops across both NER (e.g., −6.4 in Indo-Aryan, −2.0 in African) and POS (−2.0 in Germanic). Local Fusion adds complementary fine-grained routing, yielding further performance gains (e.g., +4.8 in Indo-Aryan NER, +1.3 in Germanic POS). Finally, inference-time EM Refinement drives substantial improvements (+4.7 in Indo-Aryan NER, +2.4 in Germanic POS). EM marginally degrades performance only for isolated target LRLs like Kinyarwanda and Luganda; this typically occurs when the base model is confidently wrong, causing the entropy minimization loop to reinforce errors rather than correct them.

Ultimately, these ablations validate our core hypothesis: global typological grounding, local token routing, and inference-time adaptation act as highly complementary mechanisms for robust 0-CLT transfer across diverse LRLs.

3.5.2 Robustness in U-CLT Setting

To test ZGUL's robustness, we use unlabeled Wikipedia data ($\mathcal{U}_t$) for 12 target languages. For each language $\mathcal{L}_t$, we use $\mathcal{U}_t$ to (1) train its Language Adapter $\text{LA}_{\mathcal{L}_t}$, and (2) pre-train mBERT model using Masked Language Modeling (MLM) objective (Devlin et al., 2019), denoted as $\mathcal{M}_{\Theta_t}$. We extend ZGUL to U-CLT, denoted as ZGUL-U, by initializing the base model with $\mathcal{M}_{\Theta_t}$

	African							Indo-Aryan		
	Hau	Ibo	Kin	Lug	Luo	Pcm	**Avg**	As	Bh	**Avg**
SFT-En	43.2	47.1	49.1	47.5	29.0	64.4	46.7	32.1	35.7	33.9
MADX-En	41.3	51.2	46.0	49.8	29.2	64.5	47.0	33.0	44.8	38.9
MADX-Rel	39.1	49.1	46.3	48.7	29.7	65.1	46.3	42.4	46.9	44.7
MADX-EMEA	42.0	52.9	50.2	50.6	30.7	65.5	48.7	41.6	47.6	44.6
MADX$^{\text{multi}}$-Rel	47.8	49.6	49.6	46.8	31.8	62.4	48.0	61.7	61.9	61.8
MADX$^{\text{multi}}$-Uniform	47.0	51.1	50.8	49.9	31.7	62.7	48.7	60.0	59.0	59.5
MADX$^{\text{multi}}$-EMEA	47.4	54.4	55.0	51.1	32.8	63.9	50.8	62.4	59.7	61.1
SFT-M	51.3	55.7	**57.9**	**56.0**	36.0	65.0	53.7	70.8	61.4	66.1
CPG	49.0	51.1	55.1	54.0	34.2	65.7	51.5	62.0	63.3	62.7
ZGUL	**53.6***	**56.8***	56.2	54.2	**40.2***	**66.5***	**54.6***	**74.4***	**64.1**	**69.3***
ZGUL w/o EM	52.4	55.6	57.3	55.1	38.6	65.5	54.1	71.1	58.0	64.6
ZGUL w/o Global Fusion	49.0	54.7	57.3	53.5	35.5	65.5	52.6	61.5	64.2	62.9
ZGUL w/o Local Fusion	53.1	54.1	57.7	55.7	36.6	65.6	53.8	67.2	61.7	64.5

Table 3.4: **0-CLT NER results** (micro-F1) on *African* and *Indo-Aryan* target languages. Best results in **bold**; * denotes statistical significance (McNemar's $p < 0.005$) on aggregated group-level predictions.

	Fo	Got	Gsw	Hsb	Cu	Hau	Ibo	Kin	Lug	Pcm	As	Bh	**Avg**
MADX$^{\text{multi}}$-U	81.8	15.5	76.1	87.6	32.8	69.7	64.0	53.1	54.8	61.8	57.7	64.4	59.9
SFT-U	81.9	16.5	76.1	**91.5**	34.6	75.8	**71.8**	68.0	64.0	66.9	68.1	65.9	65.1
ZGUL-U	**82.3**	**17.9**	**82.7**	**91.5**	**35.4**	**77.7**	71.4	69.2	**64.4**	**67.6**	73.5	**70.5**	**67.0**
ZGUL-U w/o $\text{LA}_{\mathscr{L}_t}$	82.1	17.8	79.7	90.5	35.1	77.1	71.7	**69.7**	63.7	67.2	**74.1**	67.9	66.4
ZGUL-U w/o $\mathscr{M}_{\Theta_t}$	78.5	15.8	69.5	87.1	34.2	64.1	60.4	60.5	56.2	65.8	64.6	66.7	60.3
Size of $\mathscr{U}_t$	160k	2k	100k	150k	8k	350k	230k	30k	35k	8k	250k	45k	–

Table 3.5: **U-CLT results** (micro-F1) after incorporating unlabeled target-language data. Bottom row: size of unlabeled corpus $\mathscr{U}_t$ (in sentences).

and fine-tuning it using the composition $\mathscr{M}_\Theta \circ \text{LA}^{\text{fuse}} \circ \text{TA}_{\mathscr{T}}$, where $\text{LA}^{\text{fuse}} = \{\text{LA}_{\mathscr{L}_s}\}_{s\in\mathscr{S}} \cup \{\text{LA}_{\mathscr{L}_t}\}$, i.e., the target LA is inserted along with source LAs for fusion. Importantly, training still uses only labeled source-language data, as no labeled data is available for $\mathscr{L}_t$.

We compare ZGUL-U against two baselines: (1) MADX$^{\text{multi}}$-U, which extends MADX$^{\text{multi}}$ to the U-CLT setting by using $\text{LA}_{\mathscr{L}_t}$ at inference; and (2) SFT-U, which initializes the base model with $\mathscr{M}_{\Theta_t}$ and fine-tunes on source-language data.

As shown in Table 3.5, ZGUL-U outperforms the competitive SFT-U baseline by an average of 2 F1 points, with a notable 6.6-point gain in Swiss German. Ablations reveal that while incorporating the target adapter $\text{LA}_{\mathscr{L}_t}$ yields an average gain of almost 0.6 F1 point, the dominant performance boost (~7 F1 points) stems from initialization with $\mathscr{M}_{\Theta_t}$, i.e., the target-language continual pre-training (CPT) of the base model.

Finally, we analyze performance scaling with varying amounts of target data $\mathscr{U}_t$. We evaluate models using 0%, 25%, 50%, and 100% of the target Wikipedia corpus and report the average. F1 across all 12 target languages (Figure 3.2). ZGUL-U consistently outperforms base-

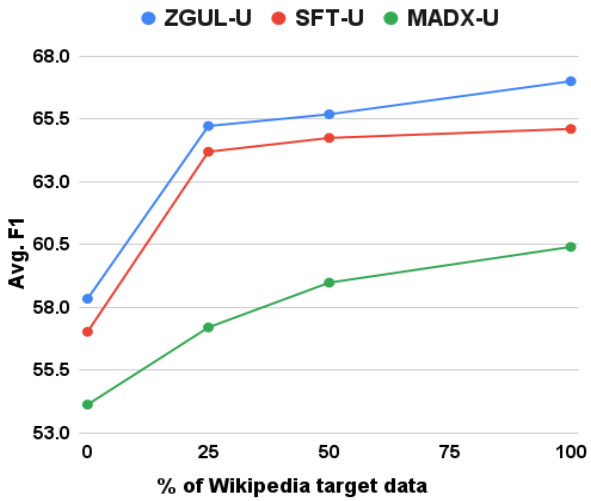


Figure 3.2: **Effect of unlabeled data scale**: average F1 over target languages vs. percentage of Wikipedia data used. *0%* denotes the 0-CLT setting.

lines across all data volumes. The gap between ZGUL-U and SFT-U widens as more data is introduced, while ZGUL-U achieves significant gains over MADX$^{\text{multi}}$-U with as little as 25% of the target corpus.

We additionally test ZGUL's robustness under the few-shot (FS-CLT) setting in the Appendix A.3.

3.5.3 Qualitative and Error Analysis

3.5.3.1 Global Fusion: Linguistic Interpretability

We analyze the interpretability of ZGUL's fusion by calculating the Spearman correlation (Spearman, 1904) between inference-time attention weights and language-relatedness (defined via `lang2vec` syntactic and genetic similarity). Since attention weights are normalized (softmax) while Lang2Vec scores are independent pairwise similarities, only a monotonic relationship is expected between them; following Brandl and Hollenstein (2022), who relate model attention to an external signal via rank correlation for the same reason, we adopt Spearman's ρ accordingly.

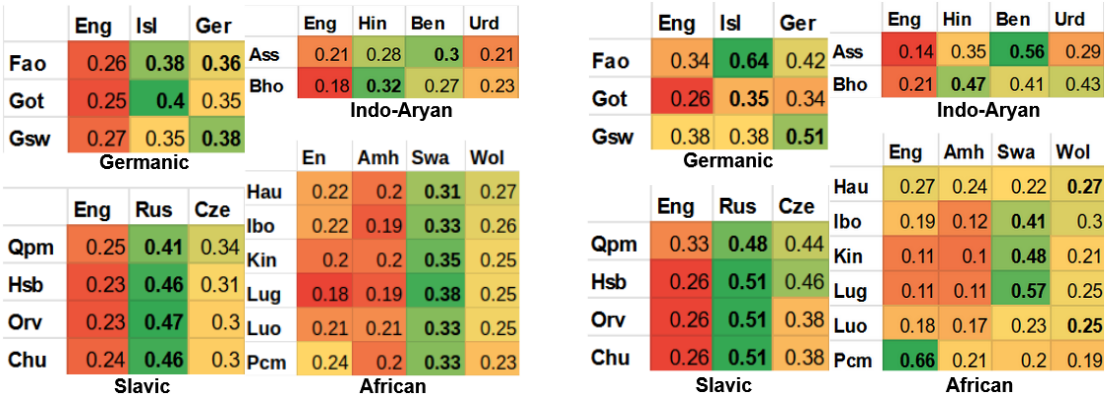
Group	Global	Local	M$^{\text{multi}}$-EM
Germanic	**0.66**	0.57	0.22
Slavic	**0.98**	0.86	0.14
African	**0.73**	0.65	0.44
Indo-Aryan	**0.84**	0.55	0.19

Table 3.6: **Spearman Correlation between attention weights and source-target linguistic similarity** (syntactic-genetic averaged). *Global/Local*: ZGUL's fusion networks; *M^{multi}-EM*: MADX$^{\text{multi}}$-EMEA baseline.

- **Strong Linguistic Correlation**: Table 3.6 shows that ZGUL's weights correlate strongly with

linguistic relatedness, particularly in Slavic (0.98) and Indo-Aryan (0.84). This suggests ZGUL successfully assigns higher importance to linguistically related source languages.

- **Global vs. Local**: Correlations are higher for global fusion than local fusion consistently across all four families. This is expected, as global fusion uses explicit language features, whereas local fusion must infer linguistic relatedness implicitly from token semantics.
- **Comparison with Baselines**: The MADX^{multi}-EMEA baseline shows substantially weaker correlation with linguistic relatedness than either of ZGUL's fusion mechanisms, in every family (lowest: Slavic, 0.14). This confirms that inference-time ensembling is considerably less effective at capturing typological structure without the typologically grounded initialization ZGUL provides.
- **Visual Alignment**: Figure 3.3 shows a strong visual alignment between ZGUL's attention heatmaps and ground-truth linguistic similarity patterns, visually corroborating the correlation results.



(a) **Global fusion** attention scores over source LAs per target language (grouped by family).

(b) **Source–target relatedness**: average of syntactic and genetic similarity (Lang2Vec).

Figure 3.3: **Interpretability of Global Fusion.** Learned attention weights (left) align with typological relatedness (right), confirming that ZGUL assigns higher weight to linguistically closer source languages.

We note that this attention-level analysis evidences *learned typological routing* rather than token-level input attribution: prior work shows that raw attention weights are not reliable input attributions (Jain and Wallace, 2019; Amini and Kosseim, 2022). This analysis would be further strengthened by a systematic per-instance input-attribution study – spanning each task and source-target pair, presented across true and error classes, applied consistently to ZGUL and its baselines, with perturbation-based faithfulness validation – which we leave to future work.

3.5.3.2 EM: Effect on Predictions and Confidence

To understand *how* EM Refinement changes model behavior beyond aggregate F1, we compare ZGUL predictions with and without EM, word-by-word, on Germanic POS (Got, Fo, Gsw) and

Model	Got			Fo			Gsw		
	True	Error	ECE	True	Error	ECE	True	Error	ECE
w/ EM	2771	7423	0.409	7850	2152	0.140	984	460	0.125
w/o EM	1905	8287	0.551	7826	2176	0.164	950	493	0.185

Table 3.7: **POS (Germanic): True/Error counts and ECE**, with and without EM.

Model	Hau			Ibo			Kin			Lug			Luo		
	True	Error	ECE	True	Error	ECE	True	Error	ECE	True	Error	ECE	True	Error	ECE
w/ EM	15479	1362	0.074	11655	**990**	0.072	13242	954	0.063	9048	**793**	0.075	4540	619	0.112
w/o EM	15400	1441	0.077	11666	979	0.065	13229	967	0.057	9063	778	0.067	4532	627	0.111

Table 3.8: **NER (African): True/Error counts and ECE**, with and without EM.

African NER (Hau, Ibo, Kin, Lug, Luo) targets, tracking correctness, confidence, and label transitions. Got is a striking case: removing EM drops its POS F1 from 20.2 to 14.9 (Table 3.3), the largest single-language degradation observed across all target languages and both tasks; the analysis below characterizes this and other languages' EM-sensitivity in detail.

POS and NER respond to EM very differently. (1) For POS, EM consistently improves calibration (ECE: Got 0.551 → 0.409; Fo 0.164 → 0.140; Gsw 0.185 → 0.125) alongside a favorable corrected-to-broken ratio in every language (11.6:1 Got, 1.85:1 Gsw, 1.19:1 Fo). For NER, this reverses: average ECE is slightly *worse* with EM than without (0.079 vs. 0.075; Table 3.10), worsening in four of five languages (Ibo 0.065 → 0.072; Kin 0.057 → 0.063; Lug 0.067 → 0.075; Luo flat), with only Hau improving, and two languages (Ibo, Lug) see EM break more predictions than it fixes (0.81:1, 0.69:1). Mean confidence also moves in opposite directions: EM *lowers* confidence across nearly all POS categories, but *raises* it in every flip category for every NER language, including CORRECT→WRONG (e.g., Hau: 0.783 → 0.863) – a reinforcement pattern that is broad and consistent for NER, unlike POS where it was confined to Got's BOTH CORRECT cases.

(2) We attribute this asymmetry to label-space size. POS corrections draw on a rich tag inventory – e.g. Got's top corrections resolve a placeholder-tag fallback (`X→NOUN`, 398 occurrences; `X→PROPN`; `X→ADP`), while Fo/Gsw show genuine `PROPN↔NOUN`, `AUX↔VERB` disambiguation – giving EM's entropy minimization many productive directions to redistribute probability mass. NER's top transitions, by contrast, are almost entirely `B-XXX→O` or `O→B-XXX` (e.g. Hau: `B-PER→O` (51), `I-PER→O` (46)) – an effectively binary entity-vs-non-entity decision. With less room to productively redistribute mass, EM still sharpens confidence (a mechanical effect of entropy minimization) without a matching accuracy gain, which is precisely what degrades calibration rather than improving it.

Model Size and Training Cost. Table 3.9 compares trainable parameter counts and per-epoch training time. ZGUL introduces substantially fewer trainable parameters than SFT (~4.5×) and CPG (~6×), while maintaining comparable training time to SFT and being significantly more efficient than hypernetwork-based CPG (~4× lesser training time per epoch).

		Per-epoch training time (min)				
Model	**Params**	Indo-Aryan	African	Germanic	Slavic	**Avg**
SFT-M	177M	9	3	36	30	19.5
CPG	253M	37	13	141	115	76.5
ZGUL	41M	14	5	36	29	21.0

Table 3.9: **Efficiency comparison**: trainable parameters and per-epoch training time (minutes) across language groups.

3.5.4 Precision and Calibration Analysis

While the overall micro-F1 scores demonstrate ZGUL's superiority in strict 0-CLT settings, a class-wise analysis reveals its key behavioral advantage: *remarkably high precision* on core semantic and syntactic classes.

As detailed in Tables 3.11 and 3.12, ZGUL consistently yields the highest average precision for critical entities like `PER` (0.66) and structural tags like `VERB` (0.65) and `ADP` (0.86). Even though ZGUL occasionally trails in recall for rarer classes (such as `DATE`), its overall micro-precision remains the highest among all methods (e.g., 0.59 for NER compared to SFT's 0.56). This indicates that ZGUL is a highly conservative and accurate predictor: when it assigns a label to a token, it is statistically more likely to be correct than the baseline models.

Beyond precision, the utility of a Stage 1 initializer heavily depends on its *calibration* – the degree to which predicted confidence scores align with actual accuracy. We quantify this using the *Expected Calibration Error* (ECE), where a lower score indicates a more trustworthy model. As shown in Table 3.10, ZGUL significantly improves calibration on the fine-grained Germanic POS tasks, reducing the average ECE to 0.22 (compared to 0.36 for SFT). For African NER, all models exhibit naturally low calibration errors due to a much smaller label space (4 labels compared to 16 for POS), with ZGUL maintaining a competitive average ECE of 0.08.

	Germanic POS				African NER					
Model	**Gsw**	**Fo**	**Got**	**Avg.**	**Hau**	**Ibo**	**Kin**	**Lug**	**Luo**	**Avg.**
SFT	0.25	0.19	0.65	0.36	**0.07**	**0.07**	**0.05**	**0.07**	**0.11**	**0.07**
CPG	0.21	0.17	0.58	0.32	**0.07**	**0.07**	0.06	**0.07**	**0.11**	0.08
ZGUL w/o EM	0.18	0.16	0.55	0.30	0.08	**0.07**	0.06	**0.07**	**0.11**	0.08
ZGUL	**0.12**	**0.14**	**0.41**	**0.22**	**0.07**	**0.07**	0.06	**0.07**	**0.11**	0.08

Table 3.10: **Expected Calibration Error (ECE)** (rounded to two decimal places) across target languages for Germanic POS and African NER.

	ZGUL			CPG			SFT		
Entity Class	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
PER	**66**	68	67	62	66	64	61	**69**	64
LOC	60	53	55	**66**	45	52	57	**53**	53
ORG	51	55	53	**53**	45	48	51	**56**	**53**
DATE	30	17	21	31	18	**22**	**36**	**18**	**22**
Micro Avg.	**59**	56	**57**	58	50	53	56	**56**	55

Table 3.11: **Class-wise performance on African NER** (averaged across all target languages) (P: Precision, R: Recall).

	ZGUL			CPG			SFT		
POS Tag	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
VERB	**65**	**47**	51	63	46	51	64	44	**52**
ADP	**86**	**75**	**79**	85	72	77	79	72	75
NOUN	**57**	**62**	**59**	56	58	57	**57**	60	58
PROPN	51	**64**	**55**	**56**	59	53	53	58	53
ADJ	52	42	44	**56**	40	44	51	**43**	44
Micro Avg.	**56**	**55**	**56**	**56**	53	54	55	52	54

Table 3.12: **Class-wise performance on Germanic & Slavic POS tags** (averaged across all target languages).

Motivation for Hybrid Prompting. These dual characteristics – high precision and reliable confidence calibration – form the foundational motivation for the next phase of this thesis. In the paradigm of In-Context Learning (ICL) with Large Language Models (LLMs), *precision is far more valuable than recall*. To construct an effective K-shot prompt (e.g., $K = 8$), an LLM does not need a Stage 1 model that covers every entity type; rather, it requires a small candidate pool where the provided pseudo-labels are reasonably correct.

Furthermore, because the exemplar selection algorithm in Chapter 4 explicitly thresholds candidates based on their confidence scores ($\hat{p}_i^{(l)}$), ZGUL's well-calibrated predictions are essential to prevent the propagation of overconfident errors into the prompt. Finally, any limitations in Stage 1 recall are directly mitigated during Stage 2: the explicit *label coverage constraint* ensures that the final prompt covers the full label space, structurally compensating for any entity types missed by the initializer.

By successfully suppressing false positives, providing trustworthy confidence scores, and deferring recall management to the prompt design, ZGUL produces an ideal, low-noise candidate pool for Self-Supervised Prompting (SSP) in Chapter 4.

3.6 Conclusion

This chapter presented ZGUL (Zero-resource Generalization to Unseen Languages), a multi-source framework for NER and POS tagging that overcomes the rigidity and high cost of prior

methods by dynamically routing knowledge from multiple source languages to unseen targets.

By combining global typological priors with local token-level routing and inference-time Entropy Minimization, ZGUL consistently outperforms baselines across four language families in the strict 0-CLT setting. Its fusion weights align with linguistic similarity, offering clear interpretability, and it scales effectively when unlabeled target data is available (U-CLT). Finally, ZGUL's high precision and well-calibrated confidence scores make it an ideal initializer for the hybrid prompting framework introduced in the next chapter. We release our code and trained checkpoints to support parameter-efficient and scalable CLT research in sequence labeling.[4]

[4] https://github.com/dair-iitd/ZGUL

Chapter 4

Self-Supervised Prompting for Cross-Lingual Transfer using Large Language Models

4.1 Introduction

The emergence of *Large Language Models (LLMs)* such as GPT-4 (Achiam et al., 2023), PaLM (Chowdhery et al., 2023), and LLaMA (Touvron et al., 2023) has fundamentally altered the NLP landscape. This trajectory has continued rapidly since, with newer model families such as Gemini (Gemini Team, 2023), Mistral (Jiang et al., 2023a), Llama 3 (Dubey et al., 2024), GPT-4o (OpenAI, 2024), and Qwen3 (Yang et al., 2025a) further advancing multilingual and reasoning capabilities. Through the paradigm of *In-Context Learning (ICL)*, these models can tackle complex tasks by relying solely on natural language instructions and a small set of labeled exemplars (Brown et al., 2020). However, despite their impressive generalization, LLMs exhibit a "curse of multilinguality": their performance remains uneven across languages, with substantial degradation for low-resource languages (LRLs) in cross-lingual transfer (CLT) settings (Ahuja et al., 2023; Asai et al., 2024a), a gap that persists even in the newest model generations, as shown throughout Chapters 4 and 6.

Chapter 3 showed that modular transfer frameworks such as ZGUL can partially mitigate this gap by leveraging linguistic relatedness and multi-source supervision. While ZGUL achieves strong performance in strict 0-CLT settings, it is ultimately built on discriminative encoders such as mBERT or XLM-R. Constrained to a fixed representational space, these models often struggle with the adaptive reasoning required for semantically complex or ambiguous inputs (Kurniawan et al., 2022), lacking the emergent world knowledge and contextual flexibility inherent in modern generative LLMs (Roberts et al., 2020; Zoph et al., 2022).

Conversely, while LLMs possess this contextual flexibility, their success relies critically on the availability of *high-quality*, *task-relevant exemplars* (Zhao et al., 2021; Liu et al., 2022).

Existing retrieval-based methods struggle to identify suitable in-context exemplars for structured prediction tasks (Blevins et al., 2023), particularly in cross-lingual settings (Ahuja et al., 2023). Consequently, even strong models like GPT-4 exhibit suboptimal performance on structured prediction problems (Asai et al., 2024a; Parekh et al., 2024).

This dichotomy motivates a *hybrid* paradigm. We posit that task-specific cross-lingual models (like ZGUL) can play a complementary role by providing structured, language-aware guidance to construct suitable exemplars for the LLM.

This chapter proposes *Self-Supervised Prompting* (SSP), a *two-stage* framework that bridges specialized cross-lingual transfer models and LLM prompting. SSP is driven by a core empirical observation: LLMs benefit more from linguistically aligned, *monolingual* (or *in-language*) exemplars – even those with moderate label noise – than from clean exemplars drawn from a different language. We operationalize this via a two-stage pipeline (Figure 4.1):

- **Stage 1 - Pseudo-Labeling:** A cross-lingual model (e.g., ZGUL) generates initial labels and confidence estimates for unlabeled target-language data.
- **Stage 2 - ILP-Based Exemplar Selection for monolingual prompting:** An Integer Linear Programming (ILP) formulation filters these pseudo-labeled instances – jointly optimizing for semantic similarity, prediction confidence, and label diversity – to construct high-quality, monolingual in-context exemplars for LLM prompting.

This hybrid design combines complementary strengths: the task-specific model supplies structured, language-aware supervision, while the LLM contributes contextual reasoning that can denoise weak labels and improve generalization. Experiments on POS tagging, NER, and Natural Language Inference across eleven low-resource languages demonstrate that SSP achieves consistent gains over both fine-tuned and prompting-based baselines.

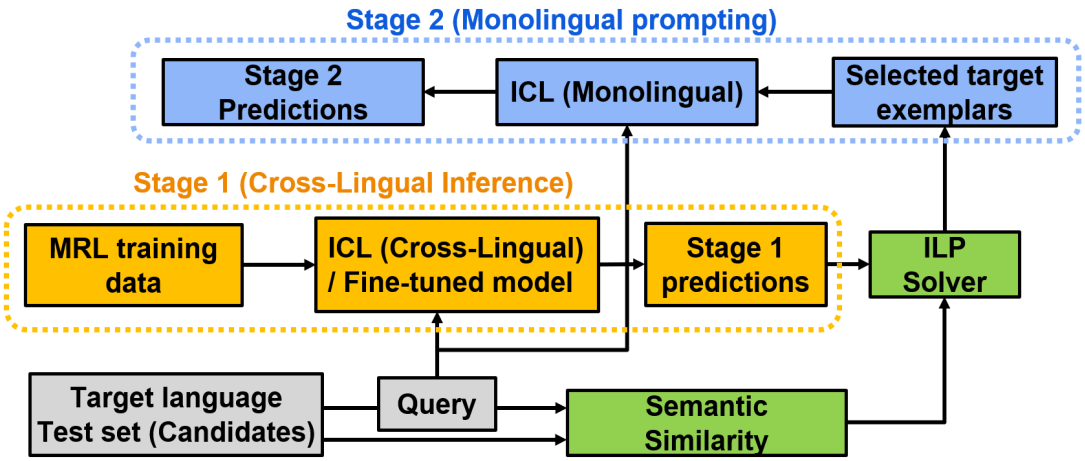


Figure 4.1: **SSP pipeline.** *Stage 1*: Fine-tune a CLT model (e.g., ZGUL) or perform cross-lingual prompting using MRL data to generate pseudo-labels. *Stage 2*: Select exemplars from the pseudo-labeled pool via ILP and perform ICL using the target-language exemplars.

4.2 Background and Related Work

This section positions Self-Supervised Prompting (SSP) within the literature on cross-lingual in-context learning (ICL) and *adaptive prompting*.

Exemplar Sensitivity in In-Context Learning. As discussed in Chapter 2, ICL performance depends heavily on the choice and ordering of exemplars, with even minor variations substantially shifting the model's inductive bias (Brown et al., 2020; Zhao et al., 2021; Min et al., 2022). To address this sensitivity, prior work has developed retrieval strategies that balance *similarity-based* retrieval – selecting exemplars closest to the query in embedding space – with *diversity-aware* techniques like Maximal Marginal Relevance or subset selection to prevent overfitting (Gupta et al., 2023; Wang et al., 2025). While exemplar relevance strongly correlates with accuracy and diversity improves robustness (Agrawal et al., 2022), most of these studies assume access to clean, gold-labeled exemplar pools and focus primarily on English benchmarks.

Challenges with Cross-lingual Prompting. When ICL is applied in cross-lingual settings, these assumptions break down. Without labeled target-language examples, systems typically rely on source-language or translated exemplars retrieved via multilingual encoders (Tanwar et al., 2023; Nambi et al., 2023). However, this retrieval process remains largely *static* and *task-agnostic*: selection is driven mainly by embedding similarity without explicitly considering task structure, label distribution, or prediction confidence. Consequently, such static strategies yield unstable behavior in low-resource languages and frequently underperform supervised transfer models for structured prediction tasks (Asai et al., 2024a; Parekh et al., 2024). This instability highlights that effective cross-lingual prompting requires mechanisms that explicitly incorporate task-specific signals into exemplar construction.

Self-Adaptive Prompting. A complementary line of research explores *self-adaptive prompting*, where ICL exemplars are iteratively refined using model-internal feedback. Approaches like *Universal Self-adaptive Prompting* (USP) (Wan et al., 2023b) assign initial pseudo-labels to unlabeled data and subsequently re-prompt the model using examples selected via entropy or self-consistency metrics. Although effective for English benchmarks, these methods face significant challenges in cross-lingual transfer: they incur high computational overhead due to repeated inference steps (Wan et al., 2023a), rely heavily on often poorly calibrated model confidence estimates (Xiao et al., 2022), and leave behavior under linguistic divergence largely unexplored.

Positioning of Self-Supervised Prompting (SSP). SSP builds on the intuition behind adaptive prompting but differs from prior work in two critical ways. First, instead of relying solely on LLM-generated feedback, SSP incorporates supervision from specialized cross-lingual models

(such as ZGUL) to guide exemplar construction. Second, SSP is designed explicitly for cross-lingual transfer under weak supervision, directly addressing the monolingual limitations of methods like USP. By jointly leveraging pseudo-label confidence, semantic similarity, and diversity constraints, SSP enables dynamic, task-aware exemplar construction that is computationally efficient and robust to cross-lingual variation in structured prediction settings.

4.3 Self-Supervised Prompting (SSP)

4.3.1 A Thought Experiment

A central motivation for SSP emerges from the following thought experiment. We prompt an LLM (GPT-4) on target LRL queries using two distinct sources of in-context exemplars:

1. **Cross-lingual (Clean):** Exemplars drawn from related MRL training data with gold-standard labels.

2. **Monolingual (Noisy):** Automatically generated target-language exemplars whose labels may contain noise.

To systematically study this effect, we inject synthetic label noise into the target-language exemplars. A fixed fraction of labels is randomly corrupted by replacing the gold tag with an incorrect one, simulating the imperfect pseudo-labels typically produced by cross-lingual models. Performance is evaluated with varying noise rates.

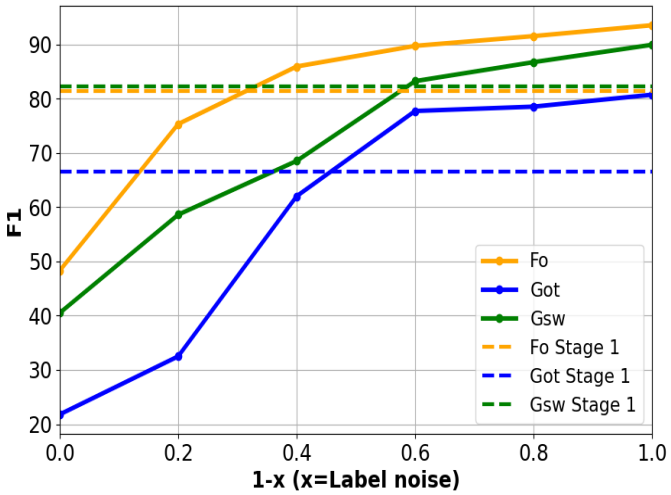


Figure 4.2: **Sensitivity to label noise**: GPT-4 POS tagging F1 on 3 Germanic LRLs with varying fractions of corrupted target-language exemplar labels. Dashed lines: performance with clean source MRL exemplars.

As shown in Figure 4.2, results reveal a consistent trend. While performance naturally improves as label noise decreases, there is a critical crossover point: once the noise rate falls below approximately 40–60%, the noisy monolingual exemplars consistently outperform the clean cross-lingual ones.

This suggests that for modern LLMs, *linguistic alignment* – the matching of script, morphology, and syntactic patterns between the exemplars and the query – can outweigh moderate imperfections in labels themselves.

This observation motivates the design of SSP's *two-stage* pipeline. Rather than relying solely on cross-lingual exemplars from MRLs, SSP first constructs a pool of pseudo-labeled target-language instances and then selects the most informative monolingual exemplars to guide in-context learning. We describe our detailed pipeline in the following sections.

4.3.2 Stage 1: Cross-Lingual Labeling

Stage 1 constructs a pseudo-labeled *candidate pool* for the target LRL using supervision from related MRLs. Depending on available resources, the MRL training data is utilized in one of two ways to annotate the target LRL test set:

1. **Fine-tuned CLT model:** The MRL task data is used to train a specialized cross-lingual model, such as mDeBERTa (He et al., 2021a) or ZGUL (Chapter 3). This model generates pseudo-labels for the unlabeled LRL test set.

2. **Cross-lingual Prompting:** An LLM is prompted using the MRL training data as few-shot exemplars to predict labels for the LRL instances.

Both variants produce pseudo-labels $\hat{y}_i^{(t)}$ and, if available, confidence scores $\hat{p}_i^{(t)}$. Even when confidence scores are not available – as is common with proprietary LLMs like GPT-4 – the labels provide linguistically relevant, albeit noisy, structural cues. The objective of Stage 1 is not perfect accuracy, but breadth: generating a rich, diverse candidate pool from which high-quality monolingual exemplars can be extracted in Stage 2.

4.3.3 Stage 2: Monolingual Prompting with Noisy Exemplars

Stage 2 transitions from cross-lingual inference to monolingual prompting. For each target query q, the LLM receives a prompt containing K pseudo-labeled exemplars drawn directly from the target language candidate pool. Despite potential label noise, these exemplars provide linguistically aligned context to the LLM, leading to potentially improved predictions – consistent with the empirical trends in the previously discussed thought experiment.

Because exemplar choice heavily dictates in-context learning performance, SSP eschews heuristic retrieval (e.g., nearest neighbors) in favor of a structured selection mechanism that ensures the exemplars are collectively informative, reliable, and representative.

ILP-Based Exemplar Selection. Exemplar selection in SSP must jointly ensure (a) *semantic relevance* to the query (Liu et al., 2022), (b) *label reliability* (Zhao et al., 2021), and (c) *label coverage* (Min et al., 2022) – a factor especially critical for sequence labeling tasks having

long-tailed label distributions (e.g., POS tagging) (Huang et al., 2021). To enforce these objectives jointly, SSP formulates exemplar selection as a constrained optimization problem using an Integer Linear Program (ILP).

Let $\mathscr{C} = \{(x_i^{(t)}, \hat{y}_i^{(t)}, \hat{p}_i^{(l)})\}_{i=1}^{n}$ denote the pseudo-labeled candidate pool for target language $\mathscr{L}_t$, where $x_i^{(t)}$ is an input instance, $\hat{y}_i^{(t)}$ is its predicted label sequence, and $\hat{p}_i^{(l)}$ is the aggregate confidence for label $l \in \mathscr{Y}$ in instance i. As detailed in §B.1.1, this confidence is computed as the average probability score over all words assigned label l in the i-th exemplar; for labels not predicted in an instance, $\hat{p}_i^{(l)} = 0$.

