

Wrap Up

Mausam

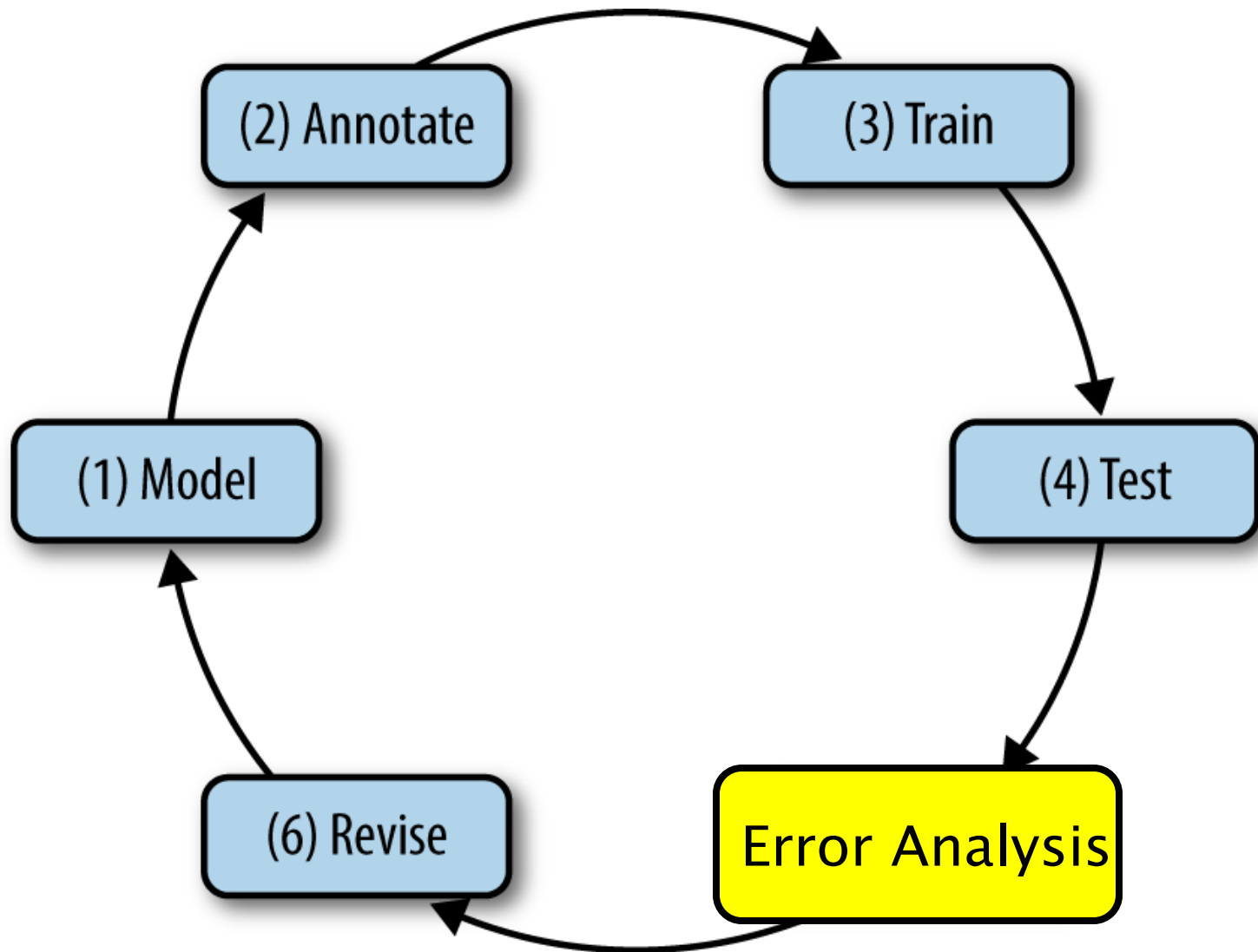
Grading

- 35% final exam
- 20% midterm (5+5+5+5)
- 10% A1
- 10% A3
- 10% A4
- 15% Quiz (each quiz 5%)
- Extra credit for class participation (upto 3%)

