Attention and its (mis)interpretation

Danish Pruthi

Acknowledgements

Mansi Gupta

Bhuwan Dhingra

Graham Neubig

Zachary C. Lipton

Outline

1. What is attention mechanism?
2. Attention-as-explanations
3. Manipulating attention weights
4. Results and discussion
5. Conclusion

Outline

1. **What is attention mechanism?**
2. Attention-as-explanations
3. Manipulating attention weights
4. Results and discussion
5. Conclusion

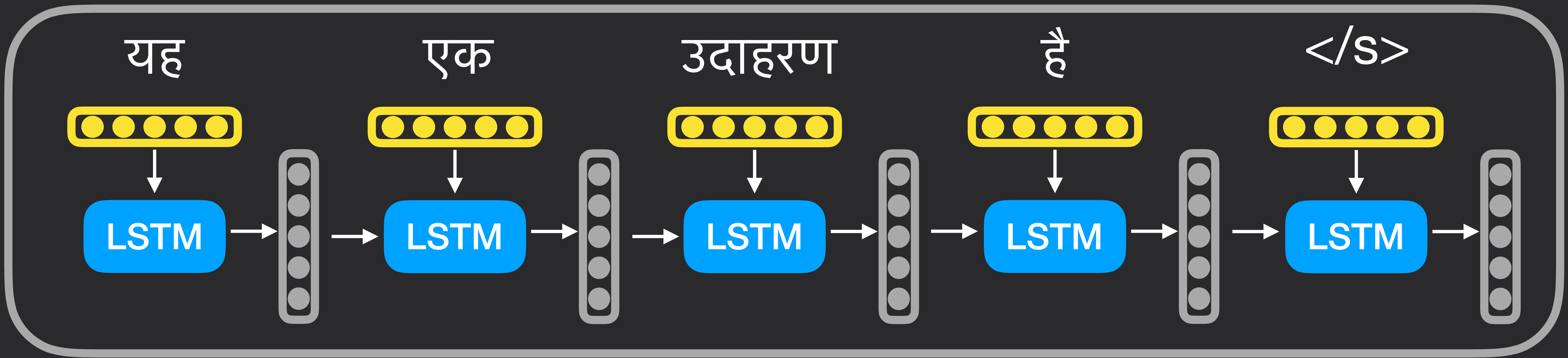
Pre-attention era

Pre-attention era

यह एक उदाहरण है </s>

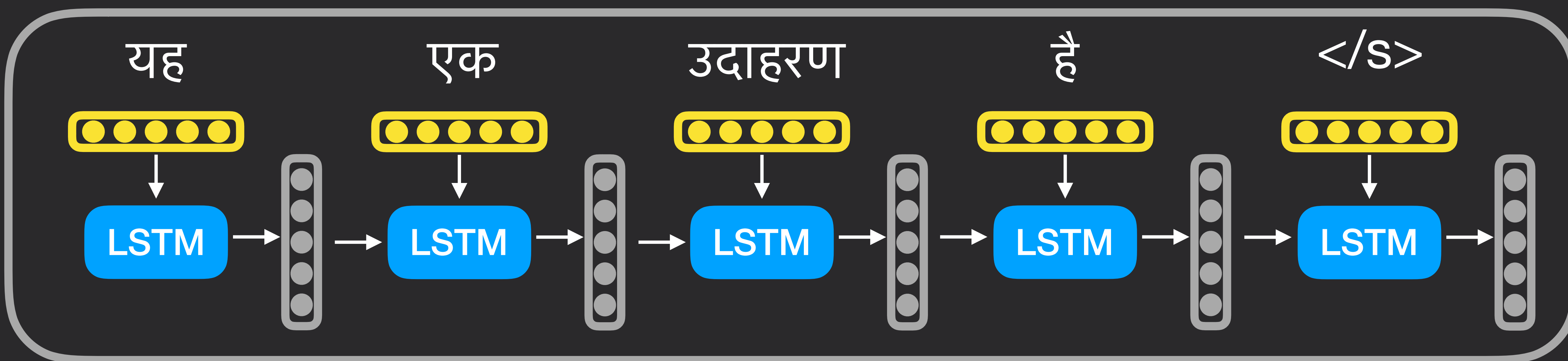
Pre-attention era