For a target query q, we compute the semantic similarity score for the i-th exemplar as $s_i = \text{sim}(x_i^{(t)}, q)$, computed as cosine similarity between query and candidate embeddings. Binary decision variables $z_i \in \{0, 1\}$ indicate whether the i-th exemplar is selected. The ILP maximizes aggregate similarity while adhering to constraints for confidence filtering (eligible exemplars must exceed a label-wise threshold τ_l) and label coverage (each label must appear at least once in the selected exemplars). The formal optimization is defined as:

$$
\begin{aligned}
\max_{z_i} \quad & \sum_{i=1}^{n} z_i \cdot s_i && \text{(maximize similarity)} \\
\text{such that} \quad & \sum_{i=1}^{n} z_i = K && \text{(select } K \text{ exemplars)} \\
& z_i \cdot (\hat{p}_i^{(l)} - \tau_l) \geq 0, \quad \forall i, \forall l \in \hat{y}_i && \text{(confidence thresholding)} \\
& \sum_{i=1}^{n} z_i \cdot \text{count}(\hat{y}_i, l) \geq 1, \quad \forall l \in \mathscr{Y} && \text{(label coverage)}
\end{aligned}
$$

where $\text{count}(\hat{y}_i, l)$ denotes the number of tokens in instance i assigned label l. We set the label-wise threshold τ_l to the 80$^{\text{th}}$ percentile of confidence scores for pseudo-label l occurrences across the development set instances in a given language. When confidence scores are unavailable (e.g., with proprietary LLMs), the confidence thresholding constraint is naturally relaxed, relying solely on the label coverage constraint.

Why ILP? Compared to iterative subset selection (Gupta et al., 2023) or RL-based methods (Wang et al., 2025), ILP performs selection in a single optimization step, is easily extensible with new constraints (e.g., token-length limits (Liu et al., 2024), translation uncertainty (Conneau et al., 2020)), and yields fully interpretable selection decisions.

4.4 Experimental Setup

This section outlines the tasks, evaluation settings, baselines, and configuration choices used to assess SSP.

Tasks. Consistent with Chapter 3, we focus primarily on sequence labeling tasks – POS tagging and NER – evaluating Germanic POS tagging and African NER across eight typologically diverse languages (details in §B.1.3). To manage the substantial inference costs incurred by evaluating proprietary LLMs across multiple prompting configurations, we randomly sample 100 instances from each language's test set for our evaluations. We restrict our study to Roman-script languages to mitigate *tokenizer fertility* – the number of subword tokens generated per word – which is substantially higher for most non-Roman scripts (Nag et al., 2024). Excessive subword fragmentation can degrade representations and reduce the ICL effectiveness (Ahuja et al., 2023). Extending SSP to non-Roman scripts remains an important direction for future work.

To assess generalization beyond token-level prediction, we additionally evaluate SSP on Natural Language Inference (NLI), a sentence-level classification task, across three indigenous American target languages (Ebrahimi et al., 2022). Detailed results for this task are provided in §B.2.

CLT Evaluation Settings. While 0-CLT serves as our primary evaluation setting, we additionally evaluate SSP under T-CLT and U-CLT regimes to assess its robustness across varying levels of resource availability. Specifically, we explore the following configurations:

1. **T-CLT:** We employ both *translate-train* and *translate-test* variants for Stage 1 labeling:
    - **Translate-train:** English NER training data is translated & projected into target African languages using CODEC (Le et al., 2024) with NLLB-200 (NLLB Team et al., 2022) translation model. This data is used either to train ZGUL or to construct an exemplar pool for prompting GPT-4 during Stage 1 labeling. This pipeline is omitted for POS tagging, as reliable token-level label projection is currently unsupported for this task.
    - **Translate-test:** The LRL query is translated into English using NLLB-200, and obtain its Stage 1 labels using ZGUL. Because accurately back-projecting these English labels onto the original LRL tokens is highly error-prone, we do not use ZGUL to output the final LRL labels directly. Instead, we adopt the *Self-Fusion* strategy (Chen et al., 2025): the LLM is prompted with the original LRL query, its English translation, and the ZGUL-predicted English labels to generate the final Stage 1 pseudo-labels.
2. **U-CLT:** For Stage 1 labeling, we utilize ZGUL-U (introduced in Chapter 3). Additionally, for African NER, we fine-tune AfriBERTa (Ogueji et al., 2021) – a language model pretrained explicitly on unlabeled African corpora – to serve as an alternative Stage 1 initializer.

Performance is reported using micro-F1, the standard metric for sequence labeling (following Chapter 3). We compare SSP against two major categories of methods:

Fine-tuned CLT models. We utilize state-of-the-art baselines, including ZGUL (Chapter 3) for Sequence labeling and mDeBERTa (He et al., 2021a) for NLI. These serve both as competitive CLT systems and as Stage 1 labelers when confidence scores ($\hat{p}_i^{(t)}$) are required for ILP selection.

Prompting baselines. We evaluate in-context learning (ICL) using three Large Language Models (LLMs) – *GPT-4x* (GPT-4 and GPT-4-Turbo) (Achiam et al., 2023), *GPT-3.5-turbo* (Achiam et al., 2023), and *LLaMA-2-70B* (Touvron et al., 2023) – under two exemplar retrieval strategies:

1. **ICL(*TopK*)-LLM:** Standard top-K similarity-based retrieval.
2. **ICL(*MMR*)-LLM:** Maximal Marginal Relevance (MMR)-based selection.

   Following Kapuriya et al. (2025), MMR selects exemplars iteratively, choosing a candidate exemplar s at time step j based on the trade-off between relevance to the query (q) and diversity from already selected exemplars (S_{j-1}):

   $$\text{MMR}(q, s, S_{j-1}) = \alpha \cdot \text{sim}(q, s) - (1 - \alpha) \cdot \max_{s' \in S_{j-1}} \text{sim}(s, s'),$$

   where α is fixed at 0.5 across all experiments.

We adopt MMR as our representative diversity-based baseline rather than more expensive alternatives such as RL-based selection (Wang et al., 2025), LM-based determinantal point process (DPP) selection (Wang et al., 2024b), or coverage-based subset selection (Gupta et al., 2023): these perform iterative or combinatorial selection and are substantially costlier at scale, while pursuing a similar diversity objective to MMR.

Semantic similarity is calculated as cosine similarity using OpenAI's `text-embedding-ada-002` multilingual embeddings.[1]

SSP Configurations. SSP is evaluated under multiple (Stage 1, Stage 2) combinations:

- **SSP(ICL)-M:** Stage 1 uses cross-lingual prompting with LLM M; Stage 2 performs monolingual prompting with the same LLM.
- **SSP(ZGUL)-M or SSP($mDeBERTa_{FT}$)-M:** Stage 1 uses a fine-tuned model (ZGUL or $mDeBERTa_{FT}$); Stage 2 performs monolingual ICL using LLM M.

Implementation Details We use the Azure OpenAI Service[2] for all experiments involving GPT-3.5-turbo (`gpt-3.5-turbo-0613`), GPT-4 (`gpt-4-0613`), and GPT-4-Turbo (`gpt-4-1106-preview`). For LLaMA-2-70B, we use the Together API[3].

[1] https://platform.openai.com/docs/guides/embeddings/embedding-model
[2] https://azure.microsoft.com/en-in/products/ai-services/openai-service
[3] https://www.together.ai/

Across all experiments, the temperature is fixed at 0.0 to ensure deterministic outputs and reproducibility. The `max_tokens` parameter is set to 1024 for POS and NER tasks, and to 15 for NLI. The number of exemplars (K) is fixed at 8 for uniform comparison, and selected exemplars are ordered by decreasing similarity to the query within each prompt. The prompt template and verbalizer for each task is described in §B.1.2.
For ILP optimization, we use Python's `gurobipy`[4] package.
Unless otherwise stated, all primary results and analyses correspond to the 0-CLT setting.

4.5 Results and Analysis

Our next goal is to empirically evaluate the proposed SSP framework for cross-lingual transfer to LRLs. Our experiments are designed around a set of key research questions (RQs) that systematically probe different aspects of SSP's effectiveness and robustness.

Research Questions. We evaluate SSP through the following research questions:

1. **RQ1 (0-CLT Efficacy):** Does SSP outperform fine-tuning and prompting baselines under 0-CLT? (§4.5.1)

2. **RQ2 (Ablations):** What is the contribution of each ILP component – confidence filtering, label coverage, and their combination? (§4.5.2)

3. **RQ3 (Robustness beyond 0-CLT):** Is SSP robust under alternative supervision regimes (T-CLT, U-CLT)? (§4.5.3)

4. **RQ4 (Qualitative Analysis):** How does ILP-based selection compare to *TopK* and *MMR* baselines in exemplar quality and downstream accuracy? (§4.5.4)

5. **RQ5 (Scalability):** How does SSP scale – with candidate pool size, exemplar count (K), and base LLM parameter size – relative to heuristic retrieval? Is SSP effective and computationally tractable with smaller exemplar budgets and smaller model sizes? (§4.5.5, §4.5.6)

4.5.1 Key Results (0-CLT)

We first address RQ1 under the strict 0-CLT setting. As detailed in Table 4.1, SSP consistently outperforms both its Stage 1 fine-tuned initializers and standard LLM prompting baselines. For African NER, SSP achieves gains of up to +19 F1 points over fine-tuned baselines and +4 F1 points over LLM prompting. These improvements are even more pronounced for Germanic POS tagging, reaching +26 and +4 F1 points, respectively.

[4] https://pypi.org/project/gurobipy/

Setting	Model	African NER						Germanic POS			
		Hau	Ibo	Kin	Lug	Luo	Avg.	Fo	Got	Gsw	Avg
0-CLT											
Stage 1 w. ICL	*TopK* (LLaMA-2)	64.3	61.2	59.2	60.1	47.3	58.4	79.1	36.0	71.8	62.3
	+ SSP	57.6	62.6	56.0	57.6	43.1	55.4	78.5	37.9	73.5	63.3
	TopK (GPT-3.5)	54.5	69.2	57.8	63.7	46.4	58.3	81.2	37.9	72.2	63.8
	+ SSP	62.8	68.4	64.0	63.8	47.6	61.3	82.4	63.2	79.4	75.0
	TopK (GPT-4x)	64.7	80.8	64.6	71.0	53.3	66.9	81.3	66.5	82.3	76.7
	+ SSP	67.2	79.6	63.3	74.1	54.4	67.7	81.8	**73.7**	85.4	**80.3**
Stage 1 w. fine-tuned CLT	FFT	49.9	54.9	55.4	56.3	40.2	51.3	77.6	17.8	62.0	52.5
	CPG	48.6	50.4	52.6	54.3	38.6	48.9	77.3	16.9	63.9	52.7
	ZGUL	52.2	56.0	53.7	54.5	44.4	52.2	77.2	21.1	65.0	54.4
	+SSP (LLaMA-2)	68.4	58.0	56.1	54.7	42.3	55.9	79.9	39.9	72.9	64.2
	+*MMR* (GPT-3.5)	56.2	69.4	64.7	69.4	39.5	59.8	81.7	30.2	70.0	60.6
	+SSP (GPT-3.5)	61.1	68.9	62.1	67.1	51.4	62.1	**82.8**	67.5	77.0	75.8
	+*MMR* (GPT-4x)	70.7	81.3	**72.0**	**76.9**	52.1	70.6	84.0	63.5	83.1	76.9
	+SSP (GPT-4x)	71.2	**82.4**	71.4	75.4	**55.1**	**71.1**	82.2	71.5	**85.6**	79.8
T-CLT											
Translate-train	ZGUL	72.5	68.5	67.9	65.5	47.3	64.3	–	–	–	–
	+SSP	75.1	76.7	72.3	79.9	54.4	71.7	–	–	–	–
	TopK (GPT-4x)	68.7	78.1	58.7	76.3	53.8	67.1	–	–	–	–
	+SSP	69.9	79.8	60.6	74.7	53.8	67.8	–	–	–	–
Translate-test	Self-Fusion(GPT-4x)	68.4	68.0	58.8	66.5	39.7	60.3	83.0	–	70.0	–
	+SSP	70.0	78.6	64.6	77.0	51.3	68.3	83.7	–	83.9	–
U-CLT											
	AfriBERTa	75.4	79.1	64.9	54.7	39.3	62.7	–	–	–	–
	ZGUL-U	78.5	68.9	62.5	66.0	50.2	65.2	81.5	18.7	80.4	60.2
	+SSP	75.6	84.7	70.3	75.4	54.6	72.1	83.9	71.7	86.0	80.5

Table 4.1: **SSP results**: micro-F1 for *African NER* (left) and *Germanic POS* (right) across CLT settings. **Bold**: best 0-CLT result; underline: overall best. Translate-train omitted for POS (due to unreliable projection); translate-test uses GPT-4 (Gothic unsupported by NLLB-200). Bold Avg. column p-values: 0.008 (NER), 0.0004 (POS).

Among all configurations, SSP(ZGUL)-GPT-4x emerges as the strongest variant, yielding an additional +3 F1 points over other SSP setups. It dominates in two of the three language groups (Tables 4.1 and B.2). Its slightly weaker performance on Germanic POS stems from poor Stage 1 predictions for Gothic (21 F1), where excessive pseudo-label noise bottlenecked the downstream benefits of SSP.

Utilizing ZGUL as the Stage 1 initializer offers two distinct advantages. First, as empirically validated by the Expected Calibration Error (ECE) analysis in §3.5.4, ZGUL provides highly calibrated prediction logits. This enables the ILP solver to safely threshold pseudo-labels and select highly reliable, confidence-aware exemplars without propagating overconfident errors. Second, it is highly cost-efficient; Stage 1 pseudo-label generation requires no additional LLM inference, significantly reducing overall computational overhead.

Furthermore, SSP consistently surpasses standard similarity-based (*TopK*) and diversity-based (*MMR*) retrieval strategies for both GPT-3.5 and GPT-4x. While LLaMA-2 shows strong *TopK* baseline performance on African NER, the joint optimization of semantic relevance, label cov-

erage, and model confidence in SSP proves vastly superior when paired with the reasoning capabilities of the larger LLMs (analyzed further in §4.5.4).

While these 0-CLT results validate SSP's strong empirical performance, understanding the exact mechanisms driving these gains requires deeper investigation. In the following sections, we address RQs 2–4 by isolating the contributions of SSP's core components, analyzing exemplar quality, and evaluating robustness across alternative transfer regimes.

4.5.2 Ablation Analysis

To quantify the impact of individual constraints in the ILP formulation, we perform an ablation study using the best-performing configuration, SSP(ZGUL)-GPT-4x. As shown in Table 4.2, removing these constraints generally leads to performance degradation, indicating that the full ILP formulation provides optimal, complementary benefits for exemplar selection.

Model	Hau	Ibo	Kin	Lug	Luo	Avg.	Fo	Got	Gsw	Avg
SSP(ZGUL)-GPT-4x (Original)	71.2	**82.4**	**71.4**	**75.4**	**55.1**	**71.1**	82.2	71.5	**85.6**	**79.8**
w/o Conf. thresholding	**71.3**	81.9	69.2	74.6	52.7	69.9	**82.8**	57	81.4	73.7
w/o Label Coverage	71.1	79.8	**71.4**	**75.4**	**55.1**	70.6	82.2	**71.6**	**85.6**	**79.8**
w/o both (similarity-only)	70.3	81.8	68	74.8	51.9	69.4	82.4	55.8	82.3	73.5
w/o ILP (Random)	64.1	77.6	61.5	66.1	46.6	63.2	80.6	54.8	80.9	72.1

Table 4.2: **Ablation of ILP components** in SSP(ZGUL)-GPT-4x on *African NER* (left) and *Germanic POS* (right). **Bold**: Best per-language results.

- **Confidence Thresholding**: Removing the confidence threshold constraint degrades performance, particularly for fine-grained token-level tasks such as POS tagging. Without it, low-confidence pseudo-labels introduce noise into the prompt, amplifying error propagation during in-context learning. For Germanic POS, this omission results in an average F1 drop of ~ 6 points.

- **Label Coverage**: Eliminating the label coverage constraint reduces class diversity in the selected exemplars, increasing the risk of label imbalance in the prompt. This effect is particularly harmful for sentence-level classification tasks such as NLI (§B.2), where missing an underrepresented label like "Neutral" severely harms performance. For NER and POS, however, the drop is negligible. As eight sentence-level exemplars inherently comprise roughly 90 to 170 words, most target labels are naturally covered by the baseline retrieval, rendering the explicit coverage constraint redundant for dense token-level tasks.

- **Necessity of Full ILP:** Removing both constraints (similarity-only retrieval) or discarding the ILP entirely (random selection) yields the most severe performance degradation (~ 8 F1 points). This confirms that optimizing for semantic relevance while enforcing structured, label-aware constraints is strictly necessary for robust exemplar selection.

4.5.3 Robustness under T-CLT and U-CLT Settings

SSP consistently delivers gains across alternative transfer regimes (Table 4.1), demonstrating robustness beyond the primary 0-CLT setting. Across all configurations, Stage 2 prompting improves upon its Stage 1 initializer, amplifying rather than distorting the underlying supervision quality.

Stage 1 performance follows a resource-driven hierarchy – translate-test < 0-CLT < translate-train < U-CLT – an ordering that is more or less preserved in Stage 2 results as well. For African NER, SSP consistently elevates performance across all regimes, yielding an average gain of 8 F1 points in translate-test, and 7 F1 points in both translate-train and U-CLT settings.

SSP is particularly resilient to translation noise. In translate-test settings, where noisy machine translations often degrade Stage 1 results, Stage 2 prompting consistently recovers accuracy, yielding gains of up to 12 F1 points for Luo and 11 points for Ibo. This behavior extends to the NLI task as well, where SSP improves F1 in Guarani (Gn) by 12 points despite highly noisy initial pseudo-labels (depicted in Figure B.1).

The strongest overall performance is observed in the U-CLT setting, where higher-quality pseudo-labels from unsupervised adaptation further improve exemplar selection. In this regime, SSP achieves its highest overall performance, including a massive 20-point F1 gain on Germanic POS, while also reducing cross-language variance. These results indicate that SSP scales favorably with improvements in Stage 1 supervision, operating as a general-purpose framework that complements a wide range of cross-lingual transfer pipelines.

4.5.4 Exemplar Quality and Selection Behavior

We evaluate the quality of exemplars selected by SSP against Top-K similarity-based retrieval and MMR-based diversity-aware retrieval. Table 4.3 reports results averaged across all queries and languages for NER and POS tagging. Exemplar sets are evaluated using four metrics:

1. **Query relevance:** Mean semantic similarity between the query and its selected exemplars.

2. **Sentence diversity:** Redundancy among selected exemplars, defined as:

$$\text{SentenceDiversity} = 1 - \frac{1}{K(K-1)/2} \sum_{i \neq j} \text{sim}(x_i, x_j)$$

   where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity. Higher values indicate less redundancy.

3. **Label diversity:** Fraction of distinct task labels covered by the selected exemplars.

4. **Label correctness:** Fraction of pseudo-labels that match the gold labels.

As expected, *TopK* achieves the highest query relevance but suffers from high redundancy and low label diversity. While *MMR* improves sentence-level diversity, its lack of explicit label awareness prevents consistent gains in pseudo-label accuracy.

Query (Gothic POS)

Gothic: hʋa allis ubilis gatawida
English: What of all evil has he done?
Gold POS: hʋa/**PRON** allis/**ADV** ubilis/**ADJ** gatawida/**VERB**

Method	hʋa	allis	ubilis	gatawida	Accuracy
Gold	PRON	ADV	ADJ	VERB	—
SSP (ours)	**PRON**	**DET**	**ADJ**	**VERB**	**3/4 (75%)**
TopK	**PUNCT**	**ADJ**	**ADJ**	**NOUN**	1/4 (25%)
MMR	**PUNCT**	**DET**	**ADJ**	**NOUN**	1/4 (25%)

***TopK* Exemplars** | Label coverage: **2** {NOUN, PUNCT} | Label correctness: **1/11 (9%)**

E1: **hʋa**$_{\text{PUNCT}\to\text{PRON}}$ **gatawides**$_{\text{NOUN}\to\text{VERB}}$ *(Eng: What have you done?)*
E2: **hʋa**$_{\text{PUNCT}\to\text{PRON}}$ **qiþau**$_{\text{NOUN}\to\text{VERB}}$ **izwis**$_{\text{NOUN}\to\text{PRON}}$ *(Eng: What do I say to you ?)*
E3: **hʋaiwa**$_{\text{PUNCT}\to\text{ADV}}$ **ussiggwis**$_{\text{NOUN}\to\text{VERB}}$ *(Eng: How shall we escape?)*
E4: **gasaihʋam**$_{\text{PUNCT}\to\text{VERB}}$ **wulþaga**$_{\text{NOUN}\to\text{ADJ}}$ **himma**$_{\text{NOUN}\to\text{DET}}$ **daga**$_{\text{NOUN}}$ *(Eng: We see glorious things today.)*

SSP Exemplars (ours) | Label coverage: **7** {ADJ, ADP, ADV, DET, INTJ, NOUN, VERB} | Label correctness: **9/28 (32%)**

E1: **galauk**$_{\text{VERB}}$ **auk**$_{\text{ADP}\to\text{ADV}}$ **guþ**$_{\text{NOUN}}$ **allans**$_{\text{DET}\to\text{ADJ}}$ **in**$_{\text{ADP}}$ **un-galaubeinai**$_{\text{NOUN}}$ **ei**$_{\text{ADV}\to\text{SCONJ}}$ **allans**$_{\text{DET}\to\text{ADJ}}$ **gaarmai**$_{\text{NOUN}\to\text{VERB}}$ *(Eng: For God has consigned all to disobedience, that he may have mercy on all.)*
E2: **jar**$_{\text{NOUN}\to\text{CCONJ}}$ **ragin**$_{\text{ADJ}\to\text{NOUN}}$ **in**$_{\text{ADP}}$ **þamma**$_{\text{DET}\to\text{ADJ}}$ **giba**$_{\text{NOUN}\to\text{VERB}}$ *(Eng: And in this I give my advice.)*
E3: **diakaunjus**$_{\text{NOUN}}$ **sijaina**$_{\text{NOUN}\to\text{AUX}}$ **ainaizos**$_{\text{NOUN}\to\text{NUM}}$ **qenais**$_{\text{NOUN}}$ **abans**$_{\text{ADV}\to\text{NOUN}}$ **barnam**$_{\text{VERB}\to\text{NOUN}}$ **waila**$_{\text{NOUN}\to\text{ADV}}$ **fauragaggan-dans**$_{\text{NOUN}\to\text{VERB}}$ **jah**$_{\text{INTJ}\to\text{CCONJ}}$ **seinaim**$_{\text{ADJ}}$ **gardim**$_{\text{NOUN}}$ *Eng: Let deacons be the husbands of one wife, managing their children and their own households well.*
E4: **laistei**$_{\text{ADJ}\to\text{VERB}}$ **afar**$_{\text{NOUN}\to\text{ADP}}$ **mis**$_{\text{DET}\to\text{PRON}}$ *Eng: Follow after me.*

Figure 4.3: **Label diversity and correctness in SSP's exemplar selection** (Gothic POS). *TopK* retrieves redundant exemplars starting with "hʋa" (*What* in English) – sentences containing only 2 POS tags, collapsing the LLM's prediction space. SSP's ILP enforces label coverage across 7 tags, enabling correct predictions for PRON and VERB – labels that were absent from *TopK*'s demonstrations. Green: Correct label; Red: Incorrect label.

Method	Query Relevance	Sentence Diversity	Label Diversity	Label Correctness
TopK	**87.4 / 86.7**	12.4 / 13.0	55.0 / 74.3	51.0 / 47.7
MMR	85.6 / 85.3	**17.2** / 17.0	70.2 / 78.3	49.9 / 54.8
SSP	84.0 / 83.3	14.8 / **17.3**	**84.2 / 81.0**	**52.4 / 62.3**

Table 4.3: **Exemplar quality** of SSP vs. the baselines across four metrics. All scores reported in the format African NER / Germanic POS.

In contrast, SSP strikes an optimal balance by leveraging the calibrated confidence of a high-precision Stage 1 model (as established in §3.5.4) and enforcing global label constraints via ILP. Strikingly, optimizing for label diversity also induces higher sentence diversity – SSP achieves the highest sentence diversity on Germanic POS (17.3), surpassing even the diversity-focused *MMR* baseline – suggesting that covering distinct label types naturally draws from semantically distinct regions of the exemplar pool. Consequently, SSP consistently selects exemplars that are substantially more label-diverse and accurate across both tasks.

This ability to reliably favor high-confidence, non-redundant, and representative exemplars directly explains the downstream performance gains observed in §4.5.1, a behavior qualitatively illustrated for Gothic POS tagging in Figure 4.3.

4.5.5 Scalability of SSP with Candidate Pool Size

We study how the size of the candidate pool used for ILP-based exemplar selection affects the downstream performance of SSP(ZGUL)-GPT-4x. For each language group, we vary the pool size by sampling subsets of the pseudo-labeled test set ({8, 32, 64, 100}), while evaluating on the full set of 100 query instances to ensure comparability. Figure 4.4 reports average F1 scores for African NER and Germanic POS.

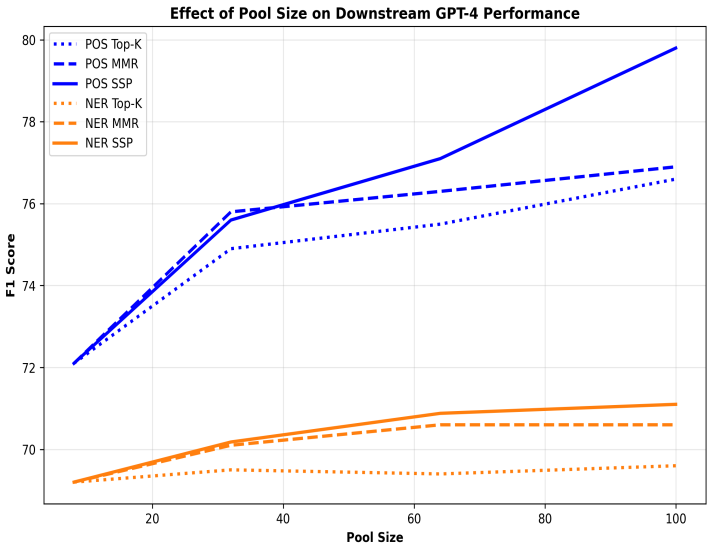


Figure 4.4: **Effect of candidate pool size** on GPT-4 F1 for *African NER* and *Germanic POS* under different retrieval strategies.

Across both tasks, SSP consistently outperforms *TopK* and *MMR* at all pool sizes. However, the scaling behavior differs by task. For Germanic POS, SSP exhibits near-monotonic improvements as the pool size increases, indicating that the ILP formulation continues to benefit from additional candidate exemplars. In contrast, African NER shows early saturation: performance improves rapidly up to a pool size of 64, after which gains taper off. Importantly, even in this regime, SSP maintains a clear margin over both *TopK* and *MMR*, whose performance plateaus earlier and at lower levels. These trends suggest that SSP exploits larger candidate pools more effectively than heuristic retrieval methods, particularly for fine-grained linguistic tasks such as POS tagging. Detailed language-wise results are provided in Table B.4 (Appendix).

Pool Size	**Avg. *TopK* Time**	**Avg. SSP Time**	**Avg. *MMR* Time**
8	2	11	5
32	6	42	22
64	13	77	45
100 (Full)	19	118	68

Table 4.4: **Selection efficiency**: average runtime (ms/query) across retrieval strategies.

We additionally analyze the runtime scaling of SSP relative to *TopK* and *MMR* methods. As shown in Table 4.4, the ILP solver scales approximately linearly with the candidate pool size within this tested range (8–100; $R^2 = 0.999$ for a linear fit), yielding predictable runtime behavior. Although ILP is consistently slower than heuristic methods (such as *TopK*), the absolute latency remains modest – approximately ~0.1s per query even at the full candidate pool size. Since LLM inference dominates end-to-end latency, this optimization cost is unlikely to pose a practical bottleneck.

Computational Complexity. Our exemplar selection formulation is an Integer Linear Program, which is NP-hard in the worst case. In practice, modern MILP solvers (branch-and-bound with cutting planes) resolve instances of this size – a small candidate pool (≤ 100) and small selection budget ($K \leq 8$) – essentially instantaneously, which is why we observe well-behaved, near-linear empirical scaling despite the formulation's worst-case hardness. This distinction matters because near-linear scaling observed here is not guaranteed to persist at substantially larger pool sizes, where solver behavior can degrade once problem size or constraint density crosses certain thresholds.

Scaling to Large-Scale Deployment. Our empirical validation covers pool sizes only up to 100, which is modest relative to realistic deployment settings where the candidate exemplar pool could span a full high-resource-language corpus (thousands to millions of sentences). We have not empirically tested ILP solve time at this scale. For such deployment settings, we propose as future work a two-stage retrieval pipeline: a cheap, embedding-based first-pass filter narrows the

full candidate pool down to a bounded shortlist (e.g., a few hundred candidates via approximate nearest-neighbor search), after which the ILP solver is applied only within that shortlist. This would bound SSP's worst-case cost regardless of underlying corpus size, at the cost of restricting the ILP's candidate set to those already deemed plausible by the cheaper first-pass filter. We leave empirical validation of this pipeline, along with direct ILP solver timing at larger pool sizes, to future work.

Overall, SSP achieves substantially higher-quality exemplar selection while maintaining manageable computational overhead, within the regimes tested, supporting its feasibility for real-world deployment, with a clear path (two-stage retrieval) for extending this to larger-scale settings.

4.5.6 Scalability with Exemplar Count and Model Size

Beyond candidate pool size (§4.5.5), we examine two further axes of scalability relevant to deployment: the number of retrieved exemplars K, and the parameter size of the base LLM. Both are directly tied to cost — smaller K reduces prompt length and API/inference cost, while smaller open-weight models reduce or eliminate inference cost altogether relative to proprietary LLMs. We therefore ask whether SSP's advantage over heuristic retrieval baselines (*MMR*, *TopK*, Random) persists under these more constrained, cost-sensitive settings, or whether it is an artifact of the larger K and larger LLMs used in our main results.

We evaluate all four exemplar-selection methods at $K \in \{2, 4, 8\}$ using **LLaMA-3.1-8B-Instruct**, a small, open-weight model, on both African NER and Germanic POS. Figures 4.5 and 4.6 report the resulting micro-F1 scores.

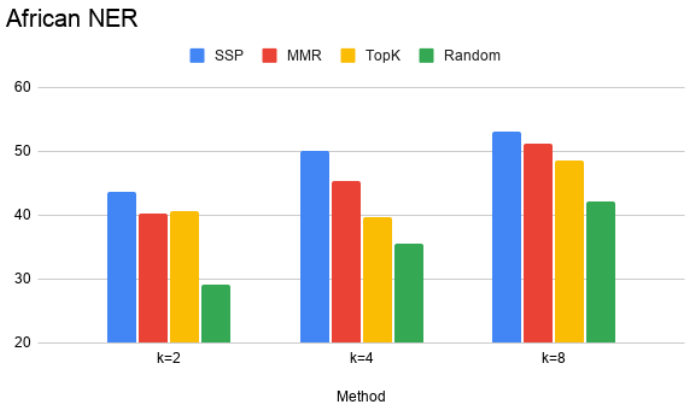


Figure 4.5: **Effect of K on African NER** with LLaMA-3.1-8B-Instruct. SSP outperforms *MMR*, *TopK*, and Random at every K.

SSP remains the strongest method at every K, on both tasks, confirming that its benefit over heuristic retrieval is not contingent on a large base LLM: the ranking observed with GPT-4x in our main results holds unchanged with an 8B open-weight model. All methods generally improve as K increases, and SSP retains its lead over every baseline even at K=2, indicating

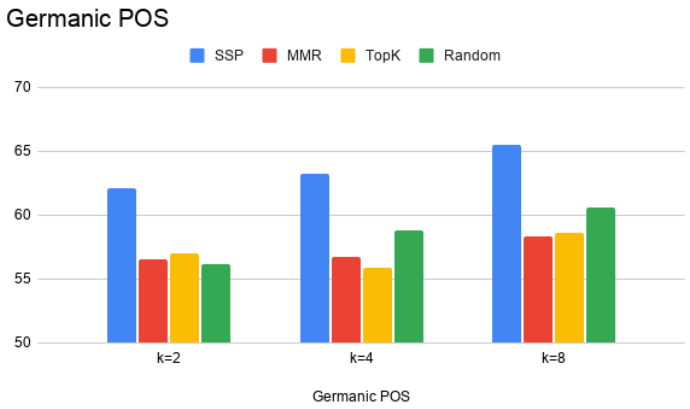


Figure 4.6: **Effect of K on Germanic POS** with LLaMA-3.1-8B-Instruct. SSP outperforms *MMR*, *TopK*, and Random at every K.

that low-K, cost-constrained settings remain usable without a fundamentally different ranking among methods — SSP's advantage does not require a large exemplar budget to manifest.

The effect of K is noticeably more pronounced for NER than for POS, across all four methods, suggesting that entity disambiguation benefits more from additional exemplars than fine-grained POS tagging does, which appears comparatively less sensitive to exemplar count. *TopK* is the one method that does not improve monotonically with K on either task, dipping at K=4 before recovering at K=8 — a non-monotonic pattern not observed for SSP, *MMR*, or Random, and one that further underscores the comparative instability of similarity-only retrieval relative to SSP's ILP-based selection.

4.6 Conclusion

This chapter introduced Self-Supervised Prompting (SSP), a hybrid framework for cross-lingual sequence labeling that enables adaptive exemplar selection without requiring target-language annotations. By leveraging pseudo-labels from existing cross-lingual models and casting exemplar retrieval as an Integer Linear Program (ILP), SSP constructs high-quality, label-aware demonstrations for in-context learning. Extensive experiments across diverse low-resource languages show consistent improvements over both prompting-based and fine-tuned baselines. To encourage reproducibility and further research, the SSP codebase and prompts have been made publicly available.[5]

Beyond these empirical gains, SSP represents a shift from parameter-centric fine-tuning toward inference-time, self-adaptive reasoning, and serves as a conceptual bridge to the hybrid cross-lingual methods developed in the remainder of this thesis. Up to this point, the focus has been on sequence labeling tasks – such as NER and POS tagging – where source-language (HRL/MRL) supervision is manually annotated.