Key Points: NLP Overview

- Challenges
 - Ambiguity, Ambiguity, Ambiguity, Sparsity
- Words: Morphology
- Sentences: Syntax
 - POS tagging, NP chunking, Parsing
- Sentences/Documents: Semantics
 - Words/Bigrams encode meanings, but are also sparse
 - Distributional Semantics, Shallow semantics
 - Patterns: bootstrapping
- Documents: Coreference, Discourse
- Applications
 - Information Extraction, Machine Translation, Summarization, Dialog

Key Points: ML Cycle



Revise Tweaks

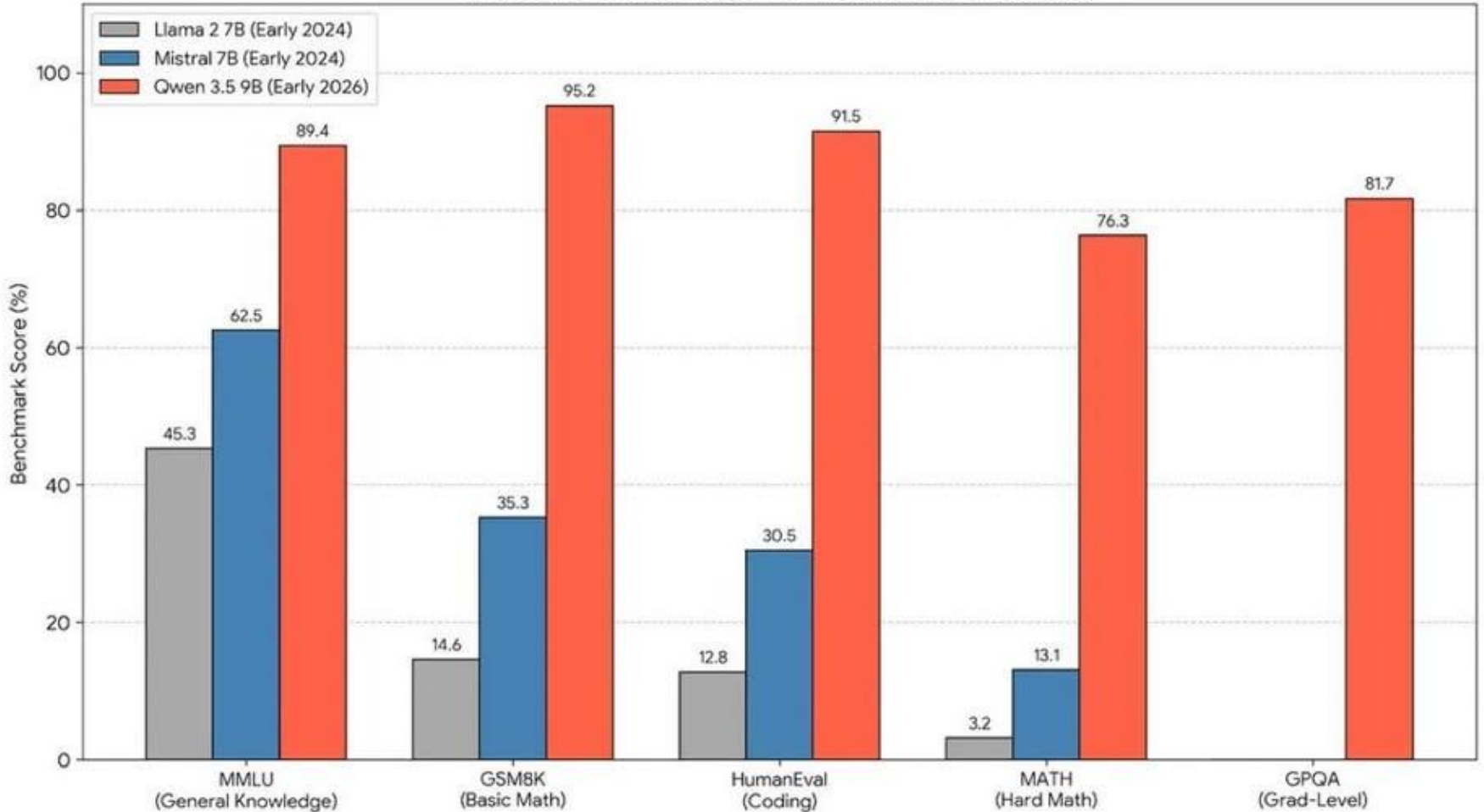
- Retrieval augmentation
- Use similar tasks; pretrain; then fine-tune few-shot
- Transfer learning; Multi-task learning
- Data augmentation
- Auxilliary loss
- Add human-features
- Add constraints
- Change architectures
- Synthetic data generation using LLMs
- LLM workflows (agents)