Encoder

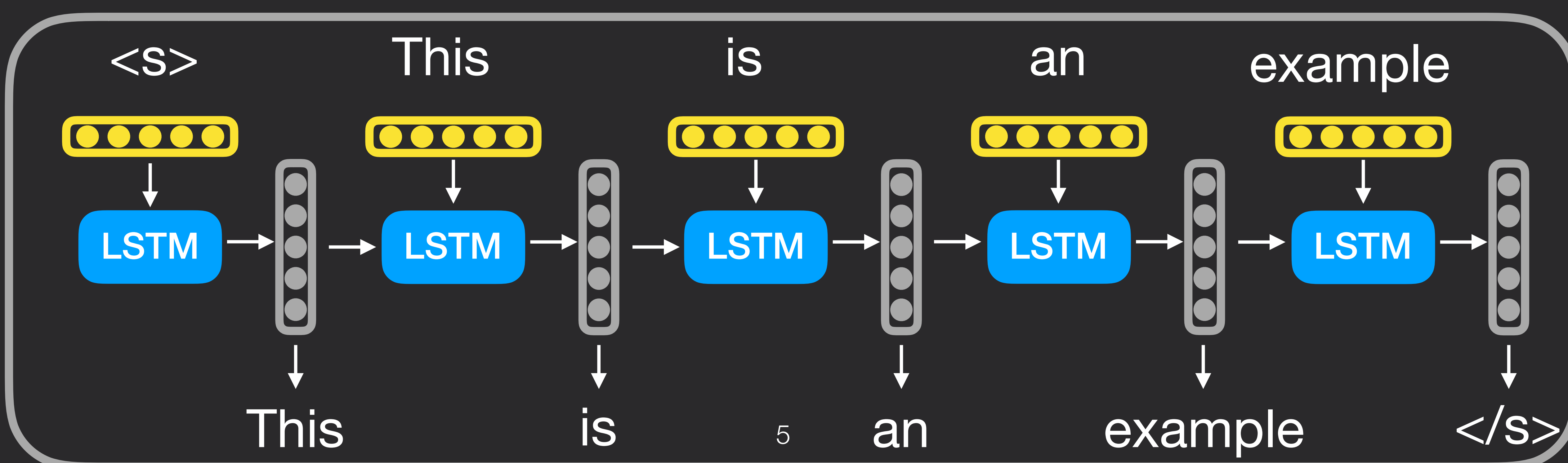


Pre-attention era

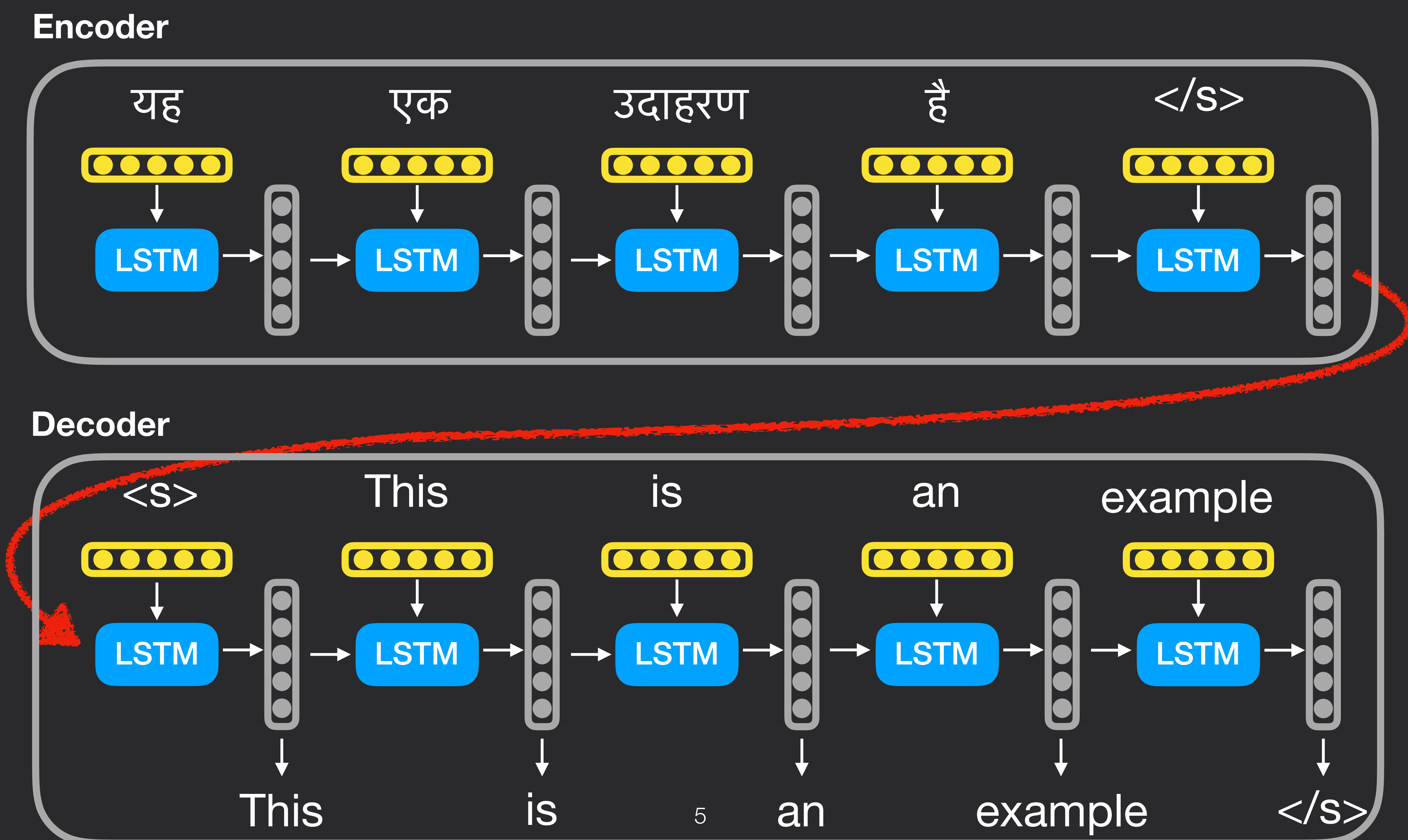
Encoder



Decoder



Pre-attention era



Sentence Representations

Problem: "You can't cram the meaning of a whole %&!$ing sentence into a single $&!*ing vector!" — Ray Mooney