[5] https://github.com/dair-iitd/ssp

The next chapter extends these principles to Distantly Supervised Relation Extraction (DSRE), where even source-language (English) training data is automatically generated and inherently noisy. Addressing this setting lays the foundation for scalable, end-to-end cross-lingual information extraction pipelines that operate with minimal human supervision.

Chapter 5

A Simple and Robust Baseline for Monolingual and Cross-Lingual DSRE

5.1 Introduction

This thesis has so far focused on cross-lingual transfer for sequence labeling. We now turn to *Distantly Supervised Relation Extraction* (DSRE) (Mintz et al., 2009), whose MIML (multi-instance multi-label) formulation and noise characteristics were established in §2.5 (Chapter 2). Figure 5.1 illustrates a typical sentence-level DSRE query: given a text containing an entity pair, the model must predict the underlying semantic relations. In cross-lingual settings, this task becomes particularly challenging when transferring to typologically distant target languages, such as Bhojpuri.

DSRE Query Example:

Entity Pair: Head: **Joe_Biden** (Q6279), Tail: **USA** (Q30)
English: Former **US** President **Joe Biden** is on a visit to India.
Bhojpuri: अमेरिका के राष्टपति रहल जो बाइडेन भारत दौरा पर आइल बाड़न ।
Gold Relations: `President_of, Nationality`

Figure 5.1: **Example sentence-level DSRE query** in English and its Bhojpuri translation for the entity pair (**Joe_Biden**, **USA**).

This chapter identifies a fundamental architectural bottleneck in existing DSRE frameworks: the *independent encoding assumption*. State-of-the-art (SoTA) methods follow the *intra-bag attention* paradigm (Surdeanu et al., 2012; Lin et al., 2016), encoding each sentence in a bag $B = \{s_1, \ldots, s_m\}$ independently before aggregation. This prevents *cross-sentence reasoning* even though the relational evidence is often distributed across multiple sentences within a bag.

We propose PARE (Passage-Attended Relation Extraction), which reformulates each bag as a unified passage and applies relation-aware attention over the joint token sequence. This enables cross-sentence interaction at the encoder level, substantially improving bag-level performance

on English benchmarks, achieving competitive sentence-level results, and enabling effective cross-lingual transfer to four low-resource Indic languages.

5.2 Related Work

This section reviews existing architectures for DSRE, and discusses the limitations of the current state of cross-lingual DSRE.

Intra-bag Attention. The dominant DSRE paradigm – intra-bag attention (Lin et al., 2016) – typically follows a two-step process:

1. **Independent Encoding:** Each sentence $s_i \in B$ is encoded independently via a base encoder (CNN (Zeng et al., 2015), GCN (Vashishth et al., 2018), or BERT (Alt et al., 2019)), producing sentence vectors $\mathbf{h}_i$.

2. **Weighted Aggregation:** For each relation r, a weighted bag representation is formed:

$$\mathbf{b}_r = \sum_i \alpha_i^r \mathbf{h}_i, \quad \text{where } \alpha_i^r = \frac{\exp\left(\mathbf{h}_i^\top \mathbf{r}\right)}{\sum_j \exp\left(\mathbf{h}_j^\top \mathbf{r}\right)}$$

While effective at suppressing noisy sentences, this design restricts the encoder's view to single sentences, preventing the model from leveraging global context across sentences.

Auxiliary Approaches. To further improve robustness, prior works augment intra-bag attention with auxiliary objectives, such as reinforcement learning (Qin et al., 2018) or contrastive learning (Chen et al., 2021) to pull together bags sharing the same relation, or masked language modeling (Alt et al., 2019; Chen et al., 2021) to regularize the encoder. However, these methods do not alter the independent encoding assumption.

The Gap in Multilingual DSRE. As noted in §2.5 (Chapter 2), existing multilingual DSRE benchmarks are limited to high-resource European languages and lack sentence-level annotations. We address this gap by introducing IndicRe4, a human-verified evaluation set for four low-resource Indic languages, in §5.4.

5.3 PARE: A Simple and Strong Supervised DSRE Method

Having motivated Pare's passage-level reformulation in §5.1, we now describe its architecture in detail. Pare implements passage-level relational reasoning through four stages: (a) passage construction, (b) passage encoding, (c) passage summarization, and (d) relation prediction.

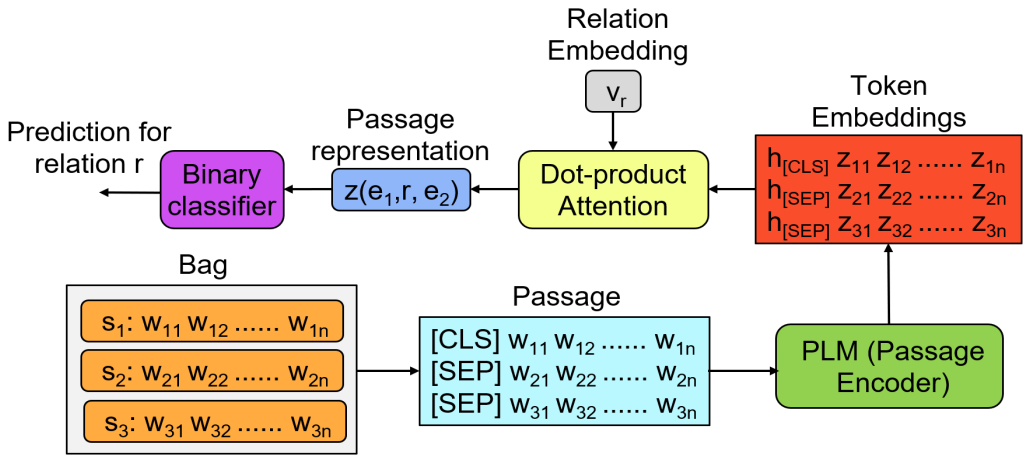


Figure 5.2: **PARE Schematic Architecture**. The model jointly encodes entire passage of bag sentences via a shared PLM encoder, with a token-level attention mechanism pooling evidence across sentences to produce the final bag representation. Entity Markers omitted for brevity.

Passage Construction. For a given bag of sentences $B(e_1, e_2)$, PARE creates a single passage $P(e_1, e_2)$ by concatenating sentences sampled randomly without replacement. This process stops if adding another sentence would exceed the encoder's limit (512 tokens for BERT) or once all sentences are included.

Passage Encoding. The passage $P(e_1, e_2)$ is fed into a pretrained encoder (e.g., BERT) to generate contextualized token representations $\mathbf{z}_i$ for each token w_i. The encoder input begins with a *[CLS]* token, followed by passage sentences separated by *[SEP]*, with remaining positions padded using *[PAD]*. Following earlier practices (Han et al., 2019), each mention of e_1 and e_2 is marked using special entity boundary tokens ⟨e1⟩, ⟨/e1⟩, ⟨e2⟩, and ⟨/e2⟩.

Passage Summarization. For each relation r_j, PARE maintains a learnable query vector v_{r_j}. Relation-specific attention weights α_i^j are computed over passage tokens using dot-product attention (Luong et al., 2015). These weights are used to form a relation-conditioned summary representation:

$$\mathbf{z}_{(e_1, r_j, e_2)} = \sum_{i=1}^{L} \alpha_i^j \mathbf{z}_i,$$

where L is the passage length. This aggregation includes representations of all tokens in the passage, including special tokens and entity markers.

Tuple Classification. Each relation-conditioned summary $\mathbf{z}_{(e_1, r_j, e_2)}$ is passed through a multilayer perceptron (MLP) – shared across relations – with a sigmoid activation to determine the probability p_j that relation r_j exists between the entity pair (e_1, e_2). At inference, the relations with $p_j \geq 0.5$ are predicted as positive. If $p_j < 0.5$ for all relations r_j, the model assigns the *NA* label.

Worked Example. Consider a bag $B(e_1, e_2)$ (Head entity - **Joe_Biden** (Q6279), Tail entity - **USA** (Q30)) with two sentences:

> s_1: "Former **US** President **Joe Biden** is on a visit to India."
> s_2: "**Biden** previously served as Vice President of the **USA**."

(a) Passage Construction: s_1 and s_2 are concatenated into a single passage:
`[CLS] Former <e2> US </e2> President <e1> Joe Biden </e1> is on a visit to India . [SEP] <e1> Biden </e1> previously served as Vice President of the <e2> USA </e2> . [SEP]`
(b) Passage Encoding: The passage is encoded jointly by BERT, producing contextualized token representations $\mathbf{z}_i$ for every token, including entity markers, `[CLS]` and `[SEP]` markers.
(c) Passage Summarization: For relation r_j = `President_of`, the learnable query v_{r_j} computes attention weights α_i^j over *all* passage tokens, so evidence from *both* sentences – despite the entity mentions themselves being split across them – can jointly inform the relation decision, unlike intra-bag methods that encode s_1 and s_2 independently before combining. The weighted sum $\mathbf{z}_{(e_1,r_j,e_2)} = \sum_i \alpha_i^j \mathbf{z}_i$ yields a single relation-conditioned summary.
Note: We illustrate this cross-sentence aggregation concretely, with real model predictions, in Figure 5.4 (§5.6.5): for the entity pair (*Turkey*, *Ankara*) spanning a six-sentence bag, the three gold relations – `/location/contains`, `/country/administrative_divisions`, `/country/capital` – each depend on evidence scattered across different sentences. PARE recovers all three correctly, while CIL, restricted to independent per-sentence encoding, captures only the dominant `contains` pattern and misses the other two.
(d) Tuple Classification: $\mathbf{z}_{(e_1,r_j,e_2)}$ is passed through the shared MLP with sigmoid activation, yielding `President_of` ≥ 0.5, so the relation is predicted positive. Repeating this for all r_j (e.g., `Nationality` ≥ 0.5 as well, matching Figure 5.1's gold labels) yields the full multi-label prediction for (e_1, e_2).

Loss Function. PARE is trained using standard binary cross-entropy (BCE) loss between predicted and gold bag-level labels. No auxiliary or regularization losses are employed.

5.4 Indic Dataset Curation

To evaluate PARE in a rigorous cross-lingual setting, we introduce a new multilingual evaluation benchmark – INDICRE4– covering four diverse Indic languages: Oriya (Ory), Santali (Sat), Manipuri (Mni), and Tulu (Tcy). These languages were selected for their typological diversity and low resource status (~3K-20K Wikipedia articles), despite having millions of native speakers (Table 5.1).

Language	Script & Family	Wiki Articles	Total Speakers	Official Status (India)
Manipuri	Meitei; Tibeto-Burman	10k	~2M	Scheduled
Santali	Ol Chiki; Austro-Asiatic (Munda)	14k	~7M	Scheduled
Oriya	Oriya; Indo-Aryan	20k	~35-40M	Scheduled
Tulu	Kannada; Dravidian	3k	~2M	Non-Scheduled

Table 5.1: **Indic test language characteristics**: typological diversity, resource availability, and official status in India.

Evaluation Set Curation via Constrained Translation. We curate a high-quality evaluation set by translating a stratified subset of the English NYT-10m test split (Gao et al., 2021), comprising 538 sentences and 722 relation labels with balanced ontology coverage, into each Indic language. Translation is performed using *CODEC* (constrained decoding) (Le et al., 2024) with *IndicTrans2* (Gala et al., 2023) as the base translation model. For Tulu, which is not supported by IndicTrans2, we use the Google Translate API.

Prior to constrained decoding, we fine-tune IndicTrans2 on synthetic parallel data augmented with explicit entity projection markers (following Chen et al. (2023)). This adaptation enables CODEC to jointly perform translation and entity-span projection while explicitly preserving head and tail entity boundaries in the target language, without which entity misalignment would invalidate relation judgments. Figure 5.3 illustrates this risk.

Example:

English: "**Barack Obama** took the oath as 44th President of **USA** in Washington D.C."
Gold Relation: `President_of`
Naive Hindi: "बराक ओबामा ने वाशिंगटन डी.सी. में अमेरिका के 44वें राष्ट्रपति के रूप में शपथ ली ।"
Error: Tail marker shifts from अमेरिका (USA) to वाशिंगटन डी.सी. (Washington D.C.) due to tail entity misalignment, yielding the pair (**Obama**, **Washington D.C.**) → model predicts **NA**.

Figure 5.3: **Entity misalignment under erroneous projection**. The tail marker attaches to वाशिंगटन डी.सी. (Washington D.C.) instead of अमेरिका (USA), invalidating the relation judgment.

Human Verification. We employ two native speakers per language. The first corrects all translated evaluation sentences across the full evaluation set, to ensure fluency and correct entity alignment. A second speaker independently reviews a randomly sampled 100 sentences per language, judging whether each correction is acceptable (Yes/No). This confirms reliability: over 70% of system translations required no correction, and corrected sentences achieved an average character-level F1 of 93% against the system output. We report the second speaker's acceptance rate, which exceeded 90% for both sentence translation and entity projection (Table C.3). We report this as percent agreement rather than a chance-corrected coefficient (Cohen's Kappa or Krippendorff's Alpha (Cohen, 1960; Krippendorff, 2018)), since this is a sequential quality-control check rather than an independent double-annotation design, and our binary judgments

are highly skewed toward one category, a regime in which chance-corrected coefficients are typically known to behave unstably (Feinstein and Cicchetti, 1990; Byrt et al., 1993).

Scalable Data Generation for Transfer Learning. For the large-scale generation of training data (the *translate-train* setting), we use EasyProject (Chen et al., 2023). Unlike the computationally intensive CODEC, EasyProject offers a lightweight pipeline for joint translation and entity projection, making it scalable for processing the full DSRE training corpus. To summarize,

- **Training Data (X_{train})**: Generated via EasyProject to maximize throughput.
- **Evaluation Data (X_{eval})**: Curated via CODEC and human verification to maximize precision.

Detailed dataset statistics and annotation guidelines are provided in §C.1.

With the evaluation benchmark established, we now describe the experimental setup used to evaluate PARE across English, multilingual, and cross-lingual settings.

5.5 Experimental Setup

Evaluation Benchmarks. We evaluate across three standard English DSRE benchmarks:

- **NYT-10d (Riedel et al., 2010):** The canonical Freebase-NYT alignment, whose test set retains distant supervision noise.
- **NYT-10m (Gao et al., 2021):** A refined version of NYT-10d with merged relations (50 → 24 relations) and a manually annotated test set, enabling reliable sentence-level evaluation.
- **Wiki-20m (Gao et al., 2021):** A Wikipedia-based dataset aligned with the Wiki80 ontology, featuring a manually annotated test set across 80 relations, providing a larger relational space for evaluating generalization.

Evaluation Settings. Our primary analysis targets bag-level performance, directly testing the passage-encoding hypothesis. We additionally report sentence-level results on NYT-10m and Wiki-20m, multilingual bag-level evaluation on DiS-ReX, and cross-lingual sentence-level evaluation on INDICRE4.

We additionally evaluate on IndoRE (Nag et al., 2021) – a supervised, sentence-level benchmark (both training and evaluation) covering English, Hindi, Bengali, and Telugu – reporting English sentence-level performance and cross-lingual transfer to Hindi, Bengali, and Telugu across 0-CLT, Translate-train, and Translate-test settings.

Cross-Lingual Adaptation. Let X_{train} denote the DSRE training corpus translated into a target language X. In the *translate-train* setting, mBERT undergoes continued pretraining on X_{train} sentences, followed by PARE (or CIL) fine-tuning. In the *translate-test* setting, target-language queries are translated to English using CODEC with IndicTrans2 as the underlying model, and the English-trained PARE model is applied directly.

Evaluation Metrics. On NYT-10d, we report AUC and P@M following Chen et al. (2021); on DiS-ReX, AUC plus micro and macro-F1 following Bhartiya et al. (2022). For manually annotated evaluations (English and Indic), we report micro and macro-F1 only, enabling direct comparison with LLM-based approaches in Chapter 6, where confidence scores are unavailable.

Baselines. We compare against the following supervised DSRE architectures, each employing intra-bag attention aggregation.

- **PCNN-Att (Lin et al., 2016):** Encodes bag sentences using a piecewise convolution neural network (PCNN).
- **RESIDE (Vashishth et al., 2018):** Augments sentence encoding via a Graph Neural Network (GNN) with side information (entity types, relation aliases) sourced from Wikidata.
- **DISTRE (Alt et al., 2019):** Uses GPT 2 (Radford et al., 2019) to encode sentences along with a masked language modeling (MLM) training loss to regularize the encoder.
- **REDSandT (Christou and Tsoumakas, 2021):** Uses a BERT-based encoder along with dependency subtree features.
- **CIL (Chen et al., 2021):** Augments intra-bag attention with contrastive instance learning, where positive pairs are constructed via TF-IDF-based word substitution on each sentence, and negative pairs are formed using bag-level representations of other relations; an auxiliary MLM objective is also employed to regularize the encoder.
- **TransRel (Nag et al., 2021, 2023):** Uses a multilingual BERT (mBERT) encoder with entity-span marker tokens to produce a sentence representation by pooling the `[CLS]`, head entity span, and tail entity span embeddings (CLS+ES), passed through a shared multi-label classifier. As adapted to bag-level DSRE here, TransRel differs from PARE in exactly two respects – attending only over entity-marker tokens rather than the full passage, and using a shared classification layer rather than per-relation learnable query vectors – effectively making it an ablation of PARE's two central architectural choices.
- **BERT+X (Gao et al., 2021):**[1] The intra-bag aggregation variants, in which *Att* refers to standard intra-bag attention, *Avg* means uniform averaging over all sentence representations in bag, and *One* refers to max-pooling over relation-specific sentence scores.

[1] X ∈ {Att, Avg, One}

- **HiCLRE & HFMRE (Li et al., 2022, 2023):** HiCLRE introduces hierarchical contrastive learning, and HFMRE employs Huffman-tree structures to denoise bags.
- **mBERT+MNRE (Bhartiya et al., 2022):** Divides a multilingual bag into individual language sub-bags and performs language-conditioned relation attention over each sub-bag's sentences. Sub-bag representations are then averaged for final prediction.

Implementation Details. For English and multilingual experiments, we use the `bert-base-uncased` and `bert-base-multilingual-cased` encoders, respectively. The framework is implemented using the OpenNRE (Han et al., 2019) codebase. Models are optimized with AdamW using a linear learning-rate schedule and weight decay of 1×10^{-5}. Hyperparameters are selected via grid search over learning rates $\{1,2\} \times 10^{-5}$ and batch sizes $\{16,32,64\}$. All experiments are performed on a single NVIDIA GeForce GTX 1080 Ti GPU, with PARE typically converging within 2–4 epochs.

5.6 Results

We analyze PARE's performance across bag-level, sentence-level, and cross-lingual evaluations.

5.6.1 Bag-level Evaluation

Bag-level results on the NYT-10d benchmark, alongside the manually annotated NYT-10m and Wiki-20m test sets, demonstrate PARE's superior ability to aggregate noisy evidence from multiple sentences.

On NYT-10d (Table 5.2), PARE achieves a 2.6-point improvement in AUC over CIL the previous SoTA, and outperforms *REDSandT* by more than 11 AUC points. Notably, while both PARE and *REDSandT* utilize the same BERT encoder, *REDSandT* requires additional external information such as entity types and dependency subtrees. This performance gap underscores the inherent advantage of passage-level encoding over traditional sentence-level aggregation for capturing complex relational signals.

On the manually annotated benchmarks (NYT-10m and Wiki-20m), PARE consistently outperforms all baseline models. As detailed in Table 5.3, it improves upon CIL by 4.2 micro-F1 points on NYT-10m and substantially surpasses the pooling-based variants (*Att, Avg, One*) originally reported in prior work (Gao et al., 2021).

Detailed analyses of PARE's attention interpretability, bag-length robustness, and model ablations are provided in §5.6.5.

Multilingual Results on DisRex. To evaluate on multilingual bags, we apply the PARE architecture on top of the mBERT encoder (mPARE). As shown in Table 5.2, mPARE outperforms

the prior SoTA (mBERT+MNRE (Bhartiya et al., 2022)) by +4.9 AUC on DiS-ReX, demonstrating that passage-level aggregation remains effective in multilingual DSRE settings.

NYT-10d (Bag-level)			**DiS-ReX (Multilingual Bags)**		
Model	**AUC**	**P@M**	**Model**	**AUC**	**F1 (Micro/Macro)**
PCNN-Att	34.1	69.4	PCNN+Att	67.8	63.4/43.7
RESIDE	41.5	77.2	mBert+Att	80.6	74.1/69.9
DISTRE	42.2	66.8	mBert+One	80.9	74.0/68.9
REDSandT	42.4	75.3	mBert+Avg	82.4	75.3/71.0
CIL	50.8	86.0	mBert+MNRE	82.1	76.1/72.7
PARE	**51.8***	**89.0***	***mPARE***	**87.0***	**79.3*****/76.0**

Table 5.2: **Distantly supervised bag-level evaluation** on the English NYT-10d dataset (left) and the multilingual DiS-ReX dataset (right). Best results in **bold**, and the second best ones underlined. * McNemar's p value $< 10^{-5}$.

Model	**Bag-level**		**Sentence-level**		
	NYT-10m	**Wiki-20m**	**NYT-10m**	**Wiki-20m**	IndoRE†
BERT+Att	54.1/25.8	66.8/64.3	24.5/17.4	59.5/52.2	–
BERT+Avg	60.9/33.9	82.7/82.0	36.4/27.2	**83.1**/79.6	–
BERT+One	60.4/35.7	81.6/81.1	26.8/19.3	82.3/78.9	–
CIL	57.3/36.3	83.2/82.6	40.9/**29.9**	81.9/78.9	–
TransRel	56.6/29	82/82	**41.7**/26	81.8/**80.6**	84.4 / 64.2
PARE	**61.5*****/39.3**	**84.6*****/83.8**	40.1/28.8	82.9/79.5	**89.8*** **/ 85.2**

Table 5.3: **English bag-level and sentence-level evaluation** (micro/macro-F1) on NYT-10m, Wiki-20m and IndoRE. Bag-level IndoRE column is omitted as IndoRE is a sentence-level-only benchmark without a multi-instance bag structure. * McNemar's p value $< 10^{-5}$. † Could not run other baselines on IndoRE due to time constraints.

5.6.2 Sentence-level Evaluation

Sentence-level performance (right three columns of Table 5.3) reveals that while PARE remains competitive on Wiki-20m (82.9 F1), it trails TransRel on NYT-10m (40.1 vs. 41.7 F1) and also trails CIL on macro-F1 specifically (28.8 vs. 29.9), though CIL's micro-F1 (40.9) is itself lower than TransRel's. This gap reflects a fundamental trade-off of passage-level modeling: PARE is optimized to consolidate relational evidence distributed across multiple sentences; it is not designed to isolate relational cues from individual sentences, and at the single-sentence level this advantage goes unused – the one setting where TransRel's simpler entity-span pooling, an ablation of PARE's full-passage attention, is not at a disadvantage.

On IndoRE (English), by contrast, PARE substantially outperforms TransRel (89.8/85.2 vs. 84.4/64.2 micro/macro-F1). Absolute scores here are notably higher than on DSRE benchmarks (e.g., PARE's 40.1/28.8 on NYT-10m), reflecting IndoRE's clean, sentence-level training rather than noisy distant supervision. We note IndoRE is a comparatively low-data setting (~6.8K English training examples), where PARE's full-passage attention and per-relation query vectors appear especially beneficial relative to TransRel's coarser CLS+entity-span pooling and shared

Passage (bag of 6 sentences)

s_1:"We have not made miracles , but we have tried to move the dossier forward a little bit , " Javier Solana, left, the foreign policy chief for the EU, said in **Ankara**, **Turkey**, at the end of two days of talks ... [SEP]
s_2: A3 Bomb Kills 6 in **Turkish Capital**: A powerful explosion tore into a shopping district in **Ankara**, **Turkey**, killing six people and wounding more than 70 ... [SEP]
s_3: **Ankara**, the **nerve center** of **Turkey** 's government, is heavily guarded and has rarely been the target of a bombing. [SEP]
s_4: Mr. Avsar, who was born in **Turkey** , was planning to move back to **Ankara** , where he had a teaching position . [SEP]
s_5: It was of no small consequence to Musharraf that he had gone to **Turkey** as a boy of 6 in 1949, when his father was assigned to his country's **embassy** in **Ankara** ... [SEP]
s_6: **Ankara**'s governor, Kemal Onal, identified the bomber as Guven Akkus, from the Sivas district in southeastern **Turkey** ... [SEP]

Predictions

Gold labels: /location/contains, /country/administrative_divisions, /country/capital
PARE: **/location/contains**, **/country/administrative_divisions**, **/country/capital** ✓ **predicts all 3**
CIL: /location/contains **misses 2 out of 3 relations**

Figure 5.4: **Cross-sentence reasoning in PARE.** The three gold relations for (**Turkey**, **Ankara**) require evidence distributed across different sentences. PARE's passage-level encoding recovers all three, while CIL captures only the dominant *contains* pattern.

classification layer – consistent with TransRel functioning as an ablation of these two design choices. IndoRE's bag-level column is omitted in Table 5.3, as IndoRE is a supervised, sentence-level RE benchmark without the multi-instance bag structure of our DSRE benchmarks.

5.6.3 CLT Evaluation

Model	English	CLT settings	
		Translate-train	Translate-test
HiCLRE	31.2/18.4	20.1/13.4	24.3/14.3
HFMRE	32.8/18.3	22.4/13.6	27.3/16
CIL	**43.1***/**31.5**	26.3/18	34.3/**23.5**
TransRel	41.7/26	23.8/15.5	**34.6***/21.7
PARE	41.8/30.8	**29.5***/**19.5**	32.8/23.3

Table 5.4: **Sentence-level cross-lingual evaluation on** INDICRE4 evaluation sets and English evaluation on the corresponding stratified NYT-10m evaluation set (micro/macro-F1). Language-wise results in §D.2.1. * McNemar's p value $< 10^{-5}$.

We evaluate PARE on the newly curated Indic benchmark INDICRE4 across four languages under *Translate-train* and *Translate-test* settings. We omit the 0-CLT setting, as mBERT-based supervised DSRE models fail to generalize to native Indic scripts, yielding near-random performance with an average of less than 5 micro-F1 points.

Model	Hindi	Bengali	Telugu	Avg
TransRel	35.2/25.8	28/18.3	18.7/8.9	27.3/17.7
PARE	41.5/33.1	32.1/24	17.7/11.7	**30.4*/22.9**

(a) Zero-shot CLT (0-CLT)

Model	Translate-Train				Translate-Test			
	Hindi	Bengali	Telugu	Avg	Hindi	Bengali	Telugu	Avg
TransRel	54.6/41.1	49.2/36.4	36.2/26	46.7/34.5	52.1/40.8	50.9/42.1	42.6/27.2	48.5/36.7
PARE	57.9/47.5	54.1/46.1	48.5/33.5	**53.5*/42.4**	52.9/43.4	53/46.5	42.5/31	**49.5/40.3**

(b) Translate-Train and Translate-Test

Table 5.5: **Sentence-level cross-lingual evaluation on IndoRE** for Hindi, Bengali and Telugu. * McNemar's p value $< 10^{-5}$.

Table 5.4 reports sentence-level results on English (NYT-10m) and averaged performance across the Indic languages. On English inputs and under translate-test, CIL marginally outperforms PARE (1.3 micro-F1), consistent with CIL's explicit optimization for sentence-level discrimination. TransRel shows a similar pattern on aggregate Translate-test (34.6*/21.7 vs. 32.8/23.3) while being competitive for English sentence-level extraction (41.7/26.0 vs. 41.8/30.8) – settings where the model reasons over a single (English or translated) sentence rather than passage-level evidence, precisely where TransRel's simpler entity-span pooling is not at a disadvantage.

In contrast, under the translate-train setting, PARE outperforms CIL by 3.2 micro-F1 points. This indicates PARE's higher robustness when inference is performed directly in the target language. We attribute this to CIL's reliance on English-specific tf-idf augmentation and contrastive objectives, which do not transfer robustly across Indic scripts. PARE similarly outperforms TransRel in Translate-train (29.5/19.5 vs. 23.8/15.5, Table 5.4), by its largest margin of any setting in this table (+5.7 micro-F1) – consistent with the pattern observed in the language-wise IndoRE breakdown (Table 5.5), where PARE also outperforms TransRel by its largest margin specifically in Translate-train, and with TransRel's restricted attention and shared classifier acting as an ablation of PARE.

Further, PARE also outperforms TransRel in both the 0-CLT (30.4*/22.9 vs. 27.3/17.7) and Translate-test (49.5/40.3 vs. 48.5/36.7) settings (Table 5.5), the same pattern of TransRel underperforming as an ablation of PARE's full-passage attention and per-relation representations.

Overall, while PARE is not the top-performing model for English sentence-level extraction, it demonstrates superior robustness for native-language (translate-train) inference in cross-lingual settings.

5.6.4 Qualitative & Error Analysis

We conduct a confusion analysis of PARE across manually identified error categories for NYT-10m sentence-level extraction, summarized in Table 5.6, with specific representative examples

Gold	Predicted	English	Manipuri	Santali
place_lived	nationality	10	12	9
place_of_birth	nationality	5	4	5
place_of_death	place_lived	4	14	12
place_of_birth	place_lived	6	8	7
place_lived	place_of_birth	0	2	4
company_advisors	company_founders	6	0	0
non-NA label	NA	286	322	327

Table 5.6: **PARE's frequent relation confusions** on NYT-10m (English) and IndicRe4 (Manipuri and Santali) under translate-train. Below dashed line: false abstention (non-NA → NA), the dominant error mode.

depicted in Table 5.7.

False Abstention. The most frequent error across languages is the prediction of "NA" label for valid relations (e.g., 286 instances in English). This conservative behavior is most prevalent in cases requiring bridging inference – where the model must logically connect indirect mentions (e.g., inferring company/advisors from "chief executive").

Proximity Bias. Confusions frequently occur between semantically related pairs, such as place_lived and nationality. These are frequently driven by *proximity bias*, where local keywords overshadow distant semantic triggers. As shown in Figure C.1a, the local salience of "home" outweighs the distant trigger "died." A similar bias manifests differently in Manipuri (Table 5.7), where the locative postposition ꯗꯤ ("at/in") suffixed to the tail entity (**Greenwich**) creates a surface pattern that mimics a residence context, leading the model to incorrectly predict place_lived instead of place_of_death.

Lexical Ambiguity and Granularity Mismatch. In the business domain, salient keywords can easily mislead the model; for instance, "founder" frequently overshadows structural cues like "partner at," causing confusion between company_advisors and company_founders. Additionally, for location-based labels, Pare exhibits a bias toward coarse-grained, high-frequency relations. It often defaults to *contains_location* instead of nuanced labels like administrative_divisions or /country/capital, even when explicit cues such as ꯔꯥꯡꯇꯤꯜ ("capital") are present.

Success Case: Cross-Sentence Reasoning. The core strength of Pare lies in its ability to resolve relations using evidence distributed across the entire bag – a task where intra-bag models like CIL fail. As an example, for the (**Turkey**, **Ankara**) pair (Fig. 5.4), the three gold relations are signaled by cues in separate sentences: "Turkish Capital" (s_2), "embassy" (s_5), and "governor" (s_6).

Error Type	Lang	Sentence (abridged)	Gold	Predicted
Lexical ambiguity	EN	... one of the most outspoken is **Vinod Khosla**, a founder of Sun Microsystems and now a partner at **Khosla Ventures**.	**advisors**	**founders**
False abstention	EN	"**MySpace** is dedicated to ..." **Chris DeWolfe**, the chief executive of MySpace, said ...	**advisors**	**NA**
Granularity mismatch	EN	Mr. Putin will continue his Middle Eastern trip ... traveling to **Amman**, **Jordan**, where he will meet with King Abdullah II ...	**admin_div**	**contains**
Proximity bias	MNI	...ꯃꯩꯔꯤꯁꯔꯤꯁꯗ ꯃꯦꯔꯊꯨ ꯃꯇꯝ ꯃꯊꯪꯕꯤ ... (*murder of **Martha Moxley** in **Greenwich***)	**place_of_death**	**place_lived**
Granularity mismatch	MNI	...ꯚꯦꯠꯦ ꯂꯥꯔꯤꯑꯦꯖꯥꯔꯤꯌꯦ ... ꯂꯥ ꯄꯥꯡ ꯇꯦ ꯒꯥꯛꯠ ... (*La Paz, the capital of **Baja California***)	**capital, contains**	contains

Table 5.7: **PARE's representative error cases** on English and Manipuri (translate-train). Underlined tokens indicate the key semantic trigger. In Manipuri, the locative postposition ꯗꯤ ("at/in") suffixed to **Greenwich** mimics a residence pattern (row 4), and the explicit cue ꯂꯥꯡ ("capital") is ignored in favor of the high-frequency `contains` relation (row 5). Colors: **Head entity**, **Tail entity**, **Correct**, **Wrong**, **Partially Correct**.

These distributed cues enrich PARE's bag encoding, allowing it to correctly resolve multiple relations simultaneously, whereas CIL remains overwhelmed by the dominant *contains* pattern prevalent in the individual sentences.

5.6.5 Analysis: Efficiency, Ablation, and Attention Behaviour

Parameter Efficiency. PARE is significantly more compact than baselines (Table 5.8). While models like CIL and MNRE add over 2.4 million parameters due to entity projection layer (Soares et al., 2019), PARE adds only 46,082. This is achieved by attending directly to the passage tokens rather than using a projection layer for entity representations.

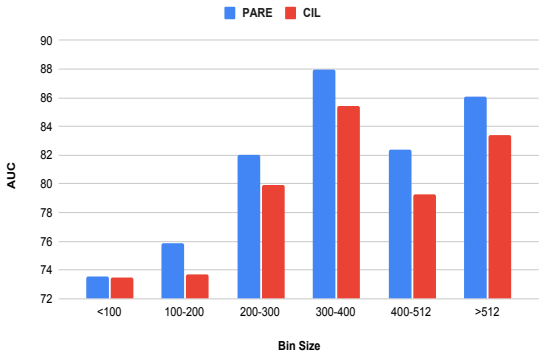
Model	#Parameters (excluding encoder)
Att	2400793
One	2399257
Avg	2399257
CIL	2453052
MNRE	2645029
PARE	46082

Table 5.8: **Comparison of trainable parameters** between PARE and other baselines

Bag Size Robustness. Despite BERT's 512-token limit, PARE maintains a consistent 2-3 AUC advantage over CIL across bag sizes, except for single-sentence bags where cross-sentence attention is unavailable (Figure 5.5).

Ablation. Ablation studies across four datasets (Table 5.9) confirm that entity markers are the most critical component, with their removal causing the largest performance drop (e.g., -36.9

Figure 5.5: **PARE's bag size robustness**: AUC on NYT-10m across bins of untruncated bag length. PARE maintains a consistent advantage over CIL except for single-sentence bags.