The World of LLMs

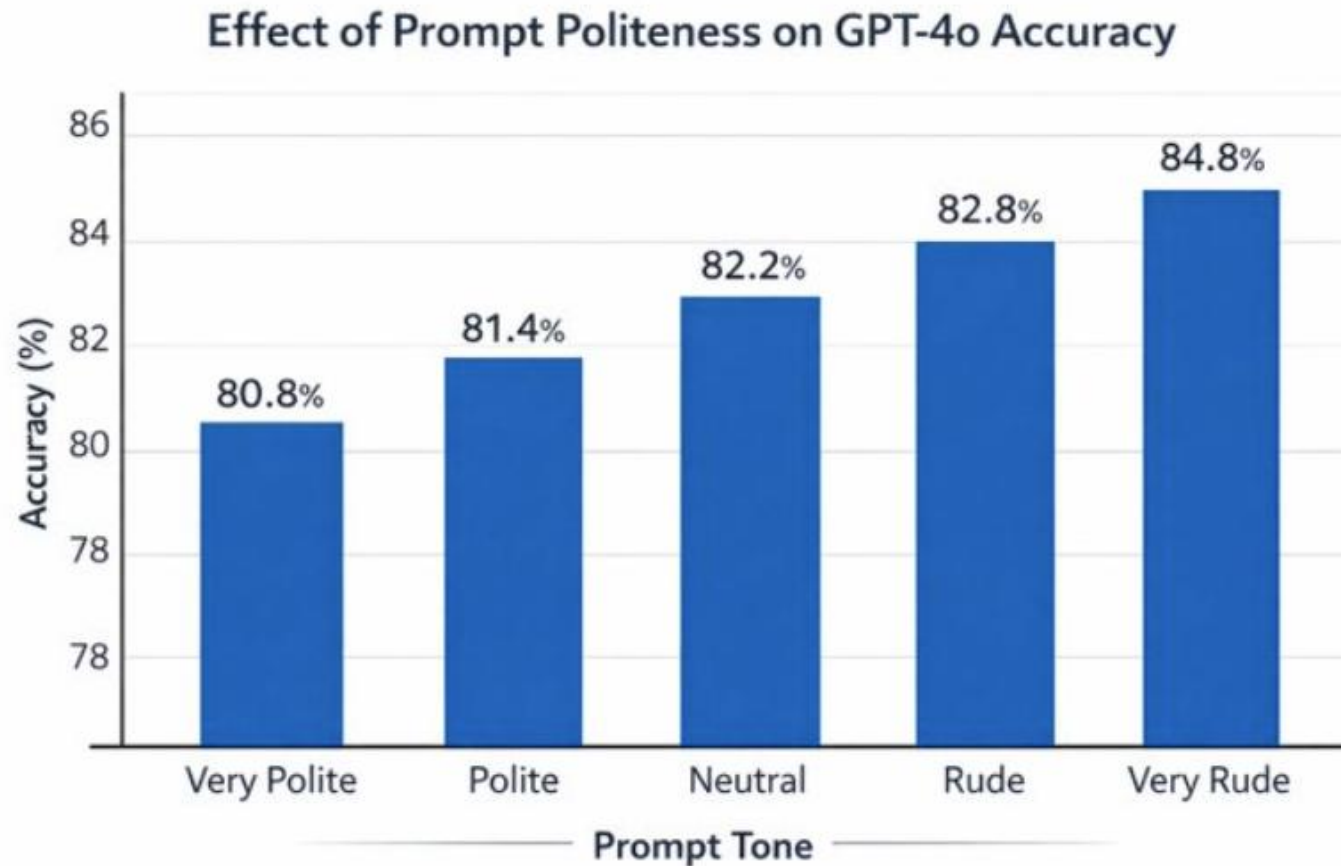
- Pre-Training → Instruction Fine Tuning → RLVR → Alignment
- Efficient Training: LoRA; Flash Attention; Multi-GPU Training
- Efficient Inference: Distillation, Speculative Sampling
- Performance
 - Generation: Diversity-Quality tradeoff
 - Retrieval Augmentation: reduce hallucinations
 - Prompting techniques; Coding/other tools; LLMs as Idea Generators

Where are We Today?