Solution: Use attention *(Bahdanau et al. 2015)*

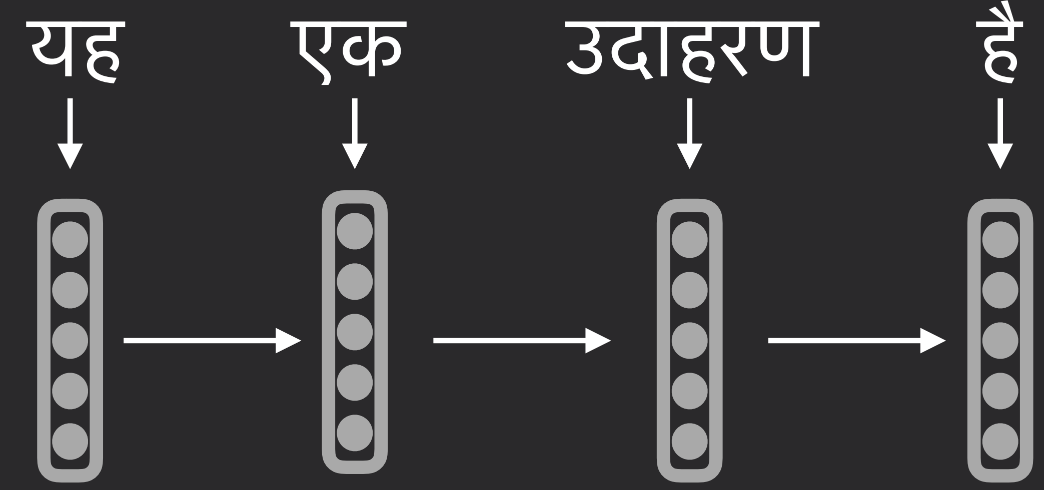
Basic Idea

Bahdanau et al. 2015

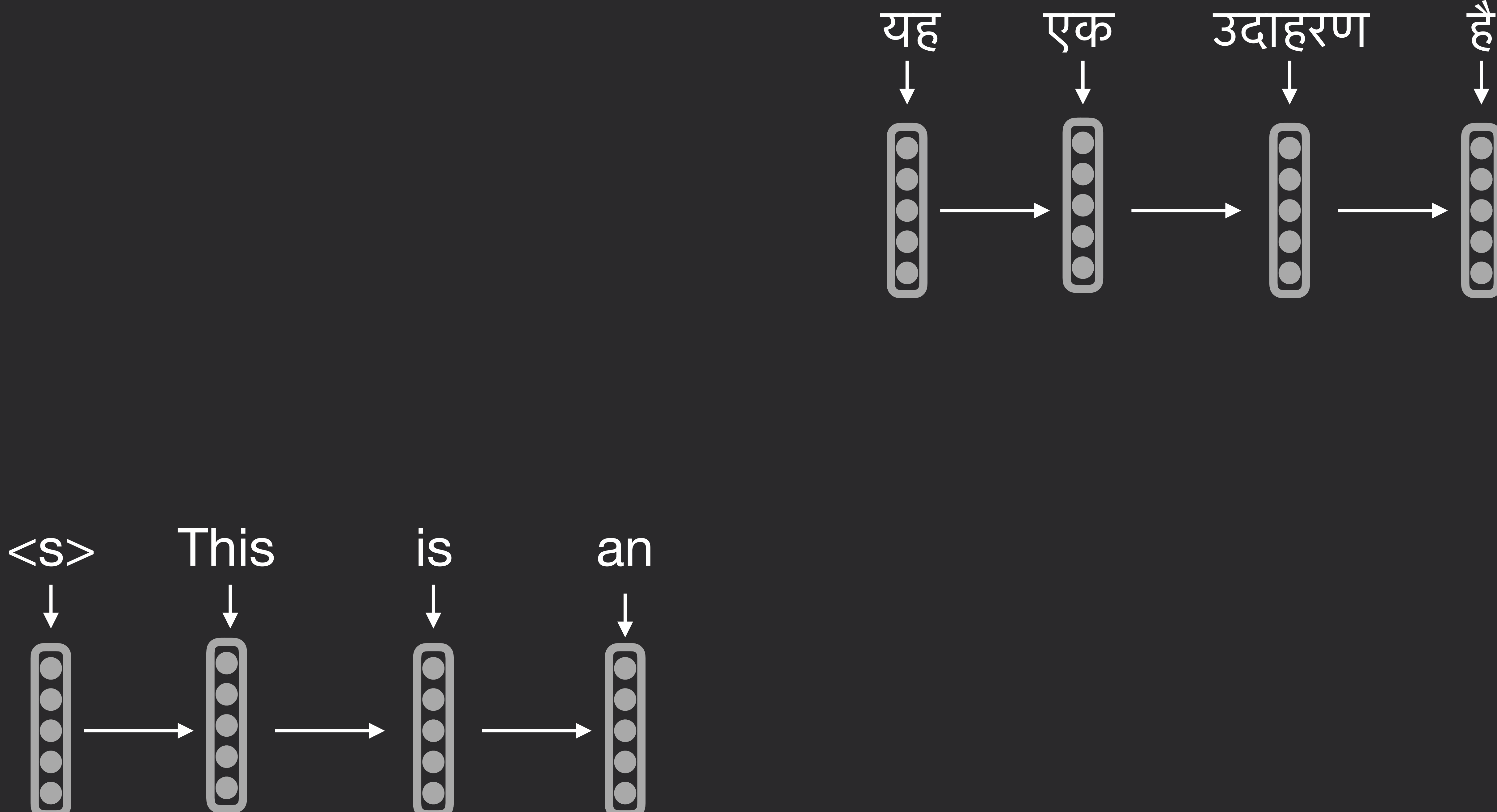
- Encode each word in the sentence into a vector
- When decoding, perform a linear combination of these vectors, weighted by “attention weights”
- Use this combination in picking the next word

Attention