Modification	NYT-10d	NYT-10m	Wiki-20m	DiS-ReX
w/o passage summarization	-4.9	-2.9	-4.2	-0.8
w/o [PAD] attention	-3.1	-2.3	-1.9	-0.1
w/o entity markers	-36.9	-16.5	-29.9	-20.5

Table 5.9: **PARE ablation**: change in AUC (Δ AUC) with different components of PARE

AUC on NYT-10d). Passage summarization also provides significant gains by enabling the model to focus on relation-relevant tokens. Notably, disallowing attention over [PAD] tokens leads to a 3.1-point AUC drop, revealing that the model uses padding as an attention "sink" to suppress low-confidence predictions: PARE assigns ~90% of attention to [PAD] tokens for NA instances, compared to only 0.1% for valid relations.

Semantic Trigger Identification. Qualitative analysis (Table 5.10) shows that PARE's relation-specific query vectors attend to pivotal linguistic *triggers* – e.g., "died" for `place_of_death` and "founder" for `company/founders` – confirming that the passage summarization mechanism learns interpretable, relation-discriminative attention patterns.

5.7 Discussion and Conclusion

This chapter presented PARE, a passage-based alternative to intra-bag attention for distantly supervised relation extraction. By jointly encoding all sentences in a bag and applying relation-aware token attention over the unified passage, PARE achieves SoTA performance on standard English and cross-lingual DSRE benchmarks with substantially fewer trainable parameters than prior approaches. We publicly release our code and trained models to facilitate research in the DSRE community.[2]

However, two critical limitations remain. First, PARE is optimized for bag-level supervision and lacks an explicit mechanism for sentence-level relation extraction. As a result, it often con-

[2] https://github.com/dair-iitd/DSRE

Input Passage (BERT-tokenized)	Predicted Relation
[CLS] she will face 11th - seeded je ##lena jan ##kovic of serbia , a 6 - 3 , 6 - 4 winner over <head> victoria az ##are ##nka </head> **of** <tail> belarus </tail> . [SEP]	/people/person/nationality
[CLS] <head> michael sm ##uin </head> , a choreographer who worked for major ballet companies and led his own , marshal ##ing eclectic dance forms , robust athletic ##ism and striking theatrical ##ity to create works that appealed to broad audiences , **died** yesterday in <tail> san francisco </tail> . [SEP]	/people/deceasedperson/place_of_death
[CLS] <tail> steve new ##comb </tail> , a <head> powers ##et </head> **founder** and veteran of several successful start - ups , said his company could become the next google . [SEP]	/business/company/founders

Table 5.10: **PARE's semantic trigger identification**: attention analysis on correctly predicted relations on NYT-10m. **Bold** tokens: highest-attended token per instance.

fuses closely related relations (e.g., *Nationality* vs. *Place_lived*) when fine-grained predictions are required. Second, despite achieving the strongest translate-train performance on IndicRe4, Pare's absolute cross-lingual scores remain substantially below English, indicating limited robustness to translation noise. Finally, the inability to perform zero-resource transfer on unseen scripts highlights the representational limits of mBERT-style encoders, necessitating alternative cross-lingual bridges.

The next chapter introduces HYDRE, a hybrid framework that addresses these gaps by augmenting Pare with LLM-based inference-time reasoning for robust relation extraction in both monolingual and cross-lingual settings.

Chapter 6

Hybrid DSRE Approaches using Large Language Models

6.1 Introduction

In Chapter 5, we introduced PARE, a supervised approach that mitigates bag-level noise in distantly supervised relation extraction (DSRE). However, while PARE achieves strong bag-level performance, it struggles to generalize to sentence-level inference – particularly under cross-lingual transfer to low-resource Indic languages. This reveals a core limitation: although effective at learning robust bag representations, PARE lacks the contextual depth required to resolve fine-grained relational ambiguity within individual sentences.

To address this sentence-level reasoning gap, we naturally turn to the prompting paradigms established in Chapter 4. While SSP demonstrated that Large Language Models (LLMs) excel at cross-lingual extraction when guided by clean, high-quality exemplars, directly extending these techniques to DSRE remains challenging. Distant supervision produces noisy bag-level annotations, whereas ICL typically requires clean, sentence-level exemplars. This fundamental mismatch makes constructing reliable in-context exemplars highly challenging, as standard retrieval is susceptible to propagating false-positive relations into the LLM's context window. Consequently, prior work has relegated LLMs primarily to offline data relabeling or refinement, rather than utilizing them for principled, end-to-end DSRE inference (Zhao et al., 2024).

To overcome these limitations, we introduce HYDRE (**HY**brid **DSRE**), a hybrid inference-time framework that integrates supervised DSRE models with LLM-based reasoning. HYDRE leverages the high recall of a fine-tuned DSRE model, such as PARE, to first filter candidate relations in a query sentence, after which an LLM performs fine-grained disambiguation. Central to this design is a *three-stage exemplar selection pipeline* that extracts reliable, relation-specific in-context exemplars from noisy bags using a joint *confidence-similarity* scoring mechanism. Crucially, HYDRE operates entirely at inference time and requires no retraining of either the LLM or the base DSRE model.

We evaluate HYDRE in both monolingual (English) and cross-lingual transfer settings, focusing on the INDICRE4 benchmark for low-resource Indic languages introduced in Chapter 5. Experimental results demonstrate that HYDRE consistently outperforms prior DSRE baselines across all settings. Furthermore, extensive ablation studies confirm the necessity of our hybrid design, revealing that removing the retrieval or filtering components degrades performance by up to 7 F1 points in English and 12 points in cross-lingual transfer.

6.2 Related Work

We now situate HYDRE within prior literature on LLM-based relation extraction and hybrid inference paradigms.

LLMs for Relation Extraction. Research on LLMs for relation extraction primarily focuses on supervised or few-shot settings with clean, sentence-level annotations. For example, Wadhwa et al. (2023) demonstrate that few-shot prompting with GPT-3 can approach fully supervised RE performance, particularly when augmented with chain-of-thought (CoT) reasoning. Other work explores zero-shot prompting using natural language relation definitions (Zhou et al., 2024), or employs LLMs as auxiliary annotators to denoise distantly supervised data (Jian et al., 2024). However, these approaches either rely on high-quality human supervision or restrict LLMs to offline data refinement, rather than integrating them directly into the DSRE inference pipeline.

Cross-Lingual RE with LLMs. While LLM-based RE has matured on English-centric benchmarks, cross-lingual applications remain underdeveloped. Most multilingual RE research still relies on the supervised fine-tuning of mPLMs such as mBERT and XLM-R (Hennig et al., 2023; Bhartiya et al., 2022). To our knowledge, HYDRE represents the first framework to leverage LLM-based in-context learning for *distantly supervised* cross-lingual RE, explicitly targeting low-resource languages that lack both clean annotated data and extensive unlabeled corpora.

LLM-as-a-Selector. LLMs increasingly serve as evaluators or "judges", demonstrating strong alignment with human preferences in complex reasoning tasks (Zheng et al., 2023). Closely related to our work is the *filter-then-rerank* paradigm introduced by Ma et al. (2023), where a small Language Model (sLM) prunes the candidate space and an LLM reranks ambiguous samples via multiple-choice prompting. While effective for supervised relation extraction on datasets like TACRED (Zhang et al., 2017), this approach assumes clean, sentence-level labels and utilizes naive embedding-based similarity for exemplar retrieval. Such assumptions fail in the DSRE setting, where labels are inherently noisy and supervision is only available at the bag level. HYDRE adapts this collaborative paradigm to the DSRE domain by leveraging PARE as a high-recall filter and introducing a bag-aware exemplar selection strategy explicitly designed to mitigate multi-instance noise.

6.3 HYDRE: The Hybrid DSRE Framework

Although PARE struggles with sentence-level extraction, our preliminary analysis (Table 6.1) shows it maintains high Recall@K: the gold relation appears among the top-K candidate predictions approximately 85% of the time at $K = 5$ on English. This makes PARE well-suited as a high-recall candidate generator, with an LLM acting as a precision-oriented selector over the filtered candidate set. The core challenge lies in constructing reliable in-context exemplars from noisy bags. To solve this, we propose a three-stage exemplar selection pipeline (Fig. 6.1) that extracts clean, representative sentences from distantly supervised data.

Language	PARE	CIL
English	**84**	82
Oriya	**75**	72
Santali	**72**	64
Manipuri	69	69
Tulu	69	**71**
Avg. (Indic)	**71**	69

Table 6.1: **Recall@5** for PARE and CIL on English (NYT-10m) and Indic languages (INDICRE4, translate-train). PARE's higher recall motivates its use as the candidate filter in HYDRE.

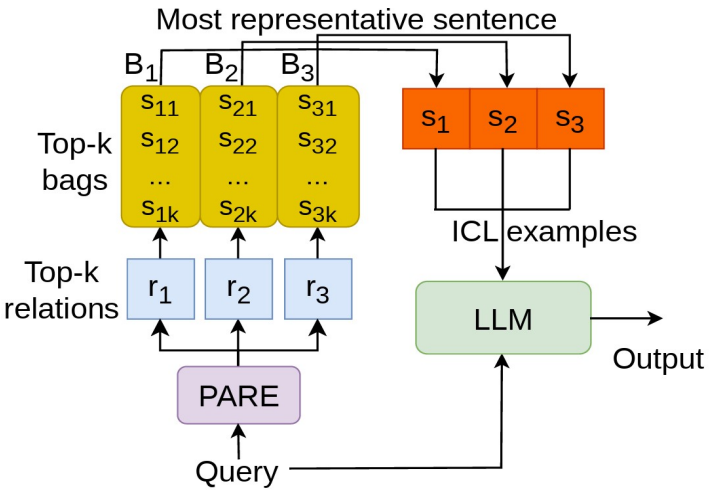


Figure 6.1: **HYDRE pipeline** ($K = 3$). Given a query, PARE produces top-K candidate relations (r_1, r_2, r_3). For each candidate, the most relevant bag (B_1, B_2, B_3) is retrieved from the DSRE corpus. The most representative sentence from each bag is selected as an ICL exemplar and passed to the LLM alongside the query.

Three-stage Exemplar Selection. The exemplar selection pipeline (Fig. 6.1) consists of three sequential stages: (1) *candidate relation selection*, (2) *bag selection*, and (3) *representative sentence selection*.

- **Stage 1 - Candidate Relation Selection:** Given a query sentence q, the supervised DSRE model PARE assigns confidence scores $f_{\text{PARE}}(q, r)$ for every relation r in the ontology. The top-K scoring relations form the candidate set $\mathcal{R}'$.

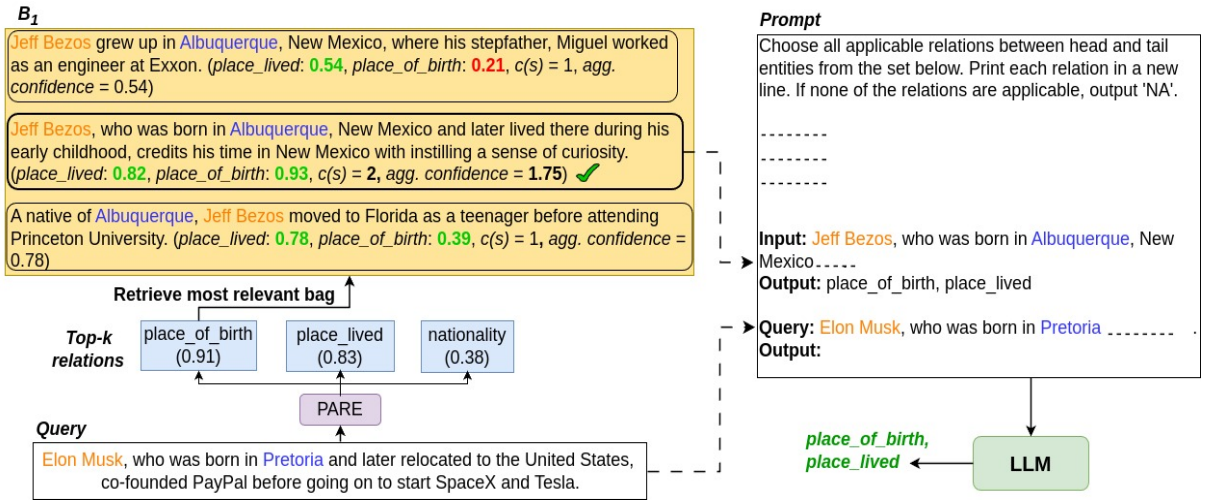


Figure 6.2: **Exemplar selection in HYDRE** (illustrated for the topmost candidate `place_of_birth`). *Left*: from bag B_1, the sentence with the highest aggregate confidence over both bag relations is selected (green tick). *Right*: the selected sentence is formatted as an in-context exemplar in the LLM prompt alongside the query.

- **Stage 2 - Bag Selection:** For each candidate relation $r \in \mathcal{R}'$, we identify the most relevant bag B_r from the subset of training bags $\mathcal{B}_r$ having r as one of its relations. Each bag $B_j \in \mathcal{B}_r$ is scored using a weighted scoring function that balances semantic similarity and model confidence:

$$\text{score}(B_j, r) = (1 - \lambda)\,\text{sim}(q, B_j) + \lambda f_{\text{PARE}}(B_j, r),$$

where $\text{sim}(q, B_j) = \max_{s \in B_j} \text{sim}(q, s)$ measures semantic similarity of bag B_j to the query, and $f_{\text{PARE}}(B_j, r) = \max_{s \in B_j} f_{\text{PARE}}(s, r)$ represents B_j's confidence score for relation r. This favors bags that are semantically aligned with the test query while providing strong evidence for the candidate relation. The trade-off hyperparameter λ controls the balance between semantic similarity and model confidence.

- **Stage 3 - Sentence Selection:** A bag B_r typically contains several sentences, and – since DSRE is multi-label under distant supervision – may itself be labeled with *multiple* relations $\textit{labels}(B_r)$, not just the candidate relation r that selected it in Stage 2. Naively picking an arbitrary sentence, or the one scoring highest on a single relation, risks giving the LLM a narrow exemplar that reflects only one of the bag's relations. We instead want the sentence that best represents the bag's *full* relation set.

  Intuition. For each sentence $s \in B_r$, we ask: how many of the bag's labeled relations does s confidently express on its own? We call this its *coverage*, and prefer the sentence with the highest coverage – even over a sentence that scores highest on any single relation in isolation.

  Worked example. Figure 6.2 illustrates this on a bag B_1 for (*Jeff_Bezos*, *Albuquerque*), labeled with two relations, `place_of_birth` and `place_lived`. The first sentence ("Jeff Bezos grew up in Albuquerque...") clears the confidence threshold τ only on `place_lived`,

Algorithm 1 Exemplar Selection for HYDRE

Input: Query sentence q; bags of sentences $\mathcal{B} = \{B_j\}$; trained DSRE model scores $f_{\text{PARE}}(s, r)$ for sentence s and relation r; semantic similarity function $\text{sim}(q, B_j)$; confidence threshold τ; trade-off parameter λ; number of exemplars K.

Output: Ordered list of selected exemplar sentences.

Compute $f_{\text{PARE}}(q, r)$ for all relations r in the ontology.
$\mathcal{R}' \leftarrow$ top-K relations ranked by $f_{\text{PARE}}(q, r)$ ▷ Candidate relation selection
Initialize exemplar list $\mathcal{E} \leftarrow [\,]$.
for each $r \in \mathcal{R}'$ **do**
    $\mathcal{B}_r \leftarrow \{B_j \in \mathcal{B} \mid r \in \textit{labelset}(B_j)\}$.
    $B_r \leftarrow \arg\max_{B_j \in \mathcal{B}_r} [(1-\lambda)\text{sim}(q, B_j) + \lambda f_{\text{PARE}}(B_j, r)]$. ▷ Select most relevant bag for r
    For each $s \in B_r$, compute $c(s) = \sum_{r_a \in \text{labels}(B_r)} \mathbb{I}[f_{\text{PARE}}(s, r_a) > \tau]$.
    Let $v_{\max} \leftarrow \max_{s \in B_r} c(s)$. ▷ max. no. of relations with confidence > threshold
    $\mathcal{S}' \leftarrow \{s \in B_r \mid c(s) = v_{\max}\}$. ▷ candidate sentences with max. label coverage
    $s^* \leftarrow \arg\max_{s \in \mathcal{S}'} \sum_{r_a \in \text{labels}(B_r)} f_{\text{PARE}}(s, r_a)$. ▷ Pick sentence with highest aggregate confidence
    Add $(s^*, \text{labels}(B_r))$ to $\mathcal{E}$.
end for
Sort $\mathcal{E}$ in ascending order of $f_{\text{PARE}}(q, r)$. ▷ keep most informative examples at the last (closer to query)
return Ordered list of exemplar sentences $\{s^*\}$ from $\mathcal{E}$.

giving $c(s) = 1$; the third ("A native of Albuquerque, Jeff Bezos moved to Florida...") similarly clears τ on only one relation. The second sentence ("Jeff Bezos, who was born in Albuquerque...and later lived there...") clears τ on *both* relations, giving $c(s) = 2$, and is selected – despite neither of its individual relation scores being the highest in the bag – because it is the only sentence that jointly, confidently expresses both of the bag's labeled relations.

Formal definition. For every sentence $s \in B_r$, we compute a coverage score $c(s)$, defined as the number of bag relations whose confidence exceeds threshold τ:

$$c(s) = \sum_{r_a \in \text{labels}(B_r)} \mathbb{I}[f_{\text{PARE}}(s, r_a) > \tau]$$

Among sentences achieving the maximum coverage, ties are broken by the highest aggregate confidence, summed over only the relations each sentence individually clears τ on:

$$s_r = argmax_{s : c(s) = \max_{s' \in B_r} c(s')} \sum_{r_a : f_{\text{PARE}}(s, r_a) > \tau} f_{\text{PARE}}(s, r_a)$$

This two-step criterion – maximize coverage first, then aggregate confidence over covered relations as a tiebreak – favors sentences that jointly express multiple valid bag relations, rather than a sentence that happens to score highest on any single relation in isolation. The selected sentence s_r, together with the bag's full label set $\textit{labels}(B_r)$, is included as an exemplar in the prompt.

This Stage 3 design offers the following benefits:

- **Multi-label support:** By selecting sentences based on the entire bag's label set, the framework captures co-occurring relations beyond the top-K candidates.
- **Noise Reduction:** Aggregate confidence scoring favors sentences with consistent re-

lational evidence, suppressing spurious high-confidence predictions. We validate this design choice via ablations in §D.2.2.

Exemplars are ordered by increasing $f_{\text{PARE}}(q, r)$, placing the most relevant exemplars closest to the query, consistent with standard ICL practice. The complete selection procedure is detailed in Algorithm 1.

6.4 Experiment Design

The experimental evaluation assesses the effectiveness of HYDRE in English and its robustness under CLT settings to four low-resource Indic languages.

CLT Evaluation Setup. We evaluate HYDRE under three CLT settings, excluding U-CLT due to the absence of unlabeled target corpora.

1. **0-CLT:** Models are trained or prompted exclusively using English DSRE data.
    - **English queries:** HYDRE uses PARE confidence scores and the *e5-large-v2* encoder (Wang et al., 2022) for semantic similarity.
    - **Indic queries:** Since PARE cannot reliably process Indic scripts to obtain confidence scores, the $f_{\text{PARE}}(\cdot)$ component in Stages 2 and 3 is omitted. HYDRE therefore relies solely on the semantic similarity – computed using the *BGE-M3* encoder (Chen et al., 2024a) – to retrieve English exemplars.
2. **Translate-train.** Let us denote the translate-train variants of PARE and *CIL*, trained using X_{train}, as PARE-X and *CIL-X*, respectively. In this setting, we use PARE-X for confidence estimation – given its higher Recall@5 compared to *CIL-X* (Table 6.1) – and *CIL-X* for semantic similarity. We choose *CIL-X* for the latter because its contrastive training objective yields task-specific representations, enabling more effective retrieval in the target language than off-the-shelf sentence embeddings.
3. **Translate-test.** Target-language queries are translated into English using CODEC powered by IndicTrans2, and the standard English HYDRE pipeline is applied to the translated queries.

Table 6.2 summarizes the configuration per setting.

Baselines. We denote a variant as HYDRE(M) when using LLM M as the selector. We compare HYDRE against three baseline categories:

- **Supervised DSRE models:** PARE and *CIL* for English, their translate-train variants: PARE*X* and *CIL-X*.

Setting	Similarity Model	Confidence Model
0-CLT (English queries)	e5-large-v2	PARE
0-CLT (Indic queries)	BGE-M3	–
Translate-train	CIL-X	PARE-X
Translate-test	e5-large-v2	PARE

Table 6.2: **HYDRE configuration** used across various settings.

- **Zero-shot prompting:** *Qwen3-235B-A22B*, *Llama3.1-8B*, *GPT-4o*, and *Llama3.1-8B-FT* (fine-tuned *Llama3.1*), evaluated with direct and definition-based prompting.
- **Few-shot prompting:** K Exemplars are drawn from bag-flattened DSRE training data (each sentence inherits its bag's full label set) and retrieved using *Random*, *TopK*, or *MMR* (§4.4, Chapter 4).[1]

Implementation Details. We set the number of exemplars to $K = 5$, following the sensitivity analysis in §6.5.4. The confidence threshold is fixed at $\tau = 0.5$, consistent with the original PARE implementation. The trade-off parameter λ is tuned via grid search over $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ on the dev set for each language using LLAMA-3.1, yielding $\lambda = 0.5$ for English and $\lambda = 0.1$ for Indic translate-train (averaged across all Indic language F1 scores). For the Wiki-20m English dev set, $\lambda = 0.7$ is optimal.

For LLM inference, we fix temperature=0.0, a maximum sequence length of 2048 tokens, and a maximum response length of 256 tokens throughout all experiments.

We fine-tune LLAMA-3.1 on X_{train} using LoRA with PARE's passage-style training, updating only adapter and embedding weights. LoRA hyperparameters are $\{lora_rank = 16, lora_\alpha = 64, \text{dropout} = 0.0\}$. Training uses a cosine learning rate scheduler of 1×10^{-4}, a warmup ratio of 0.1, and an effective batch size of 32 (per-device batch size = 8, gradient accumulation steps = 4). The model is trained for 5000 steps on a single NVIDIA A100 40GB GPU.

Evaluation Metrics. Our primary metrics are micro-F1 and macro-F1. AUC is omitted as LLMs do not produce calibrated confidence scores.

6.5 Results & Analysis

Research Questions. We investigate the following research questions to assess the efficacy of HYDRE:

1. **RQ1 (Monolingual Efficacy):** Does HYDRE outperform standard supervised baselines and zero-shot LLM prompting on English DSRE benchmarks?

[1] Following Kapuriya et al. (2025), we refer to the MMR-based prompting method as *LM-MMR*.

2. **RQ2 (Cross-Lingual Transfer):** Do the performance gains of HYDRE extend to low-resource Indic languages under diverse cross-lingual transfer settings?

3. **RQ3 (Ablation Analysis):** What is the relative contribution of each stage in HYDRE's exemplar selection pipeline to downstream performance?

4. **RQ4 (Bag-Level Robustness):** Does HYDRE maintain its effectiveness when evaluated under the traditional bag-level inference setting?

5. **RQ5 (Qualitative Analysis):** Qualitatively, how does HYDRE resolve errors made by supervised DSRE models and zero-shot LLM prompting?

Table 6.3 summarizes the comparative results. We refer to English-only evaluation as 0-CLT with English queries.

0-CLT Results (English). Among supervised baselines, CIL (43 micro-F1) and PARE (42 micro-F1) perform comparably. Zero-shot prompting with *GPT-4o* (56 F1) significantly outperforms these supervised models, confirming the strong reasoning capabilities of modern LLMs. HYDRE further elevates this performance to 63 F1. Notably, smaller LLMs such as *Llama-3.1* and *Qwen3* exhibit the largest relative gains (+15-20 F1), suggesting that HYDRE's structured exemplars are particularly effective at guiding weaker models.

0-CLT Results (Indic). Zero-shot prompting yields low performance (20-30 F1) due to the limited exposure of Indic scripts in pretraining. HYDRE consistently improves this by 3 to 11 F1 points across all LLMs. Notably, HYDRE(Llama-3.1-FT) improves from 24/14 to 35/21, indicating that even without native-script supervision, providing semantically aligned English exemplars significantly aids cross-lingual performance.

Translate-Train (*T-CLT-Train*). While supervised translate-train baselines (e.g., *PARE-X* (30/20 F1)) outperform 0-shot prompting, HYDRE surpasses them by a wide margin. For instance, HYDRE(*Llama-3.1-FT*) achieves 47/29 F1, beating the best supervised baseline by +17 micro-F1.

We observe distinct script-dependent trends (per-language results in §D.2.1): GPT-4o remains competitive on medium-resource scripts (Oriya, Tulu), but its zero-shot performance collapses on rarer scripts – Santali and Manipuri (Ol Chiki and Meitei Mayek scripts, respectively) both score just 8/6 F1, indicating near-total tokenizer breakdown rather than a merely weaker result. Even with HYDRE's exemplar selection, GPT-4o only partially recovers (20/11 Santali, 21/14 Manipuri), while HYDRE(Llama3.1-8B-FT$_X$) closes the gap decisively, reaching 39*/23 on Santali and 34*/15 on Manipuri. This is consistent with prior work showing commercial LLM tokenizers are optimized for high-resource-language vocabularies, systematically disadvantaging low-resource, rare-script languages (Nag et al., 2024).

Model	English	Indic Languages		
		0-CLT	Translate-train	Translate-test
Supervised				
CIL	43/32	–	26/18	34/24
PARE	42/31	–	30/20	33/23
0-shot (direct)				
Qwen3-235B-A22B	47/39	25/21	25/21	40/34
Llama3.1-8b	24/21	16/12	16/12	22/20
Llama3.1-8B-FT	55/37	26/17	26/17	47/31
GPT-4o	56/55	31/29	31/29	51/49
0-shot (Ontology-based)				
Qwen3-235B-A22B	49/40	25/21	25/21	44/35
Llama3.1-8b	31/17	22/16	22/16	29/25
Llama3.1-8B-FT	60/44	24/14	38/23	49/34
GPT-4o	56/57	33/31	32/30	54/51
few-shot (Random)				
Random(Qwen3-235B-A22B)	55/55	21/19	29/25	43/35
Random(Llama3.1-8B)	30/24	19/11	11/7	27/21
Random(Llama3.1-8B-FT)	55/40	24/14	25/12	46/32
Random(GPT-4o)	56/52	31/30	38/31	48/42
few-shot (Similarity-based)				
TopK(Qwen3-235B-A22B)	52/49	22/18	32/26	45/38
TopK(Llama3.1-8B)	33/27	22/14	19/9	30/22
TopK(Llama3.1-8B-FT)	55/40	27/15	29/14	46/30
TopK(GPT-4o)	58/53	29/26	41/32	48/41
few-shot (Diversity-based)				
LM-MMR(Qwen3-235B-A22B)	50/47	24/20	33/28	44/36
LM-MMR(Llama3.1-8B)	34/26	21/12	17/8	30/22
LM-MMR(Llama3.1-8B-FT)	56/40	26/16	28/15	47/32
LM-MMR(GPT-4o)	56/52	28/25	41/33	50/43
few-shot (HYDRE)				
HYDRE(Qwen3-235B-A22B)	**63***/**62**	29/26	38/31	54/46
HYDRE(Llama3.1-8B)	52/47	30/19	35/21	44/36
HYDRE(Llama3.1-8B-FT)	61/45	35/21	**47***/29	51/37
HYDRE(GPT-4o)	**63***/60	**36***/**33**	38/**34**	**56***/**54**

Table 6.3: **Main results** (micro/macro-F1) on English (NYT-10m) and Indic languages (INDICRE4, averaged over 4 languages). Indic results span three CLT settings; per-language breakdowns in §D.2.1. **Bold**: best per-column; *: statistically significant (McNemar's $p < 10^{-5}$).

This suggests a favorable tradeoff for deployment: for very low-resource, rare-script languages, locally fine-tuned open-source models are not merely competitive but close to necessary. For medium-resource scripts, proprietary models remain the stronger choice on accuracy and require no fine-tuning infrastructure, though Nag et al. (2024) suggests their per-token API cost likely still carries some script-driven penalty relative to high-resource languages.

Translate-Test (*T-CLT-Test*). Translating test queries into English introduces a mild degradation relative to native English evaluation, primarily due to translation noise. Despite this, HYDRE delivers consistent gains: Llama3.1 and Qwen3 gain +15 and +10 micro-F1 points over

their zero-shot counterparts, and GPT-4o reaches 56/54 F1—the best overall performance in this setting.

HYDRE vs. Other Few-Shot Methods. Across all settings, HYDRE consistently outperforms similarity-based (*TopK*) and diversity-based (*MMR*) baselines. Compared to *LM-MMR*, HYDRE achieves gains of up to +13 F1 in English and +19 avg. F1 in Indic Translate-train. These improvements stem from our *label-aware exemplar selection*: unlike MMR, which optimizes only for semantic diversity, HYDRE uses DSRE confidence to select exemplars that are both diverse and highly discriminative for the target relation. This mirrors the advantages of label-aware diversity observed in the SSP framework (Chapter 4), validating the principle across different structured prediction tasks.

6.5.1 Ablations

To quantify the contribution of each component within HYDRE, we conduct systematic ablations of its three-stage design. We evaluate variants removing *Candidate Selection (Stage 1)*, *Bag Selection (Stage 2)*, *Sentence Selection (Stage 3)*, and specific scoring components (Semantic Similarity vs. PARE Confidence).

Table 6.4 summarizes the results. Across all models and languages, the full HYDRE pipeline consistently outperforms ablated variants, confirming the complementary value of each stage.

Ablation	English			Indic (Translate-train)			
	L3.1	Qw3	G4o	L3.1	L3.1-FT	Qw3	G4o
HYDRE	**52/47**	**63/62**	**63**/60	31/**20**	**45/28**	37/**31**	**39/35**
w/o candidate selection (Stage 1)	39/31	**63**/61	62/**64**	–	–	–	–
w/o bag selection (Stage 2)	48/41	62/59	56/51	**33/20**	43/25	34/30	38/**35**
w/o sem. sim. ($\lambda = 1.0$)	45/39	60/59	60/62	32/**20**	**45/28**	**38/31**	38/34
w/o ICL	45/37	58/50	59/49	29/19	33/18	33/20	32/21

Table 6.4: **Ablation study for HYDRE** on English NYT-10m and INDICRE4 (translate-train), reporting micro/macro-F1. **L3.1**: LLaMA-3.1-8B; **L3.1-FT**: LLaMA-3.1-8B-FT; **Qw3**: Qwen3-235B; **G4o**: GPT-4o. "–": not applicable (omitted for Indic languages due to excessive prompt length from high token fertility).

Semantic similarity vs. DSRE confidence. In English, semantic similarity is the dominant signal: removing it degrades Llama 3.1 by up to 7 F1 points, while ablating PARE confidence leads to smaller drops (2-3 F1) across LLMs.

In contrast, under the Indic translate-train setting, DSRE confidence becomes more critical. Removing semantic similarity causes little or no degradation, highlighting the unreliability of embedding-based retrieval in low-resource and script-diverse regimes (Lin et al., 2025).

Candidate relation selection. Providing exemplars for all ontology relations (i.e., removing Stage 1) yields performance comparable to HYDRE, but at approximately 5× higher prompt cost. While Qwen3 and GPT-4o remain relatively stable, LLaMA exhibits noticeable degradation, indicating that excessive distractor relations hinder reasoning in smaller models. This confirms that candidate relation selection is essential not only for efficiency but also for stable inference.

Importance of bag selection. Omitting Bag Selection (Stage 2) and retrieving sentences directly leads to consistent performance drops, particularly in the translate-train setting (up to 4-6 F1 points). By constraining sentence selection to bags aligned with candidate relations, HYDRE reduces noise from distant supervision and preserves relation-consistent evidence.

Importance of sentence selection. Providing entire bags instead of selecting representative sentences degrades performance by up to 6 micro-F1 and 10 macro-F1 points. This suggests that LLMs struggle with longer and noisier contexts, and highlights HYDRE's ability to exploit a key property of distant supervision: precise relation inference often requires only a small subset of sentences.

Importance of in-context exemplars. Removing in-context exemplars reduces performance by up to 7 F1 points on English and up to 12 F1 points in the translate-train setting. This confirms that HYDRE's gains arise from exemplar-guided reasoning rather than exposure to candidate relations alone.

6.5.2 Qualitative & Error Analysis

We conduct a confusion analysis comparing HYDRE against PARE and zero-shot GPT-4o on the English NYT-10m dataset (Table 6.5). In addition, we manually identify 12 semantically overlapping relation pairs that frequently exhibit confusion (Figures D.1 and D.2). To illustrate HYDRE's reasoning behavior, we provide two main-text working examples (Figure 6.3) and 8 additional representative cases in §D.2.3 (Figures D.3a–D.6b).

Overall, HYDRE demonstrates a significantly stronger ability to resolve closely related relations and overcome baseline limitations. For example, it successfully disambiguates *Major-shareholders* from *Founders* (Figure 6.3a), correctly predicts *place_of_burial* where baselines default to the superficially similar *place_lived* (Figure D.3a), and accurately identifies complex corporate relations like *company/advisors* (Figure D.5b). HYDRE is also highly effective at exhaustive multi-label extraction, recovering both gold labels for a single entity pair where baselines only captured one (Figure D.3b).

Crucially, this reasoning capacity transfers to low-resource languages, and effectively resolves the predominant false negative ("NA") cases. As shown in Table 6.6, HYDRE recovers

166 gold relations that PARE defaults to NA, while losing only 6. This confirms that LLM-based disambiguation primarily addresses PARE's conservative abstention behavior. Strikingly, HYDRE can even overcome the "missing candidate bottleneck" across languages. It successfully predicts the correct gold label for English *Majorshareholders* (Figure 6.3a) and Manipuri *business/location* (Figure D.6a) even when those labels are absent from PARE's top-5 retrieved candidates. Furthermore, HYDRE correctly extracts Manipuri *founders* where 0-shot returns NA and PARE predicts an incorrect relation (Figure 6.3b).

	PARE Correct	**PARE Wrong**
HYDRE Correct	172	223
HYDRE Wrong	25	272

	0-Shot Correct	**0-Shot Wrong**
HYDRE Correct	312	83
HYDRE Wrong	30	267

Table 6.5: **Confusion analysis for HYDRE vs. the baselines**, aggregated over all relations (English, NYT-10m).

	PARE Correct	**PARE predicts NA**
HYDRE correct	149	**166**
HYDRE predicts NA	6	13

Table 6.6: **HYDRE vs. PARE confusion on gold non-NA instances** (English NYT-10m). HYDRE recovers **166** relations where PARE defaults to NA, while losing only 6.