The Generational Leap: ~9B Models (2024 vs 2026)



Where are We Today?



Penn State University, n=50 questions.

Where are We Today?

'I think you're testing me': Anthropic's new AI model asks testers to come clean

Safety evaluation of Claude Sonnet 4.5 raises questions about whether predecessors 'played along', firm says



Where are We Today?

The screenshot shows the Google Translate web interface. At the top, there is a hamburger menu icon, the Google Translate logo, a grid icon, and a blue 'Sign in' button. Below this, there are two tabs: 'Text' (selected) and 'Documents'. The language selection bar shows 'HUNGARIAN - DETECTED' on the left and 'ENGLISH' on the right, with other language options like 'POLISH' and 'PORTUGUESE' visible. The main content area is split into two columns. The left column contains the Hungarian text: 'Ő szép. Ő okos. Ő olvas. Ő mosogat. Ő épít. Ő varr. Ő tanít. Ő főz. Ő kutat. Ő gyereket nevel. Ő zenél. Ő takarító. Ő politikus. Ő sok pénzt keres. Ő süteményt süt. Ő professzor. Ő asszisztens.' with a close icon (X) to its right. The right column contains the English translation: 'She is beautiful. He is clever. He reads. She washes the dishes. He builds. She sews. He teaches. She cooks. He's researching. She is raising a child. He plays music. She's a cleaner. He is a politician. He makes a lot of money. She is baking a cake. He's a professor. She's an assistant.' with a star icon to its right. At the bottom of the interface, there are icons for voice input, voice output, a character count '194 / 5000', and icons for copy, edit, and share.

Where are We Today?

First Proof?

OpenAI

February 13, 2026

Abstract

Here we present the solution attempts our models found for the ten <https://1stproof.org/> tasks posted on February 5th, 2026. All presented attempts were generated and typeset by our models.

Contents

- | | | |
|---|--|---|
| 1 | Smooth shifts of the Φ_3^4 measure on \mathbb{T}^3 | 2 |
| 2 | A nonvanishing test vector for the twisted local Rankin–Selberg integral | 8 |

Where are We Today?

Claude's Cycles

Don Knuth, Stanford Computer Science Department
(28 February 2026; revised 02 March 2026)

Shock! Shock! I learned yesterday that an open problem I'd been working on for several weeks had just been solved by Claude Opus 4.6 — Anthropic's hybrid reasoning model that had been released three weeks earlier! It seems that I'll have to revise my opinions about “generative AI” one of these days. What a joy it is to learn not only that my conjecture has a nice solution but also to celebrate this dramatic advance in automatic deduction and creative problem solving. I'll try to tell the story briefly in this note.

Here's the problem, which came up while I was writing about directed Hamiltonian cycles for a future volume of *The Art of Computer Programming*:

Consider the digraph with m^3 vertices ijk for $0 \leq i, j, k < m$, and three arcs from each vertex, namely to i^+jk , ij^+k , and ijk^+ , where $i^+ = (i+1) \bmod m$. Try to find a general decomposition of the arcs into three directed m^3 -cycles, for all $m > 2$.

I had solved the problem for $m = 3$, and asked for a generalization as part of the answer to an exercise in [3]. My friend Filip Stappers rose to the challenge, and empirically discovered solutions for $4 \leq m \leq 16$; therefore it became highly likely that the desired decompositions do exist, except when $m \leq 2$.

Indeed, it was Filip who had the gumption to pose this question to Claude, using exactly the wording above. He also gave guidance/coaching, instructing Claude to summarize its ongoing progress:

```
** After EVERY exploreXX.py run, IMMEDIATELY update this file [plan.md]
before doing anything else. ** No exceptions. Do not start the next exploration
until the previous one is documented here.
```

Where are We Today?