Despite these improvements, an analysis of the failure cases reveals four primary error categories:

1. **Position bias and partial multi-label recall:** 34% of HYDRE's errors involve selecting the top-ranked PARE candidate, compared to 27% for zero-shot prompting. This likely stems from placing the highest-ranked exemplar closest to the query in the prompt – a known instance of *recency bias* in LLM-as-a-judge frameworks (Zheng et al., 2023) (Figures D.4a, D.4b and D.6b). A direct consequence of this bias is that HYDRE sometimes struggles with perfectly exhaustive multi-label extraction. While it drastically improves overall multi-label coverage (Table D.5), its hyper-focus on the most recent, top-ranked candidate often prevents it from predicting the remaining valid labels, resulting in partial predictions for complex queries (Figures D.4a and D.6b).

2. **Missing candidate bottleneck:** Although HYDRE occasionally recovers relations missing from the PARE's candidate relations (Figures 6.3a and D.6a), overall performance still degrades heavily when the gold relation is absent from PARE's top-5 candidates. As seen in Table D.7, HYDRE's single-label performance trails 0-shot prompting in this regime (~14% of all queries), likely due to the LLM grounding on overly noisy exemplars.

3. **Persistent semantic overlap:** Some relations remain intrinsically difficult for the LLM to separate, particularly when gold annotations themselves are ambiguous. Both HYDRE

and 0-shot predict *nationality* for "American" in Fig. D.5a, where the gold label is *ethnicity* – a distinction that is arguably ill-defined for demonyms. Similarly, both misclassify *business/location* as *place_founded* when the lexical cue "based in" is present (Fig. D.4b).

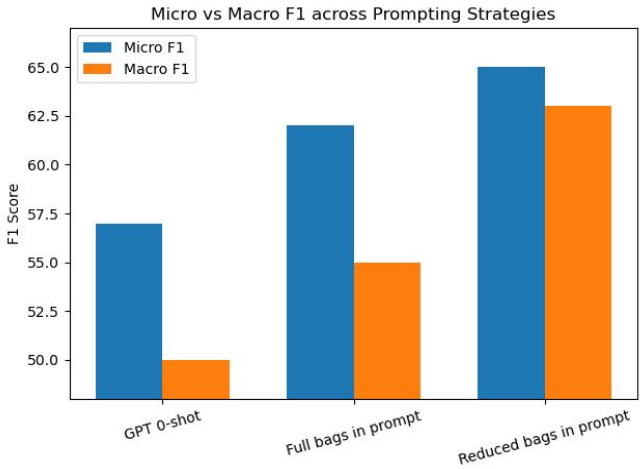


Figure 6.4: **Bag-level prompting strategies** for GPT-4o on English NYT-10m: *zero-shot* vs. *full-bag* ICL vs. *reduced-bag* ICL.

6.5.3 Robustness to Bag-level Evaluation

We further evaluate HYDRE under the bag-level protocol introduced in Chapter 5. To adapt the pipeline, we follow Algorithm 1 through *Stage 2 (bag selection)*. In *Stage 3*, instead of selecting a single representative sentence per bag, HYDRE retains one sentence *per bag relation*, choosing the sentence with the highest PARE confidence for that relation. This yields a *reduced-bag* representation that preserves relational evidence while filtering redundant sentences.

As shown in Fig. 6.4, reduced-bag prompting improves performance by approximately +3 micro-F1 and +7 macro-F1 over full-bag prompting, while nearly halving prompt length (2451 $\rightarrow$ 1242 tokens). This substantially lowers inference cost without sacrificing accuracy.

These results reinforce a key property of DSRE: effective inference depends on a small set of reliable relational cues rather than exhaustive bag-level context.

6.5.4 HYDRE Sensitivity to K

We analyze the sensitivity of HYDRE to the number of candidate relations (K) selected in Stage 1. As shown in Fig. 6.5a, PARE attains high Recall@K even for small K values ($\sim$ 85% at K = 5), with recall continuing to increase monotonically thereafter. In contrast, Fig. 6.5b shows that HYDRE's micro-F1 improves with K up to $K \in [5, 10]$, after which performance saturates or slightly degrades.

This divergence reflects a core trade-off: larger K increases coverage but also introduces distractor relations and longer prompts, which complicate LLM-based disambiguation. This

ICL Exemplars (HYDRE)

Input: Mr. Marek started the <Head> Fairfield </Head> Theater Company in 2001 ... at <Tail> Fairfield University </Tail> .
Output: /location/location/contains

Input: But Mr. Wallace, 45, a businessman and an investor in partnership for a time with <Tail> Mark Thatcher </Tail> , son of the former British Prime Minister <Head> Margaret Thatcher </Head>, has detractors, including ... . ”
Output: /people/person/children

Input: Behind the News – Founded eight years ago in a Silicon Valley garage by two <Tail> Stanford University </Tail> graduate students, <Head> Google </Head> went public two years ago at $ 85 a share.
Output: /business/company/place_founded

Input: A notch or two down-market, the 777 ’s, 767 ’s, and 757 ’s are often coveted by corporate titans, among them Larry Page and <Tail> Sergey Brin </Tail>, the co-founders of <Head> Google </Head>, who bought ... .
Output: /business/company/founders /business/company/majorshareholders

Input: When he was 32 , <Tail> Bill Hambrecht </Tail> was a co-founder of <Head> Hambrecht & Quist </Head> , a West Coast investment bank that ... .
Output: /business/company/founders

Query:

Input: Mr. Bechtolsheim , one of the first investors in <Head> Google </Head> , co-founded Kealia in 2001 with <Tail> David Cheriton </Tail> , a Stanford professor who was another early Google investor .

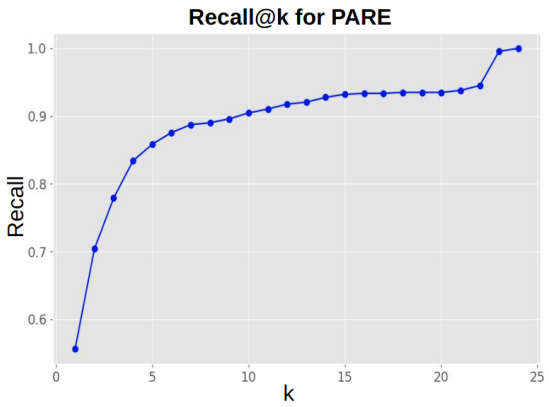
Outputs:

HYDRE: **/business/company/majorshareholders**
0-shot: **/people/person/company**
PARE: **NA**

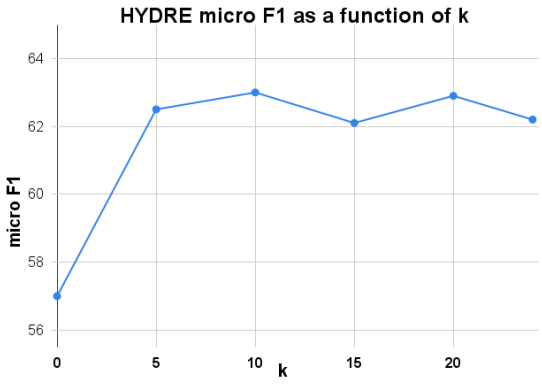
(a) **Example 1:** HYDRE correctly predicts the “*Majorshareholders*” relation while both 0-shot and PARE fail to recognize it.

ICL Exemplars (HYDRE)

Input: .... ꯑꯦꯛꯄꯣꯛꯇꯤ ꯁ꯭ꯀꯔꯑꯣꯏꯟ꯫ꯋꯥ ꯑꯦꯛ ꯅꯤ꯫꯫ꯍꯧꯔꯤꯡꯅ ꯋꯥꯍꯩ ꯁꯣꯟꯗꯣꯜ ꯑꯦꯛ ꯅꯤ꯫꯫ ꯁꯤꯡꯍꯟ ꯑꯦꯛ ꯅꯤ꯫꯫ ꯑꯦꯛ ꯅꯤ꯫꯫ ꯅꯤ꯫꯫ ꯑꯦꯛ ꯅꯤ꯫꯫ ꯑꯦꯛ ꯅꯤ꯫꯫ ꯑꯦꯛ



a PARE Recall@K vs. K



b HYDRE (GPT-4o) micro-F1 vs. K

Figure 6.5: **Effect of candidate set size** (K) on PARE's Recall@K and downstream HYDRE performance (English, NYT-10m dev set).

shows that moderate K values can provide the best balance between recall, accuracy, and inference cost.

6.5.5 Additional Results on IndoRE (Supervised RE Setting)

IndoRE's supervised, sentence-level setting – fundamentally different from the noisy, bag-level distant supervision HYDRE targets elsewhere – motivates two questions: (a) can GPT-4o outperform fine-tuned PARE on clean data, and (b) is HYDRE's multi-stage pipeline still necessary without noisy bags? Since IndoRE has no bag structure, Stage 2/3 (bag and sentence selection) do not apply and are not ablated; only Stage 1 and exemplar retrieval are.

Component ablations. Table 6.7 reports HYDRE(GPT-4o) ablations on IndoRE. Removing Stage 1 costs 2.6 micro-F1, confirming candidate-relation filtering remains valuable even without noisy bags. Removing PARE-informed confidence scoring, by contrast, *improves* results by 1.0 micro-F1 – the strongest configuration overall – while removing semantic-similarity retrieval instead costs 2.5 micro-F1, comparable to removing Stage 1. We interpret this as consistent with PARE's confidence signal plausibly serving to compensate for distant-supervision noise elsewhere in this thesis; with no such noise to correct in IndoRE's already-filtered data, this signal appears to add no value and mildly misleads, whereas semantic-similarity-based retrieval remains a generically useful exemplar-selection heuristic regardless of supervision quality. Removing exemplars entirely (zero-shot) costs the most (7.9 micro-F1), confirming in-context exemplars remain essential throughout.

Variant	Micro-F1	Macro-F1
HYDRE (full)	79.7	74.5
w/o Stage 1	77.1	72.3
w/o PARE confidence (sem-sim only)	**80.7**	**75.6**
w/o semantic-similarity retrieval	77.2	74.8
w/o ICL (zero-shot)	71.8	69.1

Table 6.7: HYDRE(GPT-4o) ablations on IndoRE. Bold: best variant.

LLM vs. fine-tuning at scale. Even the strongest HYDRE(GPT-4o) variant (80.7/75.6) remains meaningfully below PARE's 89.8/85.2 – the LLM does not outperform a fine-tuned model trained on the full, manually filtered training set. Figure 6.6 characterizes this gap as a function of training-set size, progressively sampling 10–100% of IndoRE's training split. HYDRE remains stable throughout (78–81 micro-F1) regardless of training size, while PARE improves monotonically from 69.5 (10%) to 90.0 (100%), overtaking HYDRE once roughly 70–80% of the training data is available – approximately 4.8k–5.4k of IndoRE's 6.8k total training examples.

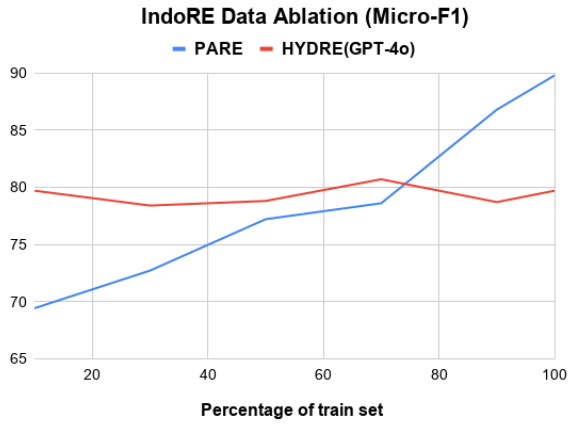


Figure 6.6: HYDRE(GPT-4o) vs. PARE micro-F1 as a function of training-set percentage on IndoRE.

Summary. In supervised, filtered settings, HYDRE's value comes from candidate filtering and similarity retrieval, not PARE's noise-correction scoring, which has nothing to correct here. HYDRE remains most valuable in low-data regimes, with fine-tuning overtaking once sufficient clean data is available – reinforcing this thesis's broader theme of data-availability-driven complementarity between LLM-based and fine-tuned approaches.

6.5.6 Additional Results on Wiki-20m

We additionally test the 's generalizability on Wiki-20m (Gao et al., 2021), which features a significantly larger ontology of 81 relations. As shown in Table 6.8, consistently outperforms zero-shot baselines across all models, with micro-F1 gains of up to 36 points for Llama-3.1 and

Ablation	Llama3.1	Qwen3	GPT-4o
0-shot	13/13	33/33	55/48
HYDRE	**49/47**	**69/64**	67/63
w/o semantic similarity	27/33	42/45	54/58
w/o PARE confidence	25/19	52/50	58/51
w/o Stage 2	41/40	59/56	65/62
w/o Stage 3	48/**47**	68/**64**	**69/64**
w/o ICL	31/34	56/56	60/56

Table 6.8: F1 scores (micro/macro) on Wiki-20m test set.

Qwen3. The *w/o ICL* variant – which restricts the LLM to PARE's top-5 candidate relations without providing exemplars –already shows a substantial improvement over the standard zero-shot baseline. This underscores the efficacy of using a specialized DSRE model to narrow the candidate space from 80 relations down to 5, significantly reducing task complexity for the LLM. Furthermore, the *w/o Stage 3* ablation (entire bag retrieval) performs comparably to the full pipeline. This is due to the dataset's low density, with an average of 1.8 sentences per training bag, which renders the additional noise-filtering and sentence-selection mechanisms of Stage 3 largely redundant compared to denser datasets like NYT-10m (5.3 sentences/bag).

6.6 Discussion and Conclusion

This chapter presented HYDRE,[2] a hybrid framework that integrates the high recall of supervised DSRE models with the precision of *LLM-as-selector*. By transforming noisy bag-level supervision into clean, discriminative exemplars through a three-stage selection pipeline, HYDRE achieves consistent gains over both supervised baselines and standard prompting strategies across English and cross-lingual transfer to four low-resource Indic languages. Extensive ablations confirm that each stage contributes complementary value.

HYDRE's Efficiency. HYDRE's design is inherently prompt-efficient. Stage 1 constrains the candidate set to at most five relations, keeping prompts compact and focused - ablations show that presenting the full ontology introduces distractors that degrade performance, particularly for smaller LLMs. Stage 3 further reduces prompt length by nearly half by selecting a single representative sentence per bag rather than providing the full bag as context, while simultaneously improving performance. These efficiency trends are especially important for low-resource languages, where higher subword fragmentation inflates token counts. For severely low-resource scripts such as Santali and Manipuri, lightweight fine-tuning of local models followed by HYDRE-based inference outperforms larger proprietary models, suggesting that modest local adaptation combined with efficient prompting can simultaneously reduce cost and improve performance.

[2] https://github.com/dair-iitd/HYDRE

Limitations & Future Directions. Despite its strengths, HYDRE exhibits position bias toward top-ranked candidates, limited multi-label recall, and degraded performance when the gold relation falls outside the candidate set.

Two broader directions emerge from this work: (a) *agentic workflows* that iteratively refine relation predictions through multiple rounds of candidate generation, exemplar retrieval, and disambiguation, potentially incorporating self-verification to reduce position bias; and (b) *real-world deployment* on evolving data sources such as Wikipedia updates or regional-language news, where HYDRE's ability to operate without retraining could support scalable knowledge base construction for under-resourced languages.

The next chapter concludes by summarizing the contributions of this thesis and discusses potential avenues for future research.

Chapter 7

Conclusion and Future Work

7.1 Conclusion and Summary

This thesis explored how hybrid models – combining supervised task-specific models with Large Language Models (LLMs) – can improve cross-lingual transfer to low-resource languages. Our work spanned two structured prediction tasks, sequence labeling and distantly supervised relation extraction (DSRE), and examined both supervised and hybrid modeling paradigms across zero-resource and low-resource cross-lingual settings.

The central thesis is that supervised models serve distinct but complementary roles: fine-tuned models provide structured, high-precision priors over the output space, while LLMs contribute broad semantic reasoning and ontological knowledge. Hybridizing these two at inference time – rather than replacing one with the other – yields a more powerful and data-efficient strategy than either alone.

Our investigation began with fine-tuning strategies for sequence labeling tasks, specifically NER and POS tagging. Existing approaches suffer from notable limitations – some entangle language representations (Muller et al., 2021; Üstün et al., 2020), while others are restricted to single-source transfer (Pfeiffer et al., 2020). To address these shortcomings, we introduced ZGUL, a multi-source language adapter ensembling framework for efficient zero-resource cross-lingual transfer (0-CLT). ZGUL enables modular composition of language-specific adapters from multiple high-resource languages, dynamically weighed at test time, without requiring any target-language supervision.

Next, we observed that existing prompting approaches often rely on static or task-agnostic exemplar selection, which limits their effectiveness for structured prediction (Ahuja et al., 2023). In contrast, task-specific models guide exemplar selection, while LLMs provide semantic reasoning and ontological knowledge. This complementary relationship motivates our hybrid approach. For sequence labeling, we proposed SSP, a two-stage framework: in Stage 1, a fine-tuned cross-lingual model annotates the target-language test set; in Stage 2, high-confidence and semantically relevant instances are selected – via an ILP-based objective – as monolingual

in-context exemplars for the LLM. This avoids the need for human-annotated target-language demonstrations entirely.

We then turned to distantly supervised relation extraction (DSRE), where even source-language supervision is inherently noisy (Mintz et al., 2009). Our contributions here form a two-step progression. PARE addresses the architectural bottleneck of independent sentence encoding by treating each bag as a unified passage with relation-aware token attention. However, this bag-level design trades off sentence-level precision: PARE frequently conflates semantically adjacent relations and falls short on cross-lingual transfer to unseen scripts – limitations that stem from the representational ceiling of mBERT-style encoders rather than the passage encoding itself.

HYDRE closes this gap at inference time by integrating fine-tuned DSRE models with LLMs, without retraining either component. By using PARE's high Recall@K (K=5) to prune the candidate relation space and then routing a carefully constructed exemplar set to an LLM for final disambiguation, HYDRE converts noisy bag-level supervision into clean, sentence-level evidence. The key insight is that PARE's limitations – overconfident abstention and coarse granularity – are precisely where LLM reasoning excels, while LLM weaknesses in exemplar selection are mitigated by PARE's candidate filtering. Together, they yield consistent gains across English and four low-resource Indic languages.

Overall, this thesis demonstrates that combining supervised models with LLM-based reasoning provides a powerful, generalizable, and annotation-efficient strategy for structured prediction, particularly in low-resource and cross-lingual settings. The results across NER, POS tagging, and DSRE – spanning typologically diverse language families including Germanic, Indic, and African languages – collectively support this hybrid paradigm. Importantly, these gains are consistent across both monolingual and cross-lingual settings, highlighting the robustness of the hybrid paradigm.

7.2 Future Work

The contributions of this thesis lay the foundation for several meaningful research directions, including the extension of hybrid CLT framework to broader tasks, more culturally nuanced environments, and computationally constrained deployment scenarios.

Generalizing to Unseen Scripts. Many low-resource languages use scripts underrepresented in current multilingual LLMs (Pfeiffer et al., 2021b; Singh et al., 2024). Although translation-based transfer partially mitigates this issue, it introduces additional noise and inference cost. Future directions include script-agnostic representations, phoneme- or transliteration-based adapters, and the use of multilingual speech or OCR signals to bridge script gaps. Investigating how matrix-factorization-based vocabulary adaptation (Pfeiffer et al., 2021b) can complement ZGUL-style ensembling is a particularly concrete next step.

Deep Agentic Workflows. Beyond single-pass hybridization, *agentic workflows* offer a promising direction for *iterative refinement* (Madaan et al., 2023). In such settings, the LLM could iteratively query task-specific models to gather additional evidence, expand its reasoning context, and critique its own intermediate predictions in a *self-refinement loop*. Moreover, multiple supervised models – offering varying precision-recall characteristics – could generate diverse candidate sets, which the LLM can consolidate to identify the most reliable and complementary signals for constructing efficient prompts (Patidar et al., 2024).

Cost-effective & Scalable Deployment. LLM inference remains computationally and financially expensive, particularly for low-resource languages where high tokenizer fertility lengthens prompts. Future work could explore lightweight or cascaded inference, *early-exit* mechanisms, *iterative prompt compression* (Jiang et al., 2024), and uncertainty-based exemplar pruning. Our preliminary cost-performance analysis, spanning both SSP (Ch. 4) and HYDRE (Ch. 6), points to two further, complementary extensions here, alongside the candidate-pool-size scalability strategy already proposed for SSP's exemplar selection specifically (§4.5.5).

Capacity-Aware Adaptation. A systematic mapping of the *LLM parameter-scale* threshold at which informed adaptation (exemplar selection or fine-tuning) stops adding value relative to raw model scale. Computing SSP's gain over TopK across the three LLM tiers evaluated in Table 4.1 (LLaMA-2, GPT-3.5, GPT-4x) shows this gain peaking at the mid-tier model rather than increasing monotonically with scale – a single-instance observation across three tiers, not yet a systematically confirmed pattern, which is precisely what motivates the proposed future work. This is a distinct scaling axis from the candidate-pool-size scaling addressed in §4.5.5.

Cost-Aware Agentic Routing. *Adaptive model routing* (Aggarwal et al., 2024) could be extended beyond instance-level complexity to route based on *when* a task exposes a model's representational weaknesses – e.g., our finding that locally fine-tuned open-weight models outperform proprietary LLMs specifically when native rare-script text must be processed directly, but not once translation removes that exposure, suggesting routing policies could condition on pipeline stage (native-script vs. translated input), not just instance difficulty. In this paradigm, simpler instances – whose complexity can be estimated via the supervised model's entropy – are handled by zero-shot or translate-train models, reserving the full hybrid pipeline for more ambiguous cases. More broadly, we envision generalizing this into a fully *cost-aware agentic pipeline*, where translation, transliteration, a locally fine-tuned task-specific model, and a proprietary LLM are treated as a heterogeneous set of callable tools, each with a distinct cost, latency, and performance profile. Since the value of invoking a costlier tool is highly instance- and script-dependent, we propose posing tool selection as a sequential decision-making problem, naturally framed as a Markov Decision Process, directly amenable to reinforcement learning-based policy optimization. This bounds cost across *which* model or tool is invoked per query, complementing

rather than duplicating §4.5.5's proposal for bounding cost as the candidate *pool* grows.

Cross-Cultural Adaptation. A persistent challenge in cross-lingual NLP is generalizing to unseen cultural contexts, where social norms and implicit world knowledge differ in spite of surface linguistic similarity (Liu et al., 2025). Building on ZGUL, future work could explore *cultural adapter ensembling*, leveraging adapters trained on culturally grounded corpora – such as Indic folktales, regional news, or community-generated content. Incorporating such cultural signals into self-adaptive exemplar selection (as in SSP) may help models mitigate bias toward dominant global perspectives (often characterized as WEIRD – *Western, Educated, Industrialized, Rich, and Democratic* – norms (Henrich et al., 2010)) and improve adaptation to diverse social and geographic contexts.

Scalable Knowledge Base Construction (KBC). Combining the cross-lingual DSRE capabilities of HYDRE with the NER pipelines enabled by SSP offers a natural path toward scalable multilingual knowledge base construction. Such a system could extract triples (e_1, r, e_2) from evolving sources – Wikipedia updates, streaming news, or social media – enabling continual KB expansion for low-resource languages.

A multilingual KB constructed in this manner supports several downstream applications (Chakrabarti et al., 2022). First, in *question answering (QA)*, entities and relations can be retrieved directly from the KB to support factual, evidence-grounded responses. Second, in *information retrieval (IR)*, entity-centric indexing over KB triples enables more precise semantic search than dense vector retrieval alone, particularly for named-entity queries common in domains such as news and law. Third, for *hallucination mitigation*, conditioning generation on verified KB triples extracted via HYDRE provides a factual scaffold that constrains the LLM's output space and improves reliability (Edge et al., 2024).

Beyond Structured Prediction: Cross-Lingual Generation. While this thesis primarily focuses on structured prediction tasks with fixed output spaces, extending hybrid methods to open-ended cross-lingual generation presents fundamentally different challenges (Singh et al., 2024; Nag et al., 2025). In generative settings, the output space spans the full vocabulary, making label-coverage and confidence-based strategies - such as those used in SSP and HYDRE - less applicable.

Future work could explore *constrained decoding* to enforce structural consistency, retrieval mechanisms grounded in structured alignment rather than purely semantic similarity, or grounding generation in knowledge-base triples (e.g., extracted using HYDRE) to reduce hallucinations and improve *faithfulness* in generation (Nandwani et al., 2023).

Appendix A

Appendix to Chapter 3

A.1 Implementation Details

A.1.1 Hyperparameters

Hyperparameter	SFT	Adapter Based
Learning Rate	{2e-5, 3e-5, 5e-5}	{5e-5, 1e-4}
Max. epochs	10 for NER, 5 for POS	10 for NER, 5 for POS
Reduction Factor	NIL	{3, 4}
Batch Size	{16, 32}	{16, 32}
EM Steps	NIL	{1, 5, 10}
EM LR	NIL	{0.05, 0.1, 0.5, 1.0}
Few-shot LR	{1e-5, 5e-5, 1e-4}	{1e-5, 5e-5, 1e-4}
Few-shot epochs	{1, 5, 10}	{1, 5, 10}
Few-shot Batch Size	{1, 4, 8}	{1, 4, 8}

Table A.1: **ZGUL hyperparameter grid details**

A.1.2 Data Statistics

Table A.2 reports training and development set sizes per language group. Table A.3 summarizes target-language test sets.

Group	Task	Train Languages	Train size	Dev size
Indo-Aryan	NER	{En,Hi,Bn,Ur}	55,026	13,003
Germanic	POS	{En,Is,De}	222,792	28,394
Slavic	POS	{En,Ru,Cs}	179,043	23,479
African	NER	{En,Amh,Swa,Wol}	19,788	4,073

Table A.2: **Training and development set sizes** (sentences) per language group. English included in all groups.

Language	Script	Size	Overlap(%)
Fo	Latin	1,208	61.7
Got	Latin	1,031	21.6
Gsw	Latin	100	54.3
Qpm	Cyrillic	635	42.9
Hsb	Latin	626	43.1
Orv	Cyrillic	4,204	45.0
Cu	Cyrillic	1,141	19.2
As	Bengali	100	44.1

Language	Script	Size	Overlap(%)
Bh	Devanagari	102	64.0
Hau	Latin	570	37.1
Ibo	Latin	642	32.0
Kin	Latin	611	32.6
Lug	Latin	419	28.3
Luo	Latin	189	35.8
Pcm	Latin	600	82.3

Table A.3: **Target-language test sets**: script, size (sentences), and token-level overlap (%, IOU) with source training data.

A.1.3 Standard Deviations

Table A.4 reports F1 standard deviations across three seeds.

	Germanic				Slavic				
	Fo	Got	Gsw	**Avg**	Qpm	Hsb	Orv	Cu	**Avg**
MADX^{multi}-EMEA	0.5	0.8	0.4	0.5	0.3	0.2	0.2	0.7	0.4
SFT-M	0.3	1.2	0.6	0.7	0.4	0.3	0.2	0.4	0.1
CPG	0.1	0.5	0.2	0.3	0.3	0.2	0.0	0.2	0.0
ZGUL	0.0	0.5	0.4	0.1	0.5	0.5	0.1	0.3	0.1

	African							Indo-Aryan		
	Hau	Ibo	Kin	Lug	Luo	Pcm	**Avg**	As	Bh	**Avg**
MADX^{multi}-EMEA	1.4	1.2	0.9	0.4	0.5	1.4	1.1	1.2	1.3	0.5
SFT-M	1.3	0.8	0.7	0.7	0.6	0.5	0.7	1.4	1	0.9
CPG	1.3	0.9	1.4	0.9	0.9	0.8	1	1.1	1.5	0.8
ZGUL	1.1	0.9	0.7	0.4	0.5	0.3	0.4	0.7	1.1	0.5

Table A.4: **Standard deviation for ZGUL vs. the baselines** over micro-F1 (rounded to 1 decimal). *Avg* column: std. dev. of the group-average F1, not the average of per-language std. devs.

A.1.4 EM Algorithm

The full EM algorithm is presented in Algo. 2.

Differences from EMEA. ZGUL modifies the original EMEA (Wang et al., 2021) in two ways: (1) token-level fusion weights are trainable independently per layer (vs. layer-wise tied as in EMEA), and (2) dual attention pathways (local + global) are initialized from trained networks (vs. a single network with uniform initialization). Table A.5 confirms a modest gain from (1).

Algorithm 2 Inference-Time EM for ZGUL

Require: Initial ensemble scores $\alpha^0 = (\alpha_F^0, \alpha_L^0)$, test input x, learning rate lr, number of EM steps T
Ensure: Prediction $\hat{y}$

function EM()
  for $t \leftarrow 0$ to $T-1$ **do**
    $\beta^t \leftarrow Softmax(\alpha^t)$ ▷ Normalize ensemble weights
    $H(x, \alpha) \leftarrow Entropy(TA \circ L_{wavg}(h, \beta^t) \circ M)$ ▷ Compute Entropy
    $g_F^t = \nabla_\alpha H(x, \alpha_F^t)$ ▷ Gradient for token-leve fusion
    $\alpha_F^{t+1} \leftarrow Update(\alpha_F^t, g_F^t)$ ▷ Update local fusion ensemble weights
    $g_L^t = \nabla_\alpha H(x, \alpha_L^t)$ ▷ Gradient for global fusion
    $\alpha_L^{t+1} \leftarrow Update(\alpha_L^t, g_L^t)$ ▷ Update global fusion ensemble weights
    $\alpha^{t+1} = (\alpha_F^{t+1}, \alpha_L^{t+1})$
  end for
  $\alpha^T \leftarrow Softmax(\alpha^T)$ ▷ Final Normalization
  $\hat{y} \leftarrow Predict(TA \circ L_{wavg}(h, \alpha^T) \circ M))$
end function

	Hau	Ibo	Kin	Lug	Luo	Pcm	As	Bh	Fo	Got	Gsw	Qpm	Hsb	Orv	Cu	**Avg.**
Layer-wise Tied																
(original EMEA)	53.2	56.9	55.2	54.1	39.8	65.8	74.1	64.1	76.4	18.6	61.9	49.5	76.7	62.3	33.4	56.1
ZGUL	53.6	56.8	56.2	54.2	40.2	66.5	74.4	64.1	76.9	20.2	64.8	50.1	77.2	63	34.1	**56.8**

Table A.5: **EM weight granularity ablation**: token-wise trainable weights (ZGUL) vs. layer-wise tied weights (used in original EMEA).

A.2 Robustness with XLM-Roberta

To verify that ZGUL's gains are not specific to mBERT, we evaluate on XLM-R-Base for Germanic (POS) and Indo-Aryan (NER) groups. Table A.6 confirms that ZGUL outperforms all baselines (+1.6 F1 on Indo-Aryan, +0.9 on Germanic), with ablation trends consistent with mBERT experiments.

	Indo-Aryan			**Germanic**			
Model	As	Bh	**Avg**	Fo	Got	Gsw	**Avg**
SFT-M	**60.4**	60.5	60.5	78	15.4	54.9	49.4
CPG	55.7	60.3	58	**78.1**	18.3	**57.6**	51.3
ZGUL	59.3	**64.9**	**62.1**	77.8	**21.7**	57.1	**52.2**
ZGUL w/o EM	58.1	61.5	59.8	77.6	15.3	54.9	49.3
ZGUL w/o local fusion	60.3	62.1	61.2	77.6	21	54.7	51.1
ZGUL w/o global fusion	59.7	58.7	59.2	77.8	21.1	54.2	51

Table A.6: **XLM-R robustness of ZGUL**: Indo-Aryan NER (trained on En,Hi,Ur) and Germanic POS (trained on En, Is, De) using XLM-R-Base adapters.

A.3 Few-Shot (FS-CLT) Performance

In this experiment, we take the trained ZGUL and other multi-source models, i.e., CPG, SFT-M, and fine-tune them further for a few labeled examples from the train set of each target language. We do this for those test languages, whose training set is available (12 out of 15). We sample training bins of sizes {0, 25, 50, 75, and 100} samples, with the 0-sample bin corresponding to each model's existing 0-CLT result, providing a consistent anchor point across the few-shot curve.

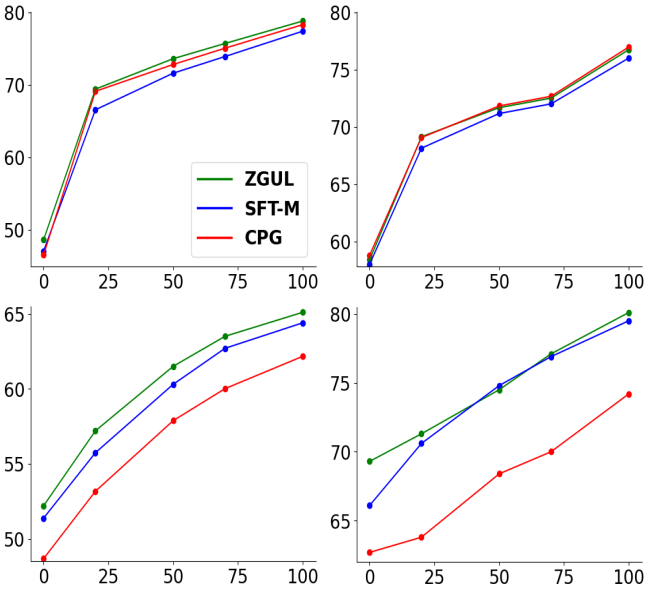


Figure A.1: Few-shot F1 averaged over languages in a group for various few-shot bins. Top row: Germanic, Slavic. Bottom row: African, Indo-Aryan.

We observe in Fig. A.1 that ZGUL scales smoothly for all language groups, maintaining its dominance over the baselines in each case, except for Slavic, where its performance is similar to the CPG baseline. The relative ordering of baselines, i.e. CPG outperforming SFT-M for Slavic and Germanic (POS), and SFT-M outperforming CPG for African and Indo-Aryan (NER), is also maintained across the regime of few-shot samples, similar to the zero-shot setting. The learning plateau is not reached in the curves for either of the language groups, showing that adding more target examples would likely result in further improvement of all the models, albeit at a smaller pace. We present the detailed language-wise few-shot plots as follows.

A.3.1 Language-wise Few Shot Performance

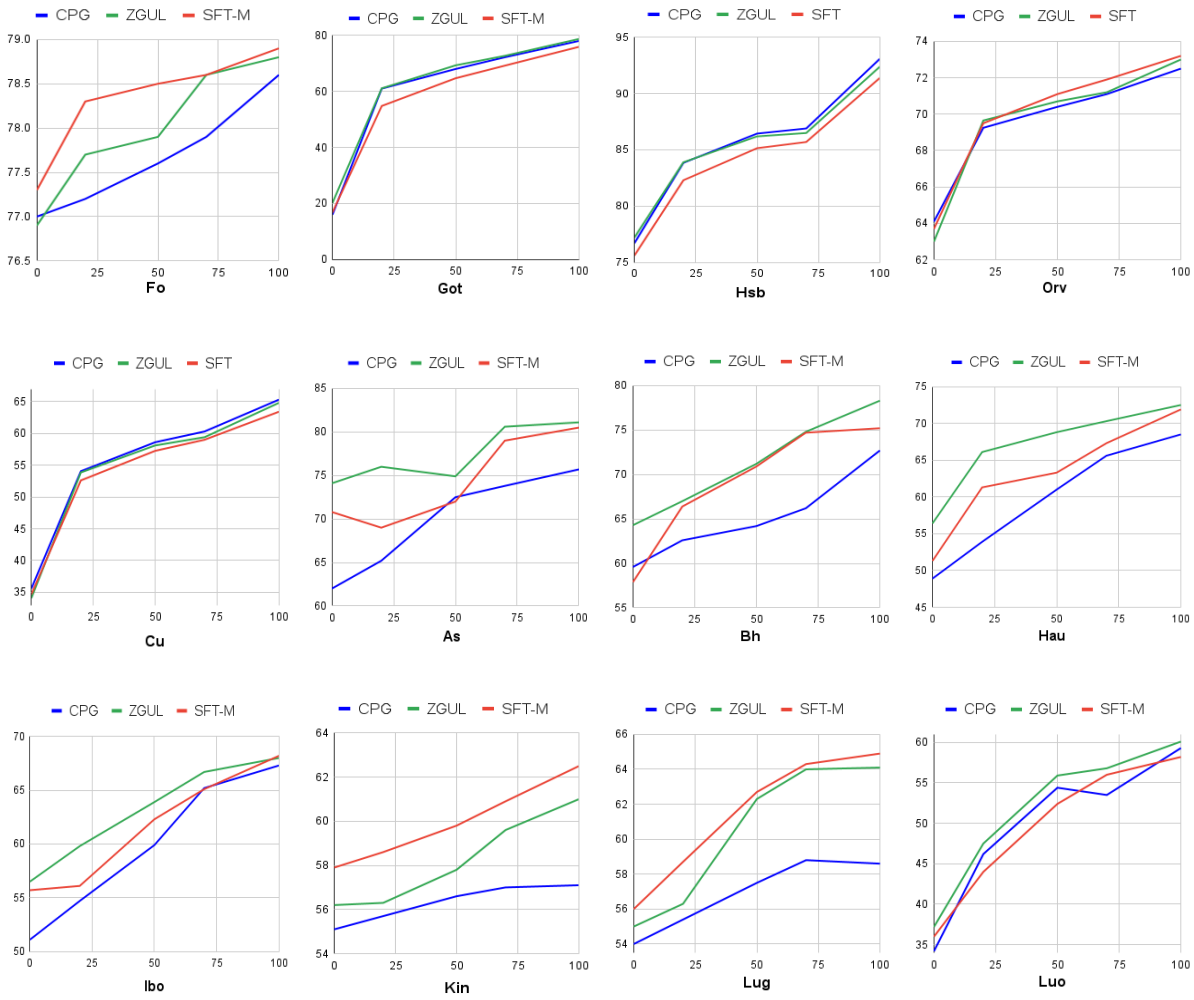


Figure A.2: Language-wise few-shot performance results

Appendix B

Appendix to Chapter 4

B.1 Implementation Details

B.1.1 Estimating Confidence $\hat{p}_i^{(l)}$

We describe how the aggregate label confidence $\hat{p}_i^{(l)}$ is computed for each task type.

Sentence-level classification (NLI). For NLI, each instance predicts a single label $l \in \{\texttt{entailment}, \texttt{neutral}, \texttt{contradiction}\}$. We compute $\hat{p}_i^{(l)}$ as the softmax probability of the predicted class.

Sequence labeling (POS, NER). For a sentence $x_i = \{w_1, \ldots, w_T\}$ with predicted labels

$$\hat{y}_i = \{o_1, \ldots, o_T\} \quad \text{and word-level confidence scores} \quad \{\hat{p}_{i,1}, \ldots, \hat{p}_{i,T}\},$$

where $\hat{p}_{i,t}$ denotes the model confidence for label o_t at word position t, we compute label-wise confidence as:

$$\hat{p}_i^{(l)} = \begin{cases} \dfrac{1}{|\mathcal{T}_{i,l}|} \displaystyle\sum_{t \in \mathcal{T}_{i,l}} \hat{p}_{i,t}, & \text{if } \mathcal{T}_{i,l} \neq \emptyset, \\ 0, & \text{otherwise,} \end{cases}$$

where $\mathcal{T}_{i,l} = \{t \mid o_t = l\}$ denotes the set of token positions assigned label l.

Thus, $\hat{p}_i^{(l)}$ represents the average confidence over all occurrences of label l in exemplar i, while labels not predicted even once in the sequence are assigned the value of 0.

B.1.2 Prompt Details

Prompt templates for all three tasks are presented below. For NLI, we adopt the optimized prompt template from Sclar et al. (2024), which improved macro-F1 by at least 10% over the initial framework of Ahuja et al. (2023) across all tested languages. The same prompt is used across all models.

Natural Language Inference (NLI). **Task Description:** "You are an NLP assistant whose purpose is to solve Natural Language Inference (NLI) problems. NLI is the task of determining the inference relation between two (short, ordered) texts: entailment, contradiction, or neutral. Answer as concisely as possible in the same format as the examples below:"

Input format:

Premise: {premise} , Hypothesis: {hypothesis} ,

Output format:

Answer: {output}

Verbalizer:

match the one-word response from the model (neutral, contradiction or entailment)

Named Entity Recognition (NER). **Task Description:** "Tag the following sentence according to the BIO scheme for the NER task, using the tags PER (person), LOC (location), ORG (organization) and DATE (date). Follow the format specified in the examples below:"

Input format:

Sentence: w_1 w_2 ... w_T

Output format:

Tags:

w_1<TAB>o_1

w_2<TAB>o_2

...

w_T<TAB>o_T

Verbalizer:

Extract the sequence of labels $o_1, o_2, ...o_3$ from generated response.

Part of Speech (PoS) tagging. **Task Description:** "Tag the following sentence according to the Part of Speech (POS) of each word. The valid tags are ADJ, ADP, ADV, AUX, CCONJ, DET, INTJ, NOUN, NUM, PART, PRON, PROPN, PUNCT, SCONJ, SYM, VERB, X. Follow the format specified in the examples below:"

Input format:

Sentence: w_1 w_2 ... w_T

Output format:

Tags:

w_1<TAB>o_1

w_2<TAB>o_2

...

w_T<TAB>o_T

Verbalizer:

Extract the sequence of labels $o_1, o_2, ...o_3$ from generated response.

Verbalizer for Sequence Labeling Tasks. The LLM outputs words alongside labels, but may omit words or produce invalid tags. We align the generated sequence to the input using Longest Common Subsequence (Bergroth et al., 2000), assign corresponding tags to matched words, and default unmatched or invalid positions to `O` (NER) or `X` (POS). For GPT-4, fewer than 1% of words required default assignment.

B.1.3 Dataset Details

Family	Source langs	Train size	Target langs	Avg. labels
Germanic	{En, Is, De}	30,000	{Fo, Got, Gsw}	2,370
African	{En, Am, Sw, Wo}	19,788	{Hau, Ibo, Kin, Lug, Luo}	1,100
American	{En, Es}	19,998	{Aym, Gn, Nah}	501

Table B.1: **Dataset statistics**: source training size (sentences) and average target test size (labels per language). Es: Spanish, Aym - Aymara, Gn: Guarani, Nah: Nahuatl.

B.2 Results on Natural language Inference (NLI)

Table B.2 presents full NLI results on the AmericasNLI (Ebrahimi et al., 2022) dataset.

Impact of Label Coverage. SSP achieves +1.3 average F1 over the ablation without label coverage. The `neutral` class – for which mDeBERTa-FT has poor recall (24.3) – is consistently omitted in selected exemplars without this constraint. Enforcing label coverage yields a 22-point recall gain for `neutral` (57.8 vs. 35.3; Table B.3). Figures B.2 and B.3 illustrate this effect.

Stage 1:

Field	Sentence in target language (Gn)	En translation (GPT-4-turbo stage 1)	En translation (gold)
Premise	Péva ha'e, eikuaáma, emaña, neapañuâima.	That is, you know, look, you're not in trouble anymore. (Error)	This is it, you know, look, you're in trouble.
Hypothesis	Ikatu reñemosê ko tetâgui.	You may be expelled from the country.	You can be deported from this country.
Label	neutral	contradiction	neutral

Stage 2:

Premise: Ha upéichako, akârasy memete, ja'ekuaa ko árape arekopaite mba'érepa cheakârasy haĝua, nde nereikuaái mba'éichapa ojejapo peteî mba'e ha he'i ndéve hikuái: péina, ejapo. , Hypothesis: Ko árape ojerure cheve ajapo haĝua tembiapo che katupyrývape. , Answer: neutral

Premise: upéichaite, ha'eséko che ko'â léi pyahukuéra rehe hasy ko'áĝa. , Hypothesis: Upevarehete, umi temimoîmby pyahu reheve, ko'áĝa hasyve. , Answer: neutral

Premise: Péva ha'e, eikuaáma, emaña, neapañuâima. , Hypothesis: Ikatu reñemosê ko tetâgui. , Answer: **contradiction**

Premise: Néi, ñaĝuahêniko ko'âvape, peteî arapokôindýpe oîha mokôi térâ mbohapy aviô ha ndoikuaái moôpa ovevéta. , Hypothesis: Hetave aviô oîramo upéva apañuâima. , Answer: neutral

Premise: Ha aha hógape ha ahenói upe papapy oje'evakue ahenói haĝua aĝuahê vove upépe. , Hypothesis: Ahenói upe papapýpe aĝuahêvo hógape. , Answer: neutral

Premise: Pe kuñataî ikatúva chepytyvô oî amo táva mboypýri. , Hypothesis: Upe mitâkuña chepytyvôtava oî águi 5km hápe. , Answer: neutral

Premise: Ha'e ou, oipe'a okê ha chemandu'a amaña che rapykuévo ha ahecha hova, ahechakuaa ndaha'éihague upe oha'arôva. , Hypothesis: Oñeha'â ani haĝua roñeñandu vai katu roikuaa orekáusa iñapañuâiha. , Answer: neutral

Premise: Ajeíma upe oje'évagui, aipo peteî kuimba'e oikutihague hembirekópe ojuka peve ha'e oke rupi ambue kuimba'e ndive, hembirekokue ha'eséko, nde reikuaa mba'érepa añe'ê. , Hypothesis: Peteî kuimba'e ojuka hembirekópe oñeno rupi ambue kuimba'e ndive ramoite ojepoi rire chupe upe haguetére. , Answer: contradiction

---

Premise: Ha'ese che ha'ekuéra orekoha amo 5 ñemoñare rupinte. Peteîva omanova'ekue. , Hypothesis: Peteîva umi 5 apytépe omano. , Answer: **entailment**

Figure B.1: **SSP's robustness to translation noise** (shown for Guarani language in NLI task). Stage 1: translate-test degrades due to NLLB translation errors. Stage 2: GPT-4-turbo correctly predicts the label despite a noisy exemplar, demonstrating SSP's ability to mitigate Stage-1 translation noise.

Premise: Ah, huk chaypi allinqa apakurqa allin qawasqayqa paniypa ñawpaq yuyariyninmi, chaypas hina hipa pampapim karqa.
Hypothesis: Yuyaruniqa hipa pampapi huk ima apakusqantam.
Answer: entailment

Premise: Yaykuykuptiykuqa punkukunaqa wichqasqam kachkarqa.
Hypothesis: Punku wichqasqa kachkaptinpas yaykurqanikum.
Answer: entailment

Premise: Yanapawaqniy atiq sispasmi hatun llaqtapa waklawninpiraq tiyan.
Hypothesis: Yanapawaqniy warmi warman 5 millas nisqan karupiraq tiyan.
Answer: **neutral**

Premise: Manam mayman risqanta yacharqanikuchu.
Hypothesis: Mayman risqantam yacharqaniku.
Answer: entailment

Premise: Chayna kaptinqa hamutachkanim huktapiwan Ramonawan rimariyta.
Hypothesis: Ramonawanmi huktapiwan rimarqani.
Answer: entailment

Premise: Ripukusqañam hinaspam amaña llakikunaypaq niwarqa.
Hypothesis: Ama llakikunaytam niwarqa.
Answer: entailment

Premise: Ichapasyá huk kaq mana yachasqaymanta hamun ichaqa
Hypothesis: Apurawtam hamun, ichaqa maymanta hamusqanta yachanim.
Answer: entailment

Premise: Locust Hill oh awriki, ari, kusa
Hypothesis: Locust Hill nisqaqa allinmi.
Answer: contradiction

Premise: Oh, payllam isqun iskay iskayraq regulador nisqapi inyecciónta qinaq karqa.
Hypothesis: Martes punchawtam inyector nisqata hinarqani.
Answer: **neutral**

Premise: Ah, huk chaypi allinqa apakurqa allin qawasqayqa paniypa ñawpaq yuyariyninmi, chaypas hina hipa pampapim karqa.
Hypothesis: Yuyaruniqa hipa pampapi huk ima apakusqantam.
Answer: entailment

Premise: Yaykuykuptiykuqa punkukunaqa wichqasqam kachkarqa.
Hypothesis: Punku wichqasqa kachkaptinpas yaykurqanikum.
Answer: entailment

Premise: Manam mayman risqanta yacharqanikuchu.
Hypothesis: Mayman risqantam yacharqaniku.
Answer: entailment

Premise: Chayna kaptinqa hamutachkanim huktapiwan Ramonawan rimariyta.
Hypothesis: Ramonawanmi huktapiwan rimarqani.
Answer: entailment

Premise: Manam pachay karqachu ima kaqpas ruranaypaq.
Hypothesis: Mana pacha llapan qinanaypaq haypawarqachu
Answer: entailment

Premise: Ripukusqañam hinaspam amaña llakikunaypaq niwarqa.
Hypothesis: Ama llakikunaytam niwarqa.
Answer: entailment

Premise: Ichapasyá huk kaq mana yachasqaymanta hamun ichaqa
Hypothesis: Apurawtam hamun, ichaqa maymanta hamusqanta yachanim.
Answer: entailment

Premise: Locust Hill oh awriki, ari, kusa
Hypothesis: Locust Hill nisqaqa allinmi.
Answer: contradiction

Premise: Oh, payllam isqun iskay iskayraq regulador nisqapi inyecciónta qinaq karqa. Hypothesis: Martes punchawtam inyector nisqata hinarqani.
Answer: **contradiction**

Figure B.2: **Label coverage effect**: SSP with label constraint (left) correctly samples a `neutral` exemplar; without this constraint (right), no `neutral` is selected, causing a misprediction.

Premise: Ah, one there good thing took away is my best view is my sister's old memory, which was also on the same hip floor.
Hypothesis: I remember something carrying on the floor.
Answer: entailment

Premise: The doors were locked when we entered.
Hypothesis: We got in even though the door was locked.
Answer: entailment

Premise: The sister who can help me lives just on the other side of the big city.
Hypothesis: My assistant lives 5 miles away.
Answer: **neutral**

Premise: We didn't know where he was going.
Hypothesis: We knew where he was going.
Answer: entailment

Premise: In that case I'm coming up with another conversation with Ramona.
Hypothesis: I talked to Ramona again.
Answer: entailment

Premise: He had left and told me not to worry.
Hypothesis: He told me not to worry.
Answer: entailment

Premise: Maybe it comes from something I don't know though
Hypothesis: It comes quickly, but I know where it comes from.
Answer: entailment

Premise: Locust Hill oh yeah, yeah, great
Hypothesis: Locust Hill is good.
Answer: contradiction

Premise: Oh, he was the only one who still injected nine seconds into the regulator.
Hypothesis: I applied the injector on Tuesday.
Answer: **neutral**

Premise: Ah, one there good thing took away is my best view is my sister's old memory, which was also on the same hip floor.
Hypothesis: I remember something carrying on the floor.
Answer: entailment

Premise: The doors were locked when we entered.
Hypothesis: We got in even though the door was locked.
Answer: entailment

Premise: We didn't know where he was going.
Hypothesis: We knew where he was going.
Answer: entailment

Premise: In that case I'm coming up with another conversation with Ramona.
Hypothesis: I talked to Ramona again.
Answer: entailment

Premise: I didn't have time to do anything.
Hypothesis: I didn't have enough time to cover everything
Answer: entailment

Premise: He had left and told me not to worry.
Hypothesis: He told me not to worry.
Answer: entailment

Premise: Maybe it comes from something I don't know though
Hypothesis: It comes quickly, but I know where it comes from.
Answer: entailment

Premise: Locust Hill oh yeah, yeah, great
Hypothesis: Locust Hill is good.
Answer: contradiction

Premise: Oh, he was the only one who still injected nine seconds into the regulator. Hypothesis: I applied the injector on Tuesday.
Answer: **contradiction**

Figure B.3: **English translations** for Exemplars shown in Fig. B.2

Model	Aym	Gn	Nah	Avg.
0-CLT				
mDeBERTa-NLI (Laurer et al., 2022)	34.9	43.9	48.9	42.6
mDeBERTa^{FT}	33.9	47	46.9	42.6
ICL-GPT-3.5	38.2	41.7	35.3	38.4
ICL-GPT-4-turbo	32.8	55.8	42.2	43.6
SSP(ICL)-GPT-3.5	38.4	38.8	43.2	40.1
SSP(ICL)-GPT-4-turbo	37.5	58.5	51.8	49.3
SSP(mDeBERTa^{FT})-Llama-2	36.5	37.8	41	38.4
SSP(mDeBERTa^{FT})-GPT-3.5	**43.1**	46	46.8	45.3
SSP(mDeBERTa^{FT})-GPT-4-turbo	36	**61.3**	**59.2**	**52.2**
w/o Conf.	42.9	60.1	50.3	51.1
w/o Label	37	58.2	57.4	50.9
w/o both	34.3	59.7	57.1	50.4
w/o ILP (Random)	33.4	53.8	53.4	46.9
Translate Train				
ICL-GPT-4-turbo	42.4	49.5	-	-
SSP(ICL)-GPT-4-turbo	44.4	58.6	-	-
Translate Test				
ICL-GPT-4-turbo	36.4	45.5	-	-
SSP(ICL)-GPT-4-turbo	42.4	57.6	-	-

Table B.2: **NLI results for SSP** (micro-F1 for three indigenous languages in AmericasNLI dataset). **Bold**: best 0-CLT; underline: overall best. * Nahuatl language not supported in NLLB-200. McNemar's $p = 0.054$ for bold results.

Model	Neu.	Ent.	Con.	Overall
mDeBERTa^{FT}	24.3	**72.7**	38.7	45.2
SSP(mDeBERTa^{FT})	**57.8**	46.5	51.5	52
(w/o Label)	35.3	43.8	68.5	49.2

Table B.3: **Labelwise Recall** for fine-tuned stage 1 model (mDeBERTa^{FT}) and SSP variants – w. and w/o Label coverage (GPT-4-Turbo).

Pool Size	Hau	Ibo	Kin	Lug	Luo	Avg.	Fo	Got	Gsw	Avg
8	68.8	80.2	67.6	75.6	53.8	69.2	80.6	54.8	80.9	72.1
32	70.1	80.8	71.3	74.8	53.9	70.2	81.8	62.8	82.2	75.6
64	70.5	79.7	**72**	**77**	**55.2**	70.9	82.1	63.4	**85.7**	77.1
100	**71.2**	**82.4**	71.4	75.4	55.1	**71.1**	**82.2**	**71.5**	85.6	**79.8**

Table B.4: **Effect of candidate pool size** on SSP's per-language F1.

Appendix C

Appendix to Chapter 5

Relation	Count	Definition (Head → Tail)
/location/location/contains	102	Larger location → smaller location within it
/location/country/admin_divisions	58	Country → subdivision (state, province)
/people/person/place_lived	73	Person → location of residence
/business/company/advisors	37	Company → person (advisor)
/people/person/nationality	34	Person → country
/business/person/company	34	Person → associated company
/business/location	33	Business → location
/location/admin_division/country	31	Administrative division → country
/people/person/place_of_birth	31	Person → birthplace
/business/company/founders	31	Company → person (founder)
/location/neighborhood/neighborhood_of	30	Neighborhood → larger location (city, town)
/location/country/capital	30	Country → capital city
/people/person/children	30	Person → person (child)
/people/deceasedperson/place_of_death	30	Deceased person → location of death
/people/ethnicity/geographic_dist	30	Ethnicity → location of common distribution
/people/person/ethnicity	23	Person → ethnicity
/people/person/religion	21	Person → religion
/location/region/capital	12	Region → capital city
/location/us_county/county_seat	6	US county → county seat
/business/company/majorshareholders	4	Company → person/org (major shareholder)
/people/deceasedperson/place_of_burial	4	Deceased person → burial site
/business/company/place_founded	4	Company → founding location
/time/event/locations	3	Event → location
/film/film/featured_film_locations	1	Film → featured location
NA	30	No relation from the ontology
Total (labels / sentences)	**722 / 538**	

Table C.1: **Relation ontology and INDICRE4 label distribution.** Label count (722) exceeds sentence count (538) due to multi-label instances.

C.1 Data Annotation and Quality Assessment

Annotator details and Protocol: We recruited native speakers (students and IT professionals) for four Indic languages to verify 538 translated sentences per language. Annotators answered

five binary questions assessing sentence translation quality and correctness of head/tail entity translation and projection. Annotators were compensated approximately $60 and informed of research-only usage.

Translation Quality: Automatic translations (IndicTrans2 / Google Translate) were already high quality: 74% of system-generated translations required no correction. For corrected cases, character-level F1 between original and corrected translations averaged 93%, indicating minor edits (Table C.2).

Inter-annotator Agreement: To ensure reliability, a second native speaker independently evaluated 100 random samples per language. Agreement exceeded 90% for translation and entity projection across all languages (Table C.3), confirming annotation reliability.

Language	No Rectification Needed (%)	Char-F1 Match (if rectified)
Oriya	69	92
Santali	72	88
Manipuri	83	96
Tulu	73	95
Average	74	93

Table C.2: **INDICRE4 quality assessment**: percentage of system translations requiring no human rectification, and character-level F1 similarity for rectified translations.

Language	Translation (%)	Head entity projection	Tail entity projection
Oriya	92	92	92
Santali	89	84	87
Manipuri	85	98	100
Tulu	96	92	88
Average	91	92	92

Table C.3: **INDICRE4 inter-annotator agreement** for 100 samples in each language

C.2 Additional Qualitative Examples

Figure C.1: Additional error examples in English and Manipuri

Query:

<Head> **Rudolf Arnheim** </Head>, a distinguished psychologist , philosopher and critic whose work explored the cognitive basis of art – how we interpret it and , by extension , the world – died on Saturday at his home in <Tail> **Ann Arbor** </Tail>, Mich. .

Outputs:

Gold: **place_of_death**
Predicted: **place_lived**

(a) **Proximity Bias:** The local salience of "home" outweighs the distant semantic trigger "died," causing the model to predict residence instead of death place.

Query:

Meitei: ꯁꯥꯏꯔꯥꯛ ꯑꯍ ꯃꯆꯥꯃꯤ ꯑꯣꯟꯁꯁꯑꯣ ꯀꯨꯏꯃꯤꯟ ꯑꯦꯐꯦꯐ ꯋꯥꯍꯪ ꯋꯥꯏ ꯍꯤꯡ ꯋꯥꯏ ꯐ ꯁꯣꯔꯤ ꯋꯥꯏꯃꯤꯍ ꯑꯣꯏꯍꯅꯒꯤ ꯂꯩꯕꯥ ꯋꯥꯌꯦꯜ ꯄꯥꯞꯄꯤ ꯃꯐꯝꯁꯤꯡ ꯋꯥꯃ ꯋꯥꯍꯩꯁꯥ ꯋꯥꯌꯦꯜ **ꯋꯦꯖꯍꯃ ꯌꯦꯁꯤ ꯍꯧꯃꯁꯤꯟꯂꯤꯡꯒꯤ** ꯈꯨꯗ ꯀꯧꯔꯥꯛ ꯁꯣꯜꯂꯤ ꯃꯁꯤꯒꯤ ꯂꯥꯏꯁꯣꯜ ꯑꯃꯁꯨꯡ ꯑꯁꯤ ꯋꯥꯌꯦꯜ ꯍꯧꯃꯁꯤꯟꯂꯤꯡꯒꯤ ꯑꯣꯏꯍꯅꯒꯤ ꯃꯇꯧ ꯀꯣꯛꯄꯤꯔꯝꯕꯥ ꯑꯣꯏꯔꯕꯗꯤ ꯃꯁꯤꯒꯤ ꯃꯇꯨꯡ ꯏꯟꯗꯨꯅ ꯐꯤꯚꯃꯁꯤꯡ ꯑꯣꯏꯔꯕꯥ ꯑꯃꯁꯨꯡ ꯃꯁꯤꯒꯤ ꯃꯇꯨꯡ ꯏꯟꯗꯨꯅ ꯑꯣꯏꯔꯕꯥ ꯃꯐꯝ ꯑꯁꯤ ꯋꯥꯌꯦꯜ ꯍꯧꯃꯁꯤꯟꯂꯤꯡꯒꯤ ꯃꯇꯧ ꯀꯣꯛꯄꯤꯔꯝꯕꯥ ꯑꯣꯏꯔꯕꯗꯤ

Appendix D

Appendix to Chapter 6

D.1 Prompt Details

Task Description: Choose all applicable relations between head and tail entities from the set below. Print each relation in a new line. If none of the relations are applicable, output "NA".
/people/person/nationality : head entity is a person and tail entity is a country
/time/event/locations : head entity is an event and tail entity is a location
/people/person/children : head entity is a person and tail entity is another person (child)
/business/company/advisors : head entity is a company and tail entity is a person (advisor)
/business/location : head entity is a business and tail entity is a location
/business/company/majorshareholders : head entity is a company and tail entity is a person or organization (major shareholder)
/people/person/place_lived : head entity is a person and tail entity is a location
/business/company/place_founded : head entity is a company and tail entity is a location
/location/neighborhood/neighborhood_of : head entity is a neighborhood and tail entity is a larger location (city, town)
/people/deceasedperson/place_of_death : head entity is a deceased person and tail entity is a location
/film/film/featured_film_locations : head entity is a film and tail entity is a location
/location/region/capital: head entity is a region and tail entity is a city (capital)
/business/company/founders : head entity is a company and tail entity is a person (founder)
/people/ethnicity/geographic_distribution : head entity is an ethnicity and tail entity is a location where the ethnicity is commonly found
/location/country/administrative_divisions : head entity is a country and tail entity is a subdivision (state, province)
/people/deceasedperson/place_of_burial : head entity is a deceased person and tail entity is a burial site
/location/country/capital : head entity is a country and tail entity is a city (capital)

/business/person/company : head entity is a person and tail entity is a company they are associated with
/location/location/contains : head entity is a larger location and tail entity is a smaller location within it
/location/administrative_division/country : head entity is an administrative division (state, province) and tail entity is a country
/location/us_county/county_seat: head entity is a U.S. county and tail entity is the county seat
/people/person/religion: head entity is a person and tail entity is a religion
/people/person/place_of_birth : head entity is a person and tail entity is a location (birthplace)
/people/person/ethnicity: head entity is a person and tail entity is an ethnicity
NA : no relation from the set exists between the given entity pair
Input format:
Input: {sentence}
Output format:
Output: {relation}
Verbalizer: Extract the relation based on string match
A sample input/output format of exemplars can be seen in Figure D.3a.

D.2 Additional Results

D.2.1 Detailed Language-wise Results

Tables D.1, D.2, and D.3 report per-language results for 0-CLT, translate-train, and translate-test settings respectively.

D.2.2 Ablations to Assess Stage 3 Design

The final stage of HYDRE selects representative sentences by aggregating confidence scores across all bag-level labels, and assigns the full relation set to each chosen exemplar in the prompt. This strategy favors sentences with strong multi-label evidence over those with potentially noisy, single-label signals.

To evaluate this design, we consider two ablations:

1. **Candidate-only scoring:** Sentences are selected based solely on confidence for the target candidate relation ($r \in \mathcal{R}'$), ignoring other bag labels. The chosen exemplars still retain their full original bag relation set in the prompt.

2. **Single-label prompting:** Building on the candidate-only selection above, exemplars are presented in the prompt with *only* the target candidate label, discarding the rest of the bag's relation set.

Model	English	Orya	Santali	Manipuri	Tulu	Avg. (Micro/Macro)†
Supervised						
HICLRE (Li et al., 2022)	31/18	–	–	–	–	–
HFMRE (Li et al., 2023)	33/18	–	–	–	–	–
CIL	43/32	–	–	–	–	–
TransRel	42/26	–	–	–	–	–
PARE	42/31	–	–	–	–	–
zero-shot						
Qwen3-235B-A22B	49/40	46/40	6/5	2/1	45/39	25/21
Llama3.1-8b	31/17	26/21	21/14	10/7	29/21	22/16
Llama3.1-8B-FT$_{En}$	60/44	33/23	15/8	11/4	36/21	24/14
GPT-4o	56/57	56/56	10/5	11/7	55/**55**	33/31
5-shot						
HYDRE(QWEN3-235B-A22B)	**63*/62**	52/46	9/7	2/2	53/48	29/26
HYDRE(LLAMA3.1-8B)	52/47	37/24	**24*/16**	**21*/11**	38/25	30/19
HYDRE(LLAMA3.1-8B-FT)$_{En}$	61/45	47/32	**24***/13	**21***/8	49/31	35/21
HYDRE(GPT-4O)	**63***/60	**57/57**	18/10	11/8	**56/55**	**36*/33**

Table D.1: **HYDRE's language-wise results (0-CLT)**: micro/macro-F1 on English and Indic languages. †Averages over non-English languages. *McNemar's $p < 10^{-5}$.

As shown in Table D.4, candidate-only scoring remains competitive for GPT-4o, but aggregate scoring substantially improves smaller models such as Qwen3 and Llama3.1. For Llama3.1, macro-F1 increases from 35 to 47, indicating that label-aware selection stabilizes performance in weaker architectures. Additionally, restricting exemplars to a single label noticeably degrades GPT-4o's performance (Micro-F1: 63 → 59), demonstrating the importance of preserving the full bag relations in the prompt.

D.2.3 Additional Qualitative Analysis

We present the confusion matrices for 12 manually identified, semantically confusing relation pairs for English NYT-10m in Figures D.1 and D.2. A few working and error examples of HYDRE are presented in Figures D.3a – D.6b.

HYDRE's multi-label coverage. To test HYDRE's multi-label efficacy, we measure recall specifically on multi-label queries (Table D.5). HYDRE improves micro- and macro-recall over zero-shot prompting by +9 and +11 points, respectively, indicating it actively recovers multiple relations.

We attribute this broad coverage to HYDRE's Stage 3 design, which retains all bag relations in the prompt. Comparing HYDRE against the single-label prompting ablation (§D.2.2) confirms this behavior (Table D.6). While single-label prompting yields a marginal 1-point micro-F1 gain on single-label queries, it triggers a severe 12-point drop on multi-label queries.

Given that relation extraction is inherently a multi-label problem, HYDRE's ability to preserve multi-label signals in the prompt is critical for robust overall performance.

Model	ory_Orya	sat_Olck	mni_Mtei	tcy_Tulu	Avg. (Micro)	Avg. (Macro)
Supervised						
CIL	29/18	22/15	25/19	29/20	26	18
TransRel	25/16	23/14	21/14	27/18	24	16
PARE	31/20	29/20	28/19	30/19	30	20
zero-shot						
Qwen3-235B-A22B	46/40	6/5	2/1	45/39	25	21
Llama3.1-8b	26/21	21/14	10/7	29/21	22	16
GPT-4o	56/**56**	8/6	8/6	56/53	32	30
Llama3.1-8B-FT$_{En}$	33/23	10/7	2/1	36/21	20	13
Llama3.1-8B-FT$_{X}$	51/37	33/18	24/11	44/26	38	23
5-shot						
HYDRE(Qwen3-235B-A22B)	54/52	23/14	15/5	57/54	37	31
HYDRE(Llama3.1-8B)	36/24	25/18	24/13	39/26	31	20
HYDRE(GPT-4o)	**57/56**	20/11	21/14	**58/57**	39	**35**
HYDRE(Llama3.1-8B-FT$_{En}$)	48/33	27/15	21/8	47/31	36	22
HYDRE(Llama3.1-8B-FT$_{X}$)	56/40	**39*/23**	**34*/15**	50/34	**45***	28

Table D.2: **HYDRE's language-wise results (translate-train)**: micro/macro-F1 on Indic languages. *McNemar's $p < 10^{-5}$.

Model	ory_Orya	sat_Olck	mni_Mtei	tcy_Tulu	Avg. (Micro/Macro)
Supervised					
HiCLRE	27/15	23/13	25/15	22/14	24/14
HFMRE	27/15	25/14	29/18	28/17	27/16
CIL	34/23	33/21	34/25	36/25	34/24
TransRel	37/23	35/23	33/20	34/21	35/22
PARE	35/25	31/21	33/25	32/22	33/23
zero-shot					
Qwen3-235B-A22B	44/37	43/34	43/35	44/34	44/35
Llama3.1	29/25	26/22	29/25	30/26	29/25
Llama3.1-8B-FT$_{En}$	49/35	45/29	49/35	52/39	49/34
GPT-4o	53/51	50/45	55/**57**	56/51	54/51
5-shot					
HYDRE(Qwen3-235B-A22B)	52/46	51/42	55/47	56/48	54/46
HYDRE(Llama3.1-8B)	46/38	41/31	46/38	44/38	44/36
HYDRE(Llama3.1-8B-FT$_{En}$)	51/37	50/33	50/37	54/40	51/37
HYDRE(GPT-4o)	**55*/54**	**53*/49**	**56/57**	**59*/55**	**56*/54**

Table D.3: **HYDRE's language-wise results (translate-test)**: micro/macro-F1 on Indic languages. *McNemar's $p < 10^{-5}$.

The Missing Candidate Bottleneck. Finally, we explicitly evaluate HYDRE's resilience when the gold relation is absent from PARE's top-5 candidates, breaking down performance by single- and multi-label queries (Table D.7).