Accelerating Scientific Research with Gemini: Case Studies and Common Techniques

David P. Woodruff^{*, †, †1,2}, Vincent Cohen-Addad^{†, †1}, Lalit Jain^{†1}, Jieming Mao^{†1}, Song Zuo^{†, †1}, MohammadHossein Bateni^{†1}, Simina Brânzei^{†3,1}, Michael P. Brenner^{†1,5}, Lin Chen^{†1}, Ying Feng^{†6}, Lance Fortnow^{†7}, Gang Fu^{†1}, Ziyi Guan^{†13}, Zahra Hadizadeh^{†10}, Mohammad T. Hajiaghayi^{†1,14}, Mahdi JafariRaviz^{†14}, Adel Javanmard^{†1,4}, Karthik C. S.^{†8}, Ken-ichi Kawarabayashi^{†12}, Ravi Kumar^{†1}, Silvio Lattanzi^{†1}, Euiwoong Lee^{†9}, Yi Li^{†15}, Ioannis Panageas^{†10}, Dimitris Paparas^{†1}, Benjamin Przybocki^{†2}, Bernardo Subercaseaux^{†2}, Ola Svensson^{†13}, Shayan Taherijam^{†10}, Xuan Wu^{†15}, Eylon Yogev^{†16}, Morteza Zadimoghaddam^{†1}, Samson Zhou^{†11}, and Vahab Mirrokni^{*, †, †1}

¹Google Research

²Carnegie Mellon University

³Purdue University

⁴University of Southern California

⁵Harvard University

⁶MIT

⁷Illinois Institute of Technology

⁸Rutgers University

⁹University of Michigan

¹⁰University of California, Irvine

¹¹Texas A&M University

¹²National Institute of Informatics, Tokyo and The University of Tokyo

¹³EPFL

¹⁴University of Maryland, College Park

¹⁵Nanyang Technological University

¹⁶Bar-Ilan University

Abstract

Recent advances in large language models (LLMs) have opened new avenues for accelerating scientific research. While models are increasingly capable of assisting with routine tasks, their ability to contribute to novel, expert-level mathematical discovery is less understood. We present a collection of case studies demonstrating how researchers have successfully collaborated with advanced AI models, specifically Google's Gemini-based models (in particular Gemini Deep Think and its advanced variants), to solve open problems, refute conjectures, and generate new proofs across diverse areas in theoretical computer science, as well as other areas such as economics, optimization, and physics. Based on these experiences, we extract common techniques for effective human-AI collaboration in theoretical research, such as iterative refinement, problem decomposition, and cross-disciplinary knowledge transfer. While the majority of our results stem

Where are We Today?

← Post



CLS 
@ChengleiSi



Automating AI research is exciting! But can LLMs actually produce novel, expert-level research ideas?

After a year-long study, we obtained the first **statistically significant** conclusion: LLM-generated ideas are **more novel** than ideas written by expert human researchers.

Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers

Chenglei Si, Diyi Yang, Tatsunori Hashimoto
Stanford University
{clsi, diyi, thashim}@stanford.edu

Abstract

Recent advancements in large language models (LLMs) have sparked optimism about their potential to accelerate scientific discovery, with a growing number of works proposing research agents that autonomously generate and validate new ideas. Despite this, no evaluations have shown that LLM systems can take the very first step of producing novel, expert-level ideas, let alone perform the entire research process. We address this by establishing an experimental design that evaluates research idea generation while controlling for confounders and performs the first head-to-head comparison between expert NLP researchers and an LLM ideation agent. By recruiting over 100 NLP researchers to write novel ideas and blind reviews of both LLM and human ideas, we obtain the first statistically significant conclusion on current LLM capabilities for research ideation: we find LLM-generated ideas are judged as more novel ($p < 0.05$) than human expert ideas while being judged slightly weaker on feasibility. Studying our agent baselines closely, we identify open problems in building and evaluating research agents, including failures of LLM self-evaluation and their lack of diversity in generation. Finally, we acknowledge that human judgements of novelty can be difficult, even by experts, and propose an end-to-end study design which recruits researchers to execute these ideas into full projects, enabling us to study whether these novelty and feasibility judgements result in meaningful differences in research outcome.¹

 Tatsunori Hashimoto and 2 others

What we Couldn't Cover But Wanted to

- Deep dive on ethical issues (hint: next course)
- Deep dive on LLM eval
- Non-Transformer Architectures
 - Diffusion language models
 - State space models -- Mamba
 - Recursive language models
 - Hybrids of Recurrent/Attention/State-space models
- ...