Unsurprisingly, HYDRE consistently outperforms both baselines when the gold label is present in the top-5. However, even when the gold label is missing, HYDRE maintains meaningful performance. While it trails zero-shot prompting on single-label queries in this regime—likely due to the LLM grounding on incorrectly retrieved exemplars—this scenario accounts for only ~14% of all queries. Crucially, for multi-label queries where the gold label is missing, HYDRE still

Ablation Variant	GPT-4o	Qwen3	Llama3.1
HYDRE (Aggregate Scoring)	**63**/60	**63/62**	**52/47**
Candidate-only Scoring	**63/62**	62/60	47/35
+ Single-Label Prompting	59/58	58/56	41/41

Table D.4: **HYDRE's stage 3 design ablation** (micro/macro-F1, English NYT-10m).

Method	Micro Recall	Macro Recall
PARE	24	19
0-shot	38	30
HYDRE	**47**	**41**

Table D.5: **Multi-label coverage.** Comparison of recall metrics on multi-label queries, demonstrating HYDRE's ability to actively recover multiple relations.

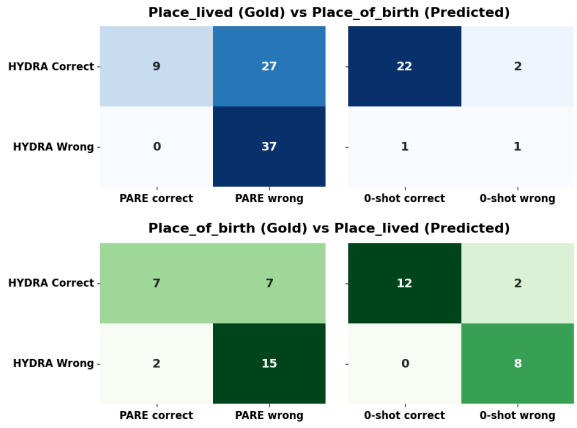
surpasses zero-shot prompting (+3 F1). This demonstrates that exemplar grounding remains beneficial for complex queries even when the initial DSRE candidate recall is imperfect.

Method	Single-label	Multi-label
0-shot prompting	59	52
single-label prompting	**65**	49
HYDRE	64	**61**

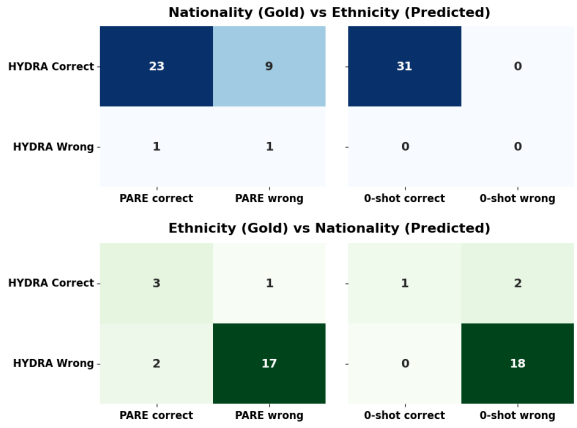
Table D.6: **Impact of single-label prompting.** Performance breakdown (Micro-F1) of GPT-4o on single-label vs. multi-label queries when restricting exemplars to a single relation label.

	Single-Label		Multi-Label	
Method	**Micro-F1**	**Macro-F1**	**Micro-F1**	**Macro-F1**
Gold in top-5 (427 queries: 323 single-label, 104 multi-label)				
PARE	49.6	27.5	43.0	16.8
0-shot	63.7	48.6	54.5	23.5
HYDRE	**69.6**	**53.6**	**63.1**	**31.5**
Gold out of top-5 (111 queries: 74 single-label, 37 multi-label)				
PARE	0.0	0.0	0.0	0.0
0-shot	**42.3**	**15.9**	51.6	30.6
HYDRE	27.5	10.6	**54.5**	**34.6**

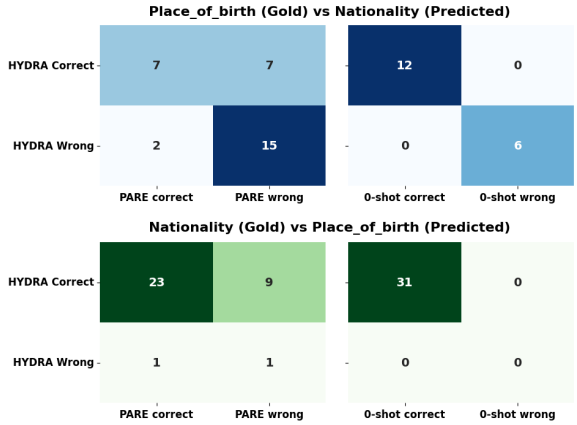
Table D.7: **Impact of DSRE candidate recall.** Performance breakdown based on whether the gold relation is present in PARE's top-5 retrieved candidates.



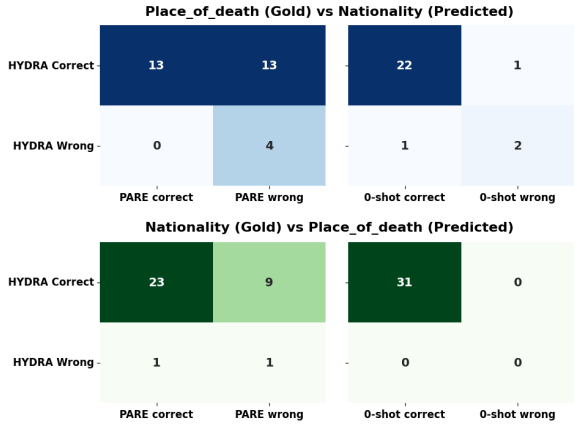
a Place of birth vs Place lived



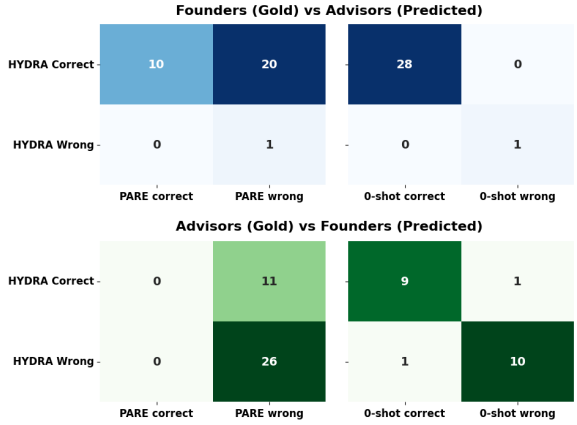
b Nationality vs Ethnicity



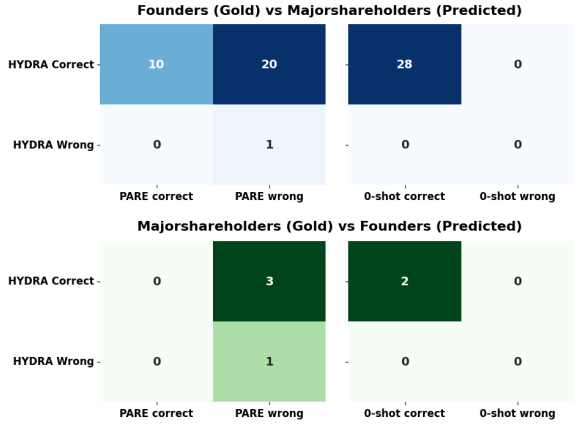
c Nationality vs Place of Birth



d Nationality vs Place of Death

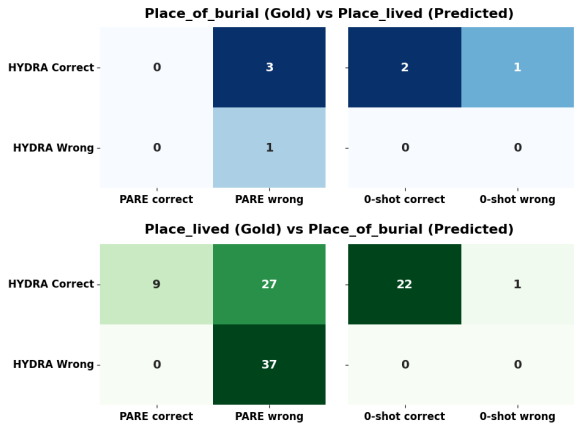


e Company founders vs Company Advisors

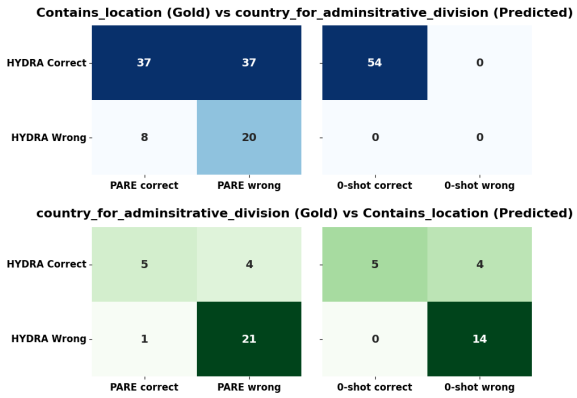


f Company founders vs Company Majorshareholders

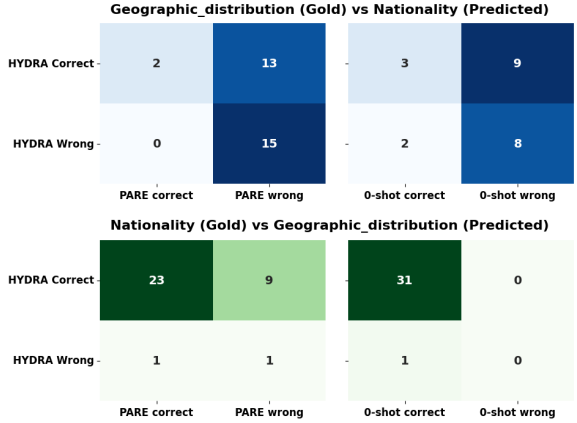
Figure D.1: **Confusion matrices for HYDRE vs. the baselines (part 1 of 2)**.



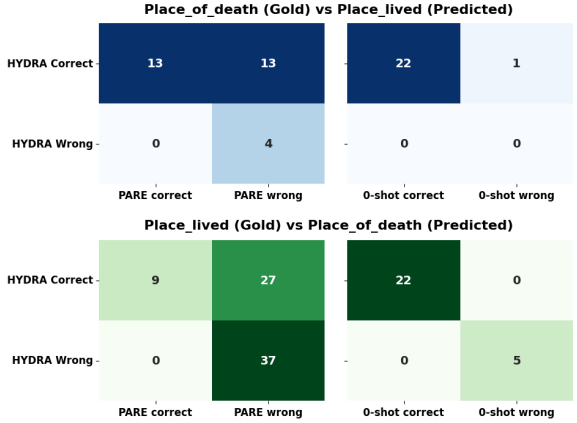
a Place of Burial vs Place lived



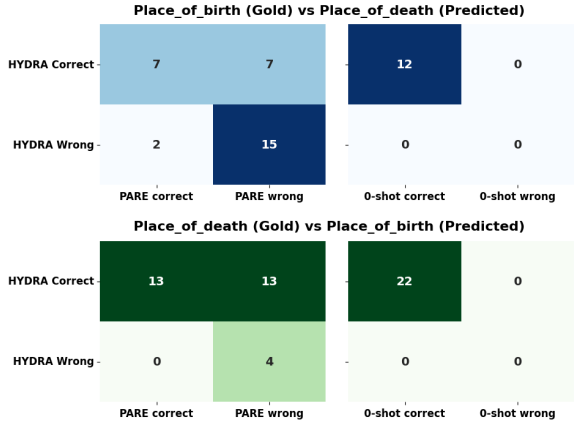
b Country for Administrative divisions vs Contains



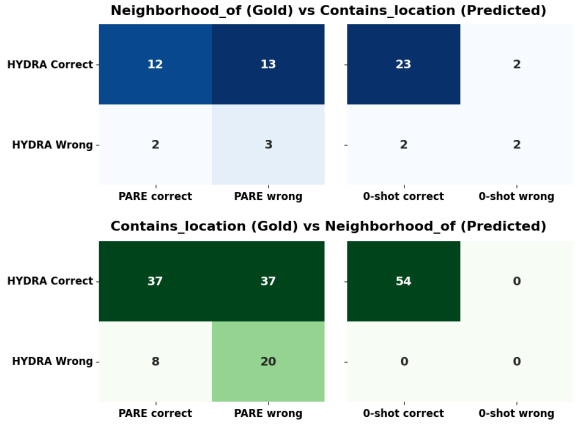
c Nationality vs Geographic distribution



d Place lived vs Place of Death



e Place of Birth vs Place of Death



f Neighborhood_of vs Contains_location

Figure D.2: **Confusion matrices for HYDRE vs. the baselines (part 2 of 2).**

ICL Exemplars (HYDRE)

Input: Mr. Marek started the <Head> Fairfield </Head> Theater Company in 2001 ... at <Tail> Fairfield University </Tail> .
Output: /location/location/contains

Input: But Mr. Wallace , 45 , a businessman and an investor in partnership for a time with <Tail> Mark Thatcher </Tail> , son of the former British Prime Minister <Head> Margaret Thatcher </Head> , has detractors , including ... . ”
Output: /people/person/children

Input: Behind the News – Founded eight years ago in a Silicon Valley garage by two <Tail> Stanford University </Tail> graduate students , <Head> Google </Head> went public two years ago at $ 85 a share .
Output: /business/company/place_founded

Input: But <Tail> Chicago </Tail> , the birthplace of <Head> Curtis Mayfield </Head> and Quincy Jones and the locus of King Oliver ’s Creole Jazz Band , has its own clanging , speakeasy sway .
Output: /people/person/place_of_birth

Input: This ocean of gray stone is <Tail> Paris </Tail> ’s most literary cemetery , holding the graves of Maupassant , Beckett , Duras and Man Ray , as well as arrivistes like Susan Sontag and <Head> Serge Gainsbourg </Head> .
Output: /people/person/place_lived

Query:

Input: The restaurant is across the street from <Tail> Atlanta </Tail> ’s oldest and loveliest cemetery , Oakland Cemetery (404-688-2107 ; www.oaklandcemetery.com) , where <Head> Margaret Mitchell </Head> , 25 mayors and thousands of unidentified Confederate soldiers are buried .

Outputs:

Gold: place_of_burial
HYDRE: **place_of_burial**
0-shot: **place_lived**
PARE: **place_lived**

(a) **Example 1:** HYDRE correctly predicts “place_of_burial” – absent from PARE’s top-5 – while 0-shot and PARE predict “place_lived”.

ICL Exemplars (HYDRE)

Input: ... Mr. Koppel to take another look at a once-unknown man , <Head> Morrie Schwartz </Head> , a <Tail> Brandeis University </Tail> professor who ... .
Output: person/company

Input: ... as you race away from the pleasant corporate maw of <Tail> Seattle </Tail> , from Starbucks and <Head> Boeing </Head> , Amazon and Microsoft .
Output: company/place_founded

Input: ... <Head> Ernest Hemingway </Head> was born in <Tail> Oak Park </Tail> in 1899 and lived here through high school .
Output: place_of_birth, place_lived

Input: ... Mr. Narayanan ’s body will be cremated ... in <Tail> New Delhi </Tail> , near the funeral ground of <Head> Jawaharlal Nehru </Head> , India ’s first prime minister ... .
Output: place_of_death, place_lived

Input: The easiest ... way to see <Tail> Philadelphia </Tail> is to stick with the older , central parts of town , emulate <Head> Benjamin Franklin </Head> ...
Output: place_of_death, place_lived

Query:

Input: Poe , Evermore A mystery man arrived at <Head> Edgar Allan Poe </Head> ’s grave at the Westminster Burial Grounds in <Tail> Baltimore </Tail> on Friday morning , as he has on Poe ’s birthday (Jan. 19) every year since 1949 , ...

Outputs:

Gold: place_lived, place_of_burial
HYDRE: **place_of_burial, place_lived**
0-shot: **place_of_burial**
PARE: **place_lived**

(b) **Example 2:** HYDRE recovers both “place_of_burial” and “place_lived”; 0-shot and PARE each capture only one.

ICL Exemplars (HYDRE)

Input: ... Mendelssohn , who was born Jewish and converted to Christianity , and <Head> Otto Klemperer </Head> , who converted to Christianity and then back to <Tail> Judaism </Tail> .
Output: /religion

Input: As the <Head> <Tail> Baltimore </Tail> Orioles </Head> return home ... , Baltimore embraces its rich sports and maritime history .
Output: /business/location

Input: <Head> Lorenzo Da Ponte </Head> , a Bridge From <Tail> Italy </Tail> to New York " includes three vocal recitals , beginning tonight with ... .
Output: /nationality

Input: The Museum of Modern Art 's exhibition of four films starring the <Tail> Italian </Tail> actress <Head> Laura Morante </Head> concludes this weekend with four films , including ... .
Output: /ethnicity

Input: ... <Head> Madhesi </Head> ethnic group , which by some estimates represents as much as a third of <Tail> Nepal </Tail> 's population of 29 million , has been granted citizenship rights ... .
Output: /geographic_distribution

Query:

Input: Among the performances of note : ... the <Head> Italian </Head> dancer ALESSANDRA FERRI gives her final performance with the company on Saturday night , with ROBERTO BOLLE , a guest artist also from <Tail> Italy </Tail> .

Outputs:

Gold: /geographic_distribution, /nationality
HYDRE: **/geographic_distribution**
0-shot: **/nationality**
PARE: **NA**

(a) **Example 3:** HYDRE and 0-shot each partially predict one of two gold labels (`geographic_distribution` vs. `nationality`).

ICL Exemplars (HYDRE)

Input: ... owned by a <Tail> Cincinnati </Tail> company , American Financial Group , whose chairman and chief executive officer is Carl H. Lindner III , who is also an owner of the <Head> Cincinnati Reds </Head> .
Output: /business/location

Input: A biomedical research institute in <Tail> Chengdu </Tail> , <Head> China </Head> , is planning to show true commitment to scientific principles ... .
Output: /location/location/contains /location/country/administrative_divisions

Input: <Tail> Robert Bigelow </Tail> , the founder of <Head> Budget Suites of America </Head> , is likely to push forward ... .
Output: /founders

Input: ... the 777 's , 767 's and 757 's are often coveted by corporate titans , among them Larry Page and <Tail> Sergey Brin </Tail> , the co-founders of <Head> Google </Head> , who bought ... .
Output: /founders
/majorshareholders

Input: The N.F.L. ... is very popular , ... , " said Tony Ponturo , vice president for global media and sports marketing at <Head> Anheuser-Busch </Head> in <Tail> St. Louis </Tail> , ... "
Output: /place_founded

Query:

Input: <Head> Nestlé </Head> , based in <Tail> Vevey </Tail> , Switzerland , ...

Outputs:

Gold: **/business/location**
HYDRE: **/place_founded**
0-shot: **/place_founded**
PARE: **NA**

(b) **Example 4:** Both HYDRE and 0-shot misclassify `business/location` as `place_founded`.

ICL Exemplars (HYDRE)

Input: ... the epic poem " Paterson " by <Head> William Carlos Williams </Head> , a native of <Tail> Rutherford </Tail> .
Output: /people/person/place_of_birth

Input: It goes on to list notable <Tail> Mississippi </Tail> writers including William Faulkner , <Head> Richard Wright </Head> , ...
Output: /people/person/place_lived

Input: ... , including Giocangga , the founder of the <Head> Manchu </Head> dynasty in <Tail> China </Tail> , and Niall of the Nine Hostages , ... "
Output: /people/ethnicity/geographic_distribution

Input: <Head> Anthony Trollope </Head> , the brilliant depicter of the 19th-century social strata in <Tail> England </Tail> , ...
Output: /people/person/nationality

Input: ... , Mr. Delli Colli worked with generations of <Tail> Italian </Tail> directors , including <Head> Pier Paolo Pasolini </Head> , ...
Output: /people/person/ethnicity

Query:

Input: Lukacs , ... , makes very large claims for his subject in " George Kennan " : He was " a better writer and a better thinker " than <Head> Henry Adams </Head> ; he was " the best and finest <Tail> American </Tail> writer about Europe " in the interwar years ... .

Outputs:

Gold: **/people/person/ethnicity**
HYDRE: **/people/person/nationality**
0-shot: **/people/person/nationality**
PARE: **NA**

(a) **Example 5:** Both HYDRE and 0-shot confuse `ethnicity` with `nationality`.

ICL Exemplars (HYDRE)

Input: Mr. White , 54 , of <Tail> Centerport </Tail> , has held top environmental posts with <Head> Suffolk County </Head> and the Town of Huntington .
Output: /location/location/contains

Input: Being able to combine Loudeye services with <Head> Nokia </Head> terminals provides a good base for innovative and useful consumer services , " said Ilkka Raiskinen , senior vice president for multimedia experiences at Nokia , based in <Tail> Espoo </Tail> , Finland .
Output: /business/company/place_founded

Input: ... ' It 's a gentle soul with a naughty sense of humor , ' Jerry was just as responsible for that as my dad , " ' said <Tail> Brian Henson </Tail> , <Head> Jim Henson </Head> 's son , who serves with his sister Lisa as chairman and chief executive of the Jim Henson Company .
Output: /people/person/children

Input: ... the 777 's , 767 's and 757 's are often coveted by corporate titans , among them Larry Page and <Tail> Sergey Brin </Tail> , the co-founders of <Head> Google </Head> , who bought a used 767 last year and spent millions converting it into a private jet .
Output: /business/company/founders, /business/company/majorshareholders

Input: The service has more ways to shield users ' identities compared with many social networking sites , said <Tail> Antony Brydon </Tail> , <Head> Visible Path </Head> 's chief executive , above . "
Output: /business/company/founders

Query:

Input: " <Head> MySpace </Head> is dedicated to ensuring that content owners , whether large or small , can both promote and protect their content in our community , " <Tail> Chris DeWolfe </Tail> , the chief executive of MySpace , said in a statement . "

Outputs:

Gold: **/business/company/advisors**
HYDRE: **/business/company/advisors**
0-shot: **/business/person/company**
PARE: **NA**

(b) **Example 6:** HYDRE correctly predicts `company/advisors`; 0-shot and PARE fail.

ICL Exemplars (HYDRE)

Input: ꯃꯤꯁꯛ ꯃꯅꯨꯄꯥ ꯑꯦꯛ꯫ ꯴ ꯑꯣꯟꯇꯔꯦꯛꯄꯣꯔ ꯀꯣ 2001ꯅ ꯃꯇꯝ ꯃꯆꯥꯅꯨꯄꯥ ꯃꯍꯥꯛꯀꯥ ꯃꯤꯑꯣꯏꯕ ꯑꯣꯟꯊꯣꯛꯇꯥ ꯁꯤꯖꯤꯟꯅꯕ ꯑꯣꯟꯇꯔꯦꯛꯄꯣꯔ ꯑꯃꯁꯨꯡ ꯃꯍꯥꯛꯀꯥ ꯑꯣꯟꯊꯣꯛꯇꯥ ꯃꯤꯑꯣꯏ

Bibliography

J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 43, 50

D. I. Adelani, J. Abbott, G. Neubig, D. D'souza, J. Kreutzer, C. Lignos, C. Palen-Michel, H. Buzaaba, S. Rijhwani, S. Ruder, et al. Masakhaner: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131, 2021. 2, 31

R. Agerri, Y. Chung, I. Aldabe, N. Aranberri, G. Labaka, and G. Rigau. Building named entity recognition taggers via parallel corpora. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1557/. 10

P. Aggarwal, A. Madaan, A. Anand, S. P. Potharaju, S. Mishra, P. Zhou, A. Gupta, D. Rajagopal, K. Kappaganthu, Y. Yang, S. Upadhyay, M. Faruqui, and M. Mausam. Automix: Automatically mixing language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 131000–131034. Curran Associates, Inc., 2024. doi: 10.52202/079017-4164. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ecda225cb187b40ea8edc1f46b03ffda-Paper-Conference.pdf. 20, 96

S. Agrawal, C. Zhou, M. Lewis, L. Zettlemoyer, and M. Ghazvininejad. In-context examples selection for machine translation. In *Annual Meeting of the Association for Computational Linguistics*, 2022. URL https://api.semanticscholar.org/CorpusID:254246450. 45

K. Ahuja, H. Diddee, R. Hada, M. Ochieng, K. Ramesh, P. Jain, A. Nambi, T. Ganu, S. Segal, M. Ahmed, et al. Mega: Multilingual evaluation of generative ai. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, 2023. 5, 6, 9, 18, 19, 43, 44, 49, 94, 103

J. O. Alabi, D. I. Adelani, M. Mosbach, and D. Klakow. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi,

P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, and S.-H. Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea, Oct. 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.382/. 3, 13

C. Alt, M. Hübner, and L. Hennig. Fine-tuning pre-trained transformer language models to distantly supervised relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1388–1398, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1134. URL https://aclanthology.org/P19-1134. 62, 67

H. Amini and L. Kosseim. How (un)faithful is attention? In J. Bastings, Y. Belinkov, Y. Elazar, D. Hupkes, N. Saphra, and S. Wiegreffe, editors, *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 119–130, Abu Dhabi, United Arab Emirates (Hybrid), Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.blackboxnlp-1.10. URL https://aclanthology.org/2022.blackboxnlp-1.10/. 38

A. Ansell, E. M. Ponti, J. Pfeiffer, S. Ruder, G. Glavaš, I. Vulić, and A. Korhonen. MAD-G: Multilingual adapter generation for efficient cross-lingual transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4762–4781, Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.410. URL https://aclanthology.org/2021.findings-emnlp.410. 15, 27, 31

A. Ansell, E. Ponti, A. Korhonen, and I. Vulić. Composable sparse fine-tuning for cross-lingual transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1778–1796, 2022. 31

M. Artetxe, S. Ruder, and D. Yogatama. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, 2020. 17

M. Artetxe, V. Goswami, S. Bhosale, A. Fan, and L. Zettlemoyer. Revisiting machine translation for cross-lingual classification. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6489–6499, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.399. URL https://aclanthology.org/2023.emnlp-main.399/. 11

A. Asai, S. Kudugunta, X. Yu, T. Blevins, H. Gonen, M. Reid, Y. Tsvetkov, S. Ruder, and H. Hajishirzi. BUFFET: Benchmarking large language models for few-shot cross-lingual transfer. In K. Duh, H. Gomez, and S. Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1771–1800, Mexico City, Mexico, June 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.100. URL https://aclanthology.org/2024.naacl-long.100/. 18, 43, 44, 45

A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=hSyW5go0v8. 20

K. Badola, S. Dave, and P. Talukdar. Parameter-efficient finetuning for robust continual multilingual learning. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9763–9780, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.619. URL https://aclanthology.org/2023.findings-acl.619/. 13

D. Bareket and R. Tsarfaty. Neural modeling for named entities and morphology (nemo2). *Transactions of the Association for Computational Linguistics*, 9:909–928, 09 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00404. URL https://doi.org/10.1162/tacl_a_00404. 2

M. S. Bari, T. Mohiuddin, and S. Joty. UXLA: A robust unsupervised data augmentation framework for zero-resource cross-lingual NLP. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1978–1992, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.154. URL https://aclanthology.org/2021.acl-long.154/. 12

N. Bel, C. H. Koster, and M. Villegas. Cross-lingual text categorization. In *International conference on theory and practice of digital libraries*, pages 126–139. Springer, 2003. 11

E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021. 1, 2

L. Bergroth, H. Hakonen, and T. Raita. A survey of longest common subsequence algorithms. In *Proceedings Seventh International Symposium on String Processing and Information Retrieval. SPIRE 2000*, pages 39–48, 2000. doi: 10.1109/SPIRE.2000.878178. 105

A. Bhartiya, K. Badola, and Mausam. DiS-ReX: A multilingual dataset for distantly supervised relation extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Dublin, Ireland, May 2022. Association for Computational Linguistics. 7, 23, 67, 68, 69, 77

D. Blasi, A. Anastasopoulos, and G. Neubig. Systematic inequalities in language technology performance across the world's languages. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.376. URL https://aclanthology.org/2022.acl-long.376/. 1

T. Blevins, H. Gonen, and L. Zettlemoyer. Analyzing the mono- and cross-lingual pretraining dynamics of multilingual language models. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of*

the 2022 Conference on Empirical Methods in Natural Language Processing, pages 3575–3590, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.234. URL https://aclanthology.org/2022.emnlp-main.234/. 4

T. Blevins, H. Gonen, and L. Zettlemoyer. Prompting language models for linguistic structure. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6649–6663, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.367. URL https://aclanthology.org/2023.acl-long.367/. 5, 6, 18, 44

P. Borchert, I. Vulić, M. F. Moens, and J. De Weerdt. Language fusion for parameter-efficient cross-lingual transfer. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25848–25868, 2025. 16

S. Brandl and N. Hollenstein. Every word counts: A multilingual analysis of individual human alignment with model attention. In Y. He, H. Ji, S. Li, Y. Liu, and C.-H. Chang, editors, *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 72–77, Online only, Nov. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.aacl-short.10. URL https://aclanthology.org/2022.aacl-short.10/. 37

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546. 1, 2, 5, 9, 43, 45

T. Byrt, J. Bishop, and J. B. Carlin. Bias, prevalence and kappa. *Journal of clinical epidemiology*, 46(5): 423–429, 1993. 66

S. Chakrabarti, H. Singh, S. Lohiya, P. Jain, and M. . Joint completion and alignment of multilingual knowledge graphs. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11922–11938, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.817. URL https://aclanthology.org/2022.emnlp-main.817/. 97

E. C. Chau, L. H. Lin, and N. A. Smith. Parsing with multilingual BERT, a small corpus, and a small treebank. In T. Cohn, Y. He, and Y. Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1324–1334, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.118. URL https://aclanthology.org/2020.findings-emnlp.118/. 3

J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024a. 19, 81

L. Chen, M. Zaharia, and J. Zou. Frugalgpt: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research*, 2024b. 20

T. Chen, H. Shi, S. Tang, Z. Chen, F. Wu, and Y. Zhuang. CIL: Contrastive instance learning framework for distantly supervised relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6191–6200, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.483. URL https://aclanthology.org/2021.acl-long.483. 62, 67

Y. Chen and A. Ritter. Model selection for cross-lingual transfer. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5675–5687, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.459. URL https://aclanthology.org/2021.emnlp-main.459/. 27

Y. Chen, C. Jiang, A. Ritter, and W. Xu. Frustratingly easy label projection for cross-lingual transfer. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5775–5796, 2023. 10, 65, 66

Y. Chen, V. Shah, and A. Ritter. Translation and fusion improves cross-lingual information extraction. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7744–7764, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.382. URL https://aclanthology.org/2025.acl-long.382/. 5, 7, 49

A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of machine learning research*, 24(240):1–113, 2023. 43

D. Christou and G. Tsoumakas. Improving distantly-supervised relation extraction through bert-based label and instance embeddings. *IEEE Access*, 9:62574–62582, 2021. 67

J. H. Clark, D. Garrette, I. Turc, and J. Wieting. Canine: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91, 2022. 17

J. Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960. 65

A. Conneau, R. Rinott, G. Lample, A. Williams, S. Bowman, H. Schwenk, and V. Stoyanov. XNLI: Evaluating cross-lingual sentence representations. In E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii,

editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1269. URL https://aclanthology.org/D18-1269/. 1

A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, É. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, 2020. 1, 4, 5, 9, 13, 32, 48

M. De Raedt, S. K. Bitew, F. Godin, T. Demeester, and C. Develder. Zero-shot cross-lingual sentiment classification under distribution shift: an exploratory study. In D. Ataman, editor, *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 50–66, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.mrl-1.5. URL https://aclanthology.org/2023.mrl-1.5/. 2

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019. 1, 4, 13, 32, 35

A. Dubey, A. Jain, A. Gudibande, A. Lin, B. McIntosh, M. Miculicich, S. Gross, S. Krishna, A. Salemi, R. Hu, et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783. 43

C. Dyer, V. Chahuneau, and N. A. Smith. A simple, fast, and effective reparameterization of IBM model 2. In L. Vanderwende, H. Daumé III, and K. Kirchhoff, editors, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://aclanthology.org/N13-1073/. 10, 11

B. Ebing and G. Glavaš. To translate or not to translate: A systematic investigation of translation-based cross-lingual transfer to low-resource languages. In K. Duh, H. Gomez, and S. Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5325–5344, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.298. URL https://aclanthology.org/2024.naacl-long.298/. 11

B. Ebing and G. Glavaš. The devil is in the word alignment details: On translation-based cross-lingual transfer for token classification tasks. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18111–18128, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.931. URL https://aclanthology.org/2025.findings-acl.931/. 11

A. Ebrahimi, M. Mager, A. Oncevay, V. Chaudhary, L. Chiruzzo, A. Fan, J. Ortega, R. Ramos, A. Rios, I. V. Meza Ruiz, G. Giménez-Lugo, E. Mager, G. Neubig, A. Palmer, R. Coto-Solano, T. Vu, and K. Kann. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.435. URL https://aclanthology.org/2022.acl-long.435/. 49, 105

D. Edge, T. Krzyzanowski, F. Basisty, et al. A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024. 97

M. Ehrmann, M. Turchi, and R. Steinberger. Building a multilingual named entity-annotated corpus using annotation projection. In R. Mitkov and G. Angelova, editors, *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 118–124, Hissar, Bulgaria, Sept. 2011. Association for Computational Linguistics. URL https://aclanthology.org/R11-1017/. 10

J. Eronen, M. Ptaszynski, and F. Masui. Zero-shot cross-lingual transfer language selection using linguistic similarity. *Information Processing & Management*, 60(3):103250, 2023. 27

H. Fei, M. Zhang, and D. Ji. Cross-lingual semantic role labeling with high-quality translated training corpus. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7014–7026, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.627. URL https://aclanthology.org/2020.acl-main.627/. 10

A. R. Feinstein and D. V. Cicchetti. High agreement but low kappa: I. the problems of two paradoxes. *Journal of clinical epidemiology*, 43(6):543–549, 1990. 66

F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang. Language-agnostic bert sentence embedding. https://github.com/GoogleResearch/LaBSE, 2020. Technical report. 19

J. Gala, P. A. Chitale, A. K. Raghavan, V. Gumma, S. Doddapaneni, A. K. M, J. A. Nawale, A. Sujatha, R. Puduppully, V. Raghavan, P. Kumar, M. M. Khapra, R. Dabre, and A. Kunchukuttan. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=vfT4YuzAYA. 65

T. Gao, X. Han, Y. Bai, K. Qiu, Z. Xie, Y. Lin, Z. Liu, P. Li, M. Sun, and J. Zhou. Manual evaluation matters: Reviewing test protocols of distantly supervised relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1306–1318, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.112. URL https://aclanthology.org/2021.findings-acl.112. 23, 65, 66, 67, 68, 91

I. García-Ferrero. Cross-lingual transfer for low-resource natural language processing. *arXiv preprint arXiv:2502.02722*, 2025. 13

I. García-Ferrero, R. Agerri, and G. Rigau. Model and data transfer for cross-lingual sequence labelling in zero-resource settings. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6403–6416, 2022. 10

I. García-Ferrero, R. Agerri, and G. Rigau. T-projection: High quality annotation projection for sequence labeling tasks. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15203–15217, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.1015. URL https://aclanthology.org/2023.findings-emnlp.1015/. 10

Gemini Team. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 43

D. Gerz, I. Vulić, E. Ponti, J. Naradowsky, R. Reichart, and A. Korhonen. Language modeling for morphologically rich languages: Character-aware modeling for word-level prediction. *Transactions of the Association for Computational Linguistics*, 6:451–465, 07 2018. ISSN 2307-387X. doi: 10.1162/tacl_a_00032. URL https://doi.org/10.1162/tacl_a_00032. 17

A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 1

R. Guo and D. Roth. Constrained labeled data generation for low-resource named entity recognition. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4519–4533, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.396. URL https://aclanthology.org/2021.findings-acl.396/. 12

S. Gupta, M. Gardner, and S. Singh. Coverage-based example selection for in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13924–13950, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.930. URL https://aclanthology.org/2023.findings-emnlp.930. 19, 45, 48, 50

D. Ha, A. Dai, and Q. V. Le. Hypernetworks. *arXiv preprint arXiv:1609.09106*, 2016. 15

X. Han, T. Gao, Y. Yao, D. Ye, Z. Liu, and M. Sun. OpenNRE: An open and extensible toolkit for neural relation extraction. In *Proceedings of EMNLP-IJCNLP: System Demonstrations*, pages 169–174, 2019. doi: 10.18653/v1/D19-3029. URL https://www.aclweb.org/anthology/D19-3029. 63, 68

P. He, X. Liu, J. Gao, and W. Chen. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=XPZIaotutsD. 47, 50

R. He, L. Liu, H. Ye, Q. Tan, B. Ding, L. Cheng, J. Low, L. Bing, and L. Si. On the effectiveness of adapter-based tuning for pretrained language model adaptation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2208–2222, 2021b. 6

M. A. Hedderich, L. Lange, H. Adel, J. Strötgen, and D. Klakow. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, 2021. 1

L. Hennig, P. Thomas, and S. Möller. MultiTACRED: A multilingual version of the TAC relation extraction dataset. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3785–3801, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.210. URL https://aclanthology.org/2023.acl-long.210/. 23, 77

J. Henrich, S. J. Heine, and A. Norenzayan. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83, 2010. 97

R. Hoffmann, C. Zhang, X. Ling, L. Zettlemoyer, and D. S. Weld. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 541–550, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://aclanthology.org/P11-1055. 21

N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019. 14

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. In *Proceedings of the 10th International Conference on Learning Representations (ICLR)*, 2022. URL https://arxiv.org/abs/2106.09685. 16

J. Hu, S. Ruder, A. Siddhant, G. Neubig, O. Firat, and M. Johnson. Xtreme: a massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020a. 26

J. Hu, S. Ruder, A. Siddhant, G. Neubig, O. Firat, and M. Johnson. Xtreme: a massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020b. 2, 4

J. Huang, K. Huang, and T. Yu. msimcse: Multilingual contrastive sentence embeddings for zero-shot cross-lingual transfer. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3828–3845, 2023. URL https://aclanthology.org/2023.findings-acl.241/. 19

Y. Huang, B. Giledereli, A. Köksal, A. Özgür, and E. Ozkirimli. Balancing methods for multi-label text classification with long-tailed class distribution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8153–8161, 2021. 48

R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991. 30

A. Jain, B. Paranjape, and Z. C. Lipton. Entity projection via machine translation for cross-lingual NER. In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1083–1092, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1100. URL https://aclanthology.org/D19-1100/. 10

S. Jain and B. C. Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, 2019. 38

Z. Jian, S. Liu, W. Gao, and J. Cheng. Distantly supervised relation extraction based on non-taxonomic relation and self-optimization. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2024. 77

A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. L. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. Sayed, G. E. Charik, and E. Boyer. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023a. 43

D. Jiang, X. Ren, and B. Y. Lin. LLM-blender: Ensembling large language models with pairwise ranking and generative fusion. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14165–14178, Toronto, Canada, July 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.792. URL https://aclanthology.org/2023.acl-long.792/. 20

H. Jiang, Q. Wu, X. Luo, D. Li, C.-Y. Lin, Y. Yang, and L. Qiu. LongLLMLingua: Accelerating and enhancing LLMs in long context scenarios via prompt compression. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1658–1677, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.91. URL https://aclanthology.org/2024.acl-long.91/. 96

P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online,

July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.560. URL https://aclanthology.org/2020.acl-main.560/. 1, 2

J. Kaplan, S. McCandlish, T. J. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *ArXiv*, abs/2001.08361, 2020. URL https://api.semanticscholar.org/CorpusID:210861095. 2

J. Kapuriya, M. Kaushik, D. Ganguly, and S. Bhatia. Exploring the role of diversity in example selection for in-context learning. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2962–2966, 2025. 19, 50, 82

S. Khanuja, D. Bansal, S. Mehtani, S. Khosla, A. Dey, B. Gopalan, D. K. Margam, P. Aggarwal, R. T. Nagipogu, S. Dave, S. Gupta, S. C. B. Gali, V. Subramanian, and P. Talukdar. Muril: Multilingual representations for indian languages, 2021. 17

M. Khelli, S. Cahyawijaya, A. Purwarianti, and G. I. Winata. What causes knowledge loss in multilingual language models? In É. Le Ferrand, E. Klyachko, A. Postnikova, T. Shavrina, O. Serikov, E. Voloshina, and E. Vylomova, editors, *Proceedings of the Fourth Workshop on NLP Applications to Field Linguistics*, pages 15–25, Vienna, Austria, Aug. 2025. Association for Computational Linguistics. ISBN 979-8-89176-282-4. URL https://aclanthology.org/2025.fieldmatters-1.2/. 13

Y. Khemchandani, S. Mehtani, V. Patil, A. Awasthi, P. Talukdar, and S. Sarawagi. Exploiting language relatedness for low web-resource language model adaptation: An indic languages study. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1312–1323, 2021. 18

S. Kim, D. Ki, Y. Kim, and J. Lee. Cross-lingual qa: A key to unlocking in-context cross-lingual performance. In *ICML 2024 Workshop on In-Context Learning*, 2024. 6, 18

K. Kolluru, M. Mohammed, S. Mittal, S. Chakrabarti, and Mausam. Alignment-augmented consistent translation for multilingual open information extraction. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2502–2517, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.179. URL https://aclanthology.org/2022.acl-long.179. 10

K. Krippendorff. *Content analysis: An introduction to its methodology*. Sage publications, 2018. 65

S. Kudugunta, Y. Huang, A. Bapna, M. Krikun, D. Lepikhin, M.-T. Luong, and O. Firat. Beyond distillation: Task-level mixture-of-experts for efficient inference. In *Findings of the association for computational linguistics: EMNLP 2021*, pages 3577–3599, 2021. 31

A. Kunchukuttan, D. Kakwani, S. Golla, G. N.C., A. Bhattacharyya, M. M. Khapra, and P. Kumar. Ai4bharat-indicnlp corpus: Monolingual corpora and word embeddings for indic languages. *arXiv preprint arXiv:2005.00085*, 2020. 17

K. Kurniawan, L. Frermann, P. Schulz, and T. Cohn. Unsupervised cross-lingual transfer of structured predictors without source data. In M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2041–2054, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.149. URL https://aclanthology.org/2022.naacl-main.149/. 43

M. Laurer, W. v. Atteveldt, A. S. Casas, and K. Welbers. Less Annotating, More Classifying – Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT - NLI. *Preprint*, June 2022. URL https://osf.io/74b8k. Publisher: Open Science Framework. 108

A. Lauscher, V. Ravishankar, I. Vulić, and G. Glavaš. From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, 2020. 2, 4

D. M. Le, Y. Chen, A. Ritter, and W. Xu. Constrained decoding for cross-lingual label projection. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=DayPQKXaQk. 10, 49, 65

B. Lester, R. Al-Rfou, and N. Constant. The power of scale for parameter-efficient prompt tuning. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.243. URL https://aclanthology.org/2021.emnlp-main.243/. 16

M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7871–7880, 2020. 1

D. Li, T. Zhang, N. Hu, C. Wang, and X. He. Hiclre: A hierarchical contrastive learning framework for distantly supervised relation extraction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2567–2578, 2022. 68, 114

M. Li, C. Shao, G. Li, and M. Zhou. Hfmre: Constructing huffman tree in bags to find excellent instances for distantly supervised relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12820–12832, 2023. 68, 114

X. L. Li and P. Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, 2021. 16

Y. Liang, N. Duan, Y. Gong, N. Wu, F. Guo, W. Qi, M. Gong, L. Shou, D. Jiang, G. Cao, X. Fan, R. Zhang, R. Agrawal, E. Cui, S. Wei, T. Bharti, Y. Qiao, J.-H. Chen, W. Wu, S. Liu, F. Yang, D. Cam-

pos, R. Majumder, and M. Zhou. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.484. URL https://aclanthology.org/2020.emnlp-main.484/. 1

P. Lin, A. Martins, and H. Schuetze. XAMPLER: Learning to retrieve cross-lingual in-context examples. In L. Chiruzzo, A. Ritter, and L. Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3968–3977, Albuquerque, New Mexico, Apr. 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.221. URL https://aclanthology.org/2025.findings-naacl.221/. 85

X. V. Lin, T. Mihaylov, M. Artetxe, T. Wang, S. Chen, D. Simig, M. Ott, N. Goyal, S. Bhosale, J. Du, R. Pasunuru, S. Shleifer, P. S. Koura, V. Chaudhary, B. O'Horo, J. Wang, L. Zettlemoyer, Z. Kozareva, M. Diab, V. Stoyanov, and X. Li. Few-shot learning with multilingual generative language models. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.616. URL https://aclanthology.org/2022.emnlp-main.616/. 18

Y. Lin, S. Shen, Z. Liu, H. Luan, and M. Sun. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2124–2133, Berlin, Germany, Aug. 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1200. URL https://aclanthology.org/P16-1200. 61, 62, 67

R. Litschko, I. Vulić, S. P. Ponzetto, and G. Glavaš. On cross-lingual retrieval with multilingual text encoders. *Inf. Retr.*, 25(2):149–183, June 2022. ISSN 1386-4564. doi: 10.1007/s10791-022-09406-x. URL https://doi.org/10.1007/s10791-022-09406-x. 18

P. Littell, D. R. Mortensen, K. Lin, K. Kairis, C. Turner, and L. Levin. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, 2017. 2, 26

C. C. Liu, A. Korhonen, and I. Gurevych. Cultural learning-based culture adaptation of language models. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3114–3134, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.156. URL https://aclanthology.org/2025.acl-long.156/. 97

J. Liu, D. Shen, Y. Zhang, W. B. Dolan, L. Carin, and W. Chen. What makes good in-context examples for gpt-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, 2022. 43, 47

N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/. 19, 48

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 1

T. Luong, H. Pham, and C. D. Manning. Effective approaches to attention-based neural machine translation. In L. Màrquez, C. Callison-Burch, and J. Su, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal, Sept. 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1166. URL https://aclanthology.org/D15-1166/. 63

Y. Ma, Y. Cao, Y. Hong, and A. Sun. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10572–10601, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.710. URL https://aclanthology.org/2023.findings-emnlp.710/. 77

A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang, S. Gupta, B. P. Majumder, K. Hermann, S. Welleck, A. Yazdanbakhsh, and P. Clark. Self-refine: iterative refinement with self-feedback. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc. 96

C. Malaviya, G. Neubig, and P. Littell. Learning language representations for typology prediction. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Copenhagen, Denmark, September 2017. 15

A. Martínez-García, T. Badia, and J. Barnes. Evaluating morphological typology in zero-shot cross-lingual transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3136–3153, 2021. 2

Q. McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947. 24

N. Melgarejo, R. Zevallos, H. Gomez, and J. E. Ortega. WordNet-QU: Development of a lexical database for Quechua varieties. In N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, and S.-H. Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4429–4433, Gyeongju, Republic of Korea, Oct. 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.390/. 1

S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, and L. Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.759. URL https://aclanthology.org/2022.emnlp-main.759/. 45, 47

M. Mintz, S. Bills, R. Snow, and D. Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore, Aug. 2009. Association for Computational Linguistics. URL https://aclanthology.org/P09-1113. 2, 21, 61, 95

N. Muennighoff, A. Rush, B. Barak, T. Le Scao, N. Tazi, A. Piktus, S. Pyysalo, T. Wolf, and C. A. Raffel. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36: 50358–50376, 2023. 1

B. Muller, A. Anastasopoulos, B. Sagot, and D. Seddah. When being unseen from mbert is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462, 2021. 4, 94

A. Nag, B. Samanta, A. Mukherjee, N. Ganguly, and S. Chakrabarti. A data bootstrapping recipe for low-resource multilingual relation classification. In A. Bisazza and O. Abend, editors, *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 575–587, Online, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.conll-1.45. URL https://aclanthology.org/2021.conll-1.45/. 66, 67

A. Nag, B. Samanta, A. Mukherjee, N. Ganguly, and S. Chakrabarti. Transfer learning for low-resource multilingual relation classification. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(2), Mar. 2023. ISSN 2375-4699. doi: 10.1145/3554734. URL https://doi.org/10.1145/3554734. 67

A. Nag, A. Mukherjee, N. Ganguly, and S. Chakrabarti. Cost-performance optimization for processing low-resource language tasks using commercial llms. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15681–15701, 2024. 4, 49, 83, 84

A. Nag, S. Chakrabarti, A. Mukherjee, and N. Ganguly. Efficient continual pre-training of LLMs for low-resource languages. In W. Chen, Y. Yang, M. Kachuee, and X.-Y. Fu, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 304–317, Albuquerque, New Mexico, Apr. 2025. Association for Computational Linguistics. ISBN 979-8-89176-194-0. URL https://aclanthology.org/2025.naacl-industry.25/. 17, 97

A. Nambi, V. Balloli, M. Ranjit, T. Ganu, K. Ahuja, S. Sitaram, and K. Bali. Breaking language barriers with a leap: Learning strategies for polyglot llms. *arXiv preprint arXiv:2305.17740*, 2023. 45

Y. Nandwani, V. Kumar, D. Raghu, S. Joshi, and L. Lastras. Pointwise mutual information based metric and decoding strategy for faithful generation in document grounded dialogs. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10335–10347, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.639. URL https://aclanthology.org/2023.emnlp-main.639/. 97

B. Nath and Y. Gulzar. Cross-lingual sparse-moe distillation for efficient low-resource assamese–english and bodo–english translation. *Scientific Reports*, 2026. 31

J. Ni, G. Dinu, and R. Florian. Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. In R. Barzilay and M.-Y. Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1470–1480, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1135. URL https://aclanthology.org/P17-1135/. 10

J. Nivre, M.-C. de Marneffe, F. Ginter, J. Hajic, C. D. Manning, S. Pyysalo, S. Schuster, F. Tyers, and D. Zeman. Universal dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, 2020. 21, 31

NLLB Team, M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. Mejia-Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. Spruit, C. Tran, P. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzmán, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk, and J. Wang. No language left behind: Scaling human-centered machine translation. 2022. 49

A. Nzeyimana and A. Niyongabo Rubungo. KinyaBERT: a morphology-aware Kinyarwanda language model. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5347–5363, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.367. URL https://aclanthology.org/2022.acl-long.367/. 17

K. Ogueji, Y. Zhu, and J. Lin. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In D. Ataman, A. Birch, A. Conneau, O. Firat, S. Ruder, and G. G. Sahin, editors, *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.mrl-1.11. URL https://aclanthology.org/2021.mrl-1.11/. 13, 49

OpenAI. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. URL https://arxiv.org/abs/2410.21276. 43

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022. 1

X. Ouyang, S. Wang, C. Pang, Y. Sun, H. Tian, H. Wu, and H. Wang. ERNIE-M: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 27–38, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.3. URL https://aclanthology.org/2021.emnlp-main.3/. 12

F. Palma Gomez, R. Sanabria, Y.-h. Sung, D. Cer, S. Dalmia, and G. Hernandez Abrego. Transforming LLMs into cross-modal and cross-lingual retrieval systems. In E. Salesky, M. Federico, and M. Carpuat, editors, *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 23–32, Bangkok, Thailand (in-person and online), Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.iwslt-1.4. URL https://aclanthology.org/2024.iwslt-1.4/. 20

X. Pan, B. Zhang, J. May, J. Nothman, K. Knight, and H. Ji. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, 2017. 21, 31

T. Parekh, I.-H. Hsu, K.-H. Huang, K.-W. Chang, and N. Peng. Contextual label projection for cross-lingual structured prediction. In K. Duh, H. Gomez, and S. Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5738–5757, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.321. URL https://aclanthology.org/2024.naacl-long.321/. 3, 5, 44, 45

H. H. Park, K. J. Zhang, C. Haley, K. Steimel, H. Liu, and L. Schwartz. Morphology matters: A multilingual language modeling analysis. *Transactions of the Association for Computational Linguistics*, 9:261–276, 03 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00365. URL https://doi.org/10.1162/tacl_a_00365. 17

M. Patidar, R. Sawhney, A. Singh, B. Chatterjee, M. ., and I. Bhattacharya. Few-shot transfer learning for knowledge base question answering: Fusing supervised models with in-context learning. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9147–9165, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.495. URL https://aclanthology.org/2024.acl-long.495/. 19, 96

V. Patil, P. Talukdar, and S. Sarawagi. Overlap-based vocabulary generation improves cross-lingual transfer among related languages. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 219–233, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.18. URL https://aclanthology.org/2022.acl-long.18/. 17

J. Pfeiffer, I. Vulić, I. Gurevych, and S. Ruder. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.617. URL https://aclanthology.org/2020.emnlp-main.617. 3, 5, 6, 9, 14, 25, 94

J. Pfeiffer, A. Kamath, A. Rücklé, K. Cho, and I. Gurevych. Adapterfusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, 2021a. 29

J. Pfeiffer, I. Vulić, I. Gurevych, and S. Ruder. Unks everywhere: Adapting multilingual language models to new scripts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10186–10203, 2021b. 4, 15, 18, 95

K. T.-K. Phan, D. Ngoc Hao, D. V. Thin, and N. Luu-Thuy Nguyen. Exploring zero-shot cross-lingual aspect-based sentiment analysis using pre-trained multilingual language models. In *2021 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, pages 1–6, 2021. doi: 10.1109/MAPR53640.2021.9585242. 2, 4

T. Pires, E. Schlinger, and D. Garrette. How multilingual is multilingual bert? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, 2019. 13

L. Qin, Q. Chen, F. Wei, S. Huang, and W. Che. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2695–2709, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.163. URL https://aclanthology.org/2023.emnlp-main.163/. 18

P. Qin, W. Xu, and W. Y. Wang. Robust distant supervision relation extraction via deep reinforcement learning. In I. Gurevych and Y. Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2137–2147. Association for Computational Linguistics, 2018. 62

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. 2019. 1, 67

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 1

V. Rathore, K. Badola, P. Singla, and Mausam. PARE: A simple and strong baseline for monolingual and multilingual distantly supervised relation extraction. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 340–354, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.38. URL https://aclanthology.org/2022.acl-short.38/. 6

V. Rathore, R. Dhingra, P. Singla, and Mausam. ZGUL: Zero-shot generalization to unseen languages using multi-source ensembling of language adapters. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6969–6987, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.431. URL https://aclanthology.org/2023.emnlp-main.431/. 6

V. Rathore, M. H. Faisal, P. Singla, and Mausam. Combining distantly supervised models with in context learning for monolingual and cross-lingual relation extraction. *arXiv preprint arXiv:2510.18344*, 2025. 6

V. K. Rathore, A. Deb, A. K. Chandresh, P. Singla, and M. . SSP: Self-supervised prompting for cross-lingual transfer to low-resource languages using large language models. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15081–15102, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.886. URL https://aclanthology.org/2024.findings-emnlp.886/. 6

S. Riedel, L. Yao, and A. McCallum. Modeling relations and their mentions without labeled text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 148–163. Springer, 2010. 7, 21, 66

A. Roberts, C. Raffel, and N. Shazeer. How much knowledge can you pack into the parameters of a language model? In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.437. URL https://aclanthology.org/2020.emnlp-main.437/. 43

S. Ruder, I. Vulić, and A. Søgaard. A survey of cross-lingual word embedding models. *J. Artif. Int. Res.*, 65(1):569–630, May 2019. ISSN 1076-9757. doi: 10.1613/jair.1.11640. URL https://doi.org/10.1613/jair.1.11640. 2, 10

P. Rust, J. Pfeiffer, I. Vulić, S. Ruder, and I. Gurevych. How good is your tokenizer? on the monolingual performance of multilingual language models. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.243. URL https://aclanthology.org/2021.acl-long.243/. 17

V. V. Saley, R. J. Das, D. Raghu, and M. . Synergizing in-context learning with hints for end-to-end task-oriented dialog systems. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5596–5612, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.320. URL https://aclanthology.org/2024.emnlp-main.320/. 19

T. Schick, J. Dwivedi-Yu, R. Dessí, R. Raileanu, M. Lomeli, E. Hambro, L. Zettlemoyer, N. Cancedda, and T. Scialom. Toolformer: language models can teach themselves to use tools. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc. 20

F. D. Schmidt, I. Vulić, and G. Glavaš. Don't stop fine-tuning: On training regimes for few-shot cross-lingual transfer with multilingual language models. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10725–10742, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.736. URL https://aclanthology.org/2022.emnlp-main.736/. 3

M. Sclar, Y. Choi, Y. Tsvetkov, and A. Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024. URL https://api.semanticscholar.org/CorpusID:264172710. 103

R. Shah, B. Lin, A. Gershman, and R. Frederking. Synergy: a named entity recognition system for resource-scarce languages such as swahili using online machine translation. In *Proceedings of the Second Workshop on African Language Technology (AfLaT 2010)*, pages 21–26, 2010. 11

U. Shaham, J. Herzig, R. Aharoni, I. Szpektor, R. Tsarfaty, and M. Eyal. Multilingual instruction tuning with just a pinch of multilinguality. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2304–2317, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.136. URL https://aclanthology.org/2024.findings-acl.136/. 13

C. E. Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3): 379–423, 1948. 27

N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017. 30

N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, and S. Yao. Reflexion: language agents with verbal reinforcement learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc. 20

H. Singh, N. Gupta, S. Bharadwaj, D. Tewari, and P. Talukdar. IndicGenBench: A multilingual benchmark to evaluate generation capabilities of LLMs on Indic languages. In L.-W. Ku, A. Martins,

and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11047–11073, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.595. URL https://aclanthology.org/2024.acl-long.595. 17, 95, 97

L. B. Soares, N. Fitzgerald, J. Ling, and T. Kwiatkowski. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, 2019. 73

J. Sohn, H. Jung, A. Cheng, J. Kang, Y. Du, and D. R. Mortensen. Zero-shot cross-lingual NER using phonemic representations for low-resource languages. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13595–13602, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.753. URL https://aclanthology.org/2024.emnlp-main.753/. 12

C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. 37

A. Subburathinam, D. Lu, H. Ji, J. May, S.-F. Chang, A. Sil, and C. Voss. Cross-lingual structure transfer for relation and event extraction. In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 313–325, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1030. URL https://aclanthology.org/D19-1030/. 5

M. Surdeanu, J. Tibshirani, R. Nallapati, and C. D. Manning. Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 455–465, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL https://aclanthology.org/D12-1042. 22, 61

E. Tanwar, S. Dutta, M. Borthakur, and T. Chakraborty. Multilingual LLMs are better cross-lingual in-context learners with alignment. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6292–6307, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.346. URL https://aclanthology.org/2023.acl-long.346/. 45

H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. Llama: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971. 43, 50

A. Üstün, A. Bisazza, G. Bouma, and G. van Noord. UDapter: Language adaptation for truly Universal Dependency parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2302–2315, Online, Nov. 2020. Association for Computational

Linguistics. doi: 10.18653/v1/2020.emnlp-main.180. URL https://aclanthology.org/2020.emnlp-main.180. 3, 15, 27, 31, 94

S. Vashishth, R. Joshi, S. S. Prayaga, C. Bhattacharyya, and P. Talukdar. RESIDE: Improving distantly-supervised neural relation extraction using side information. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1257–1266, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1157. URL https://aclanthology.org/D18-1157. 62, 67

S. Vatsal, H. Dubey, and A. Singh. Multilingual prompt engineering in large language models: A survey across nlp tasks. 05 2025. doi: 10.48550/arXiv.2505.11665. 18

S. Wadhwa, S. Amir, and B. C. Wallace. Revisiting relation extraction in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15566–15589, 2023. 77

X. Wan, R. Sun, H. Dai, S. Arik, and T. Pfister. Better zero-shot reasoning with self-adaptive prompting. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3493–3514, Toronto, Canada, July 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.216. URL https://aclanthology.org/2023.findings-acl.216/. 45

X. Wan, R. Sun, H. Nakhost, H. Dai, J. Eisenschlos, S. Arik, and T. Pfister. Universal self-adaptive prompting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7437–7462, 2023b. 45

L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder, and F. Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022. 81

P. Wang, X. Wang, C. Lou, S. Mao, P. Xie, and Y. Jiang. Effective demonstration annotation for in-context learning via language model-based determinantal point process. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1266–1280, Miami, Florida, USA, Nov. 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.74. URL https://aclanthology.org/2024.emnlp-main.74/. 19

P. Wang, X. Wang, C. Lou, S. Mao, P. Xie, and Y. Jiang. Effective demonstration annotation for in-context learning via language model-based determinantal point process. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1266–1280, 2024b. 50

X. Wang, Y. Tsvetkov, S. Ruder, and G. Neubig. Efficient test time adapter ensembling for low-resource language varieties. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 730–737, 2021. 27, 31, 99

X. Wang, J. Wu, Y. Yichen, D. Cai, M. Li, and W. Jia. Demonstration selection for in-context learning via reinforcement learning. In *International Conference on Machine Learning*, pages 64937–64954. PMLR, 2025. 19, 45, 48, 50

C. Whitehouse, F. Huot, J. Bastings, M. Dehghani, C.-C. Lin, and M. Lapata. Low-rank adaptation for multilingual summarization: An empirical study. In K. Duh, H. Gomez, and S. Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1202–1228, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.77. URL https://aclanthology.org/2024.findings-naacl.77/. 16

G. I. Winata, A. Madotto, Z. Lin, R. Liu, J. Yosinski, and P. Fung. Language models are few-shot multilingual learners. In D. Ataman, A. Birch, A. Conneau, O. Firat, S. Ruder, and G. G. Sahin, editors, *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 1–15, Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.mrl-1.1. URL https://aclanthology.org/2021.mrl-1.1/. 18

B. Workshop, T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022. 1, 4

C.-S. Wu, S. C. Hoi, R. Socher, and C. Xiong. Tod-bert: Pre-trained natural language understanding for task-oriented dialogue. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 917–929, 2020a. 1

L. Wu, H.-R. Wei, B. Yang, and W. Lu. From English to second language mastery: Enhancing LLMs with cross-lingual continued instruction tuning. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23006–23023, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1121. URL https://aclanthology.org/2025.acl-long.1121/. 13

Q. Wu, Z. Lin, B. F. Karlsson, J.-G. Lou, and B. Huang. Single-/multi-source cross-lingual NER via teacher-student learning on unlabeled data in target language. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6505–6514, Online, July 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.581. URL https://aclanthology.org/2020.acl-main.581/. 12

S. Wu and M. Dredze. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1077. URL https://aclanthology.org/D19-1077/. 9, 13

Y. Xiao, P. P. Liang, U. Bhatt, W. Neiswanger, R. Salakhutdinov, and L.-P. Morency. Uncertainty quantification with pre-trained language models: A large-scale empirical analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7273–7284, 2022. 45

H. Xu and K. Murray. Por qué não utiliser alla språk? mixed training with gradient optimization in few-shot cross-lingual transfer. In M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, editors, *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2043–2059, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.157. URL https://aclanthology.org/2022.findings-naacl.157/. 3

W. Xu, R. Hoffmann, L. Zhao, and R. Grishman. Filling knowledge base gaps for distant supervision of relation extraction. In H. Schuetze, P. Fung, and M. Poesio, editors, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 665–670, Sofia, Bulgaria, Aug. 2013. Association for Computational Linguistics. URL https://aclanthology.org/P13-2117/. 22

L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. URL https://aclanthology.org/2021.naacl-main.41/. 17

L. Xue, A. Barua, N. Constant, R. Al-Rfou, S. Narang, M. Kale, A. Roberts, and C. Raffel. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306, 2022. doi: 10.1162/tacl_a_00461. URL https://aclanthology.org/2022.tacl-1.17/. 17

A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, C. Zheng, D. Liu, F. Zhou, F. Huang, F. Hu, H. Ge, H. Wei, H. Lin, J. Tang, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Zhou, J. Lin, K. Dang, K. Bao, K. Yang, L. Yu, L. Deng, M. Li, M. Xue, M. Li, P. Zhang, P. Wang, Q. Zhu, R. Men, R. Gao, S. Liu, S. Luo, T. Li, T. Tang, W. Yin, X. Ren, X. Wang, X. Zhang, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Zhang, Y. Wan, Y. Liu, Z. Wang, Z. Cui, Z. Zhang, Z. Zhou, and Z. Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a. 1, 43

J. Yang, W. Zhang, Y. Miao, S. Quan, Z. Wu, Q. Peng, L. Yang, T. Liu, Z. Cui, B. Hui, and J. Lin. Qwen2.5-xCoder: Multi-agent collaboration for multilingual code instruction tuning. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13121–13131, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.642. URL https://aclanthology.org/2025.acl-long.642/. 20

W. Yu, S. Zhang, X. Wang, R. Meng, and D. Radev. Sparkle: Structured prompting and refinement for knowledge-grounded language models. In *Findings of the Association for Computational Linguistics (ACL Findings)*, 2024. 20

D. Zelenko, C. Aone, and A. Richardella. Kernel methods for relation extraction. *Journal of machine learning research*, 3(Feb):1083–1106, 2003. 21

D. Zeng, K. Liu, Y. Chen, and J. Zhao. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1753–1762, 2015. 62

Y. Zhang, V. Zhong, D. Chen, G. Angeli, and C. D. Manning. Position-aware attention and supervised data improve slot filling. In M. Palmer, R. Hwa, and S. Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1004. URL https://aclanthology.org/D17-1004/. 77

X. Zhao, Y. Deng, M. Yang, L. Wang, R. Zhang, H. Cheng, W. Lam, Y. Shen, and R. Xu. A comprehensive survey on relation extraction: Recent advances and new frontiers. *ACM Computing Surveys*, 56(11): 1–39, 2024. 76

Y. Zhao, W. Zhang, H. Wang, K. Kawaguchi, and L. Bing. Adamergex: Cross-lingual transfer with large language models via adaptive adapter merging. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9785–9800, 2025. 16

Z. Zhao, E. Wallace, S. Feng, D. Klein, and S. Singh. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. PMLR, 2021. 43, 45, 47

L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023. 77, 87

G. Zhou, J. Su, J. Zhang, and M. Zhang. Exploring various knowledge in relation extraction. In *Proceedings of the 43rd annual meeting of the association for computational linguistics (acl'05)*, pages 427–434, 2005. 21

S. Zhou, Y. Meng, B. Jin, and J. Han. Grasping the essentials: Tailoring large language models for zero-shot relation extraction. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13462–13486, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.747. URL https://aclanthology.org/2024.emnlp-main.747/. 77

W. Zhou, F. Liu, I. Vulić, N. Collier, and M. Chen. Prix-LM: Pretraining for multilingual knowledge base construction. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th*

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5412–5424, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.371. URL https://aclanthology.org/2022.acl-long.371/. 1

S. Zhu, Supryadi, S. Xu, H. Sun, L. Pan, M. Cui, J. Du, R. Jin, A. Branco, and D. Xiong. Multilingual large language models: A systematic survey. *ArXiv*, abs/2411.11072, 2024. URL https://api.semanticscholar.org/CorpusID:274131470. 18

B. Zoph, C. Raffel, D. Schuurmans, D. Yogatama, D. Zhou, D. Metzler, E. H. Chi, J. Wei, J. Dean, L. B. Fedus, M. P. Bosma, O. Vinyals, P. Liang, S. Borgeaud, T. B. Hashimoto, and Y. Tay. Emergent abilities of large language models. *TMLR*, 2022. 1, 43

List of Included Papers

This thesis is based on the following publications:

1. "PARE: A Simple and Strong Baseline for Monolingual and Multilingual Distantly Supervised Relation Extraction". Vipul Rathore, Kartikeya Badola, Parag Singla, and Mausam. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Dublin, Ireland, July 2022. (Acceptance Rate (Main): 21%)

2. "ZGUL: Zero-shot Generalization to Unseen Languages using Multi-source Ensembling of Language Adapters". Vipul Rathore, Rajdeep Dhingra, Parag Singla, and Mausam. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Singapore, December 2023. (Acceptance Rate (Main): 23%)

3. "SSP: Self-Supervised Prompting for Cross-Lingual Transfer to Low-Resource Languages using Large Language Models". Vipul Rathore, Aniruddha Deb, Ankish Kumar Chandresh, Parag Singla, and Mausam. *Findings of the Association for Computational Linguistics: EMNLP 2024*. Miami, Florida, USA, November 2024. (Acceptance Rate (Main + Findings): 37%)

4. "Combining Distantly Supervised Models with In Context Learning for Monolingual and Cross-Lingual Relation Extraction". Vipul Rathore, Malik Hammad Faisal, Parag Singla, and Mausam. *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. San Diego, CA, USA. July 2026. (Acceptance Rate (Oral): 3.5%)

Biography

Vipul Kumar Rathore hails from Gwalior, Madhya Pradesh. He completed his Bachelor's from the Department of Electrical Engineering at IIT Bombay in 2016 and Master's from Computational and Data Sciences at IISc Bangalore in 2019, and joined the PhD program of the Department of Computer Science and Engineering at IIT Delhi in 2019, supported by the Prime Minister's Research Fellowship (PMRF). He has published several research papers in top-tier venues such as ACL and EMNLP, and has served as a reviewer for these conferences as well. During his PhD, he also successfully completed an industry research collaboration with Huawei Technologies on multilingual relation extraction. Prior to his PhD, he had one year of industrial experience with HP Enterprises and an internship with Ericsson Research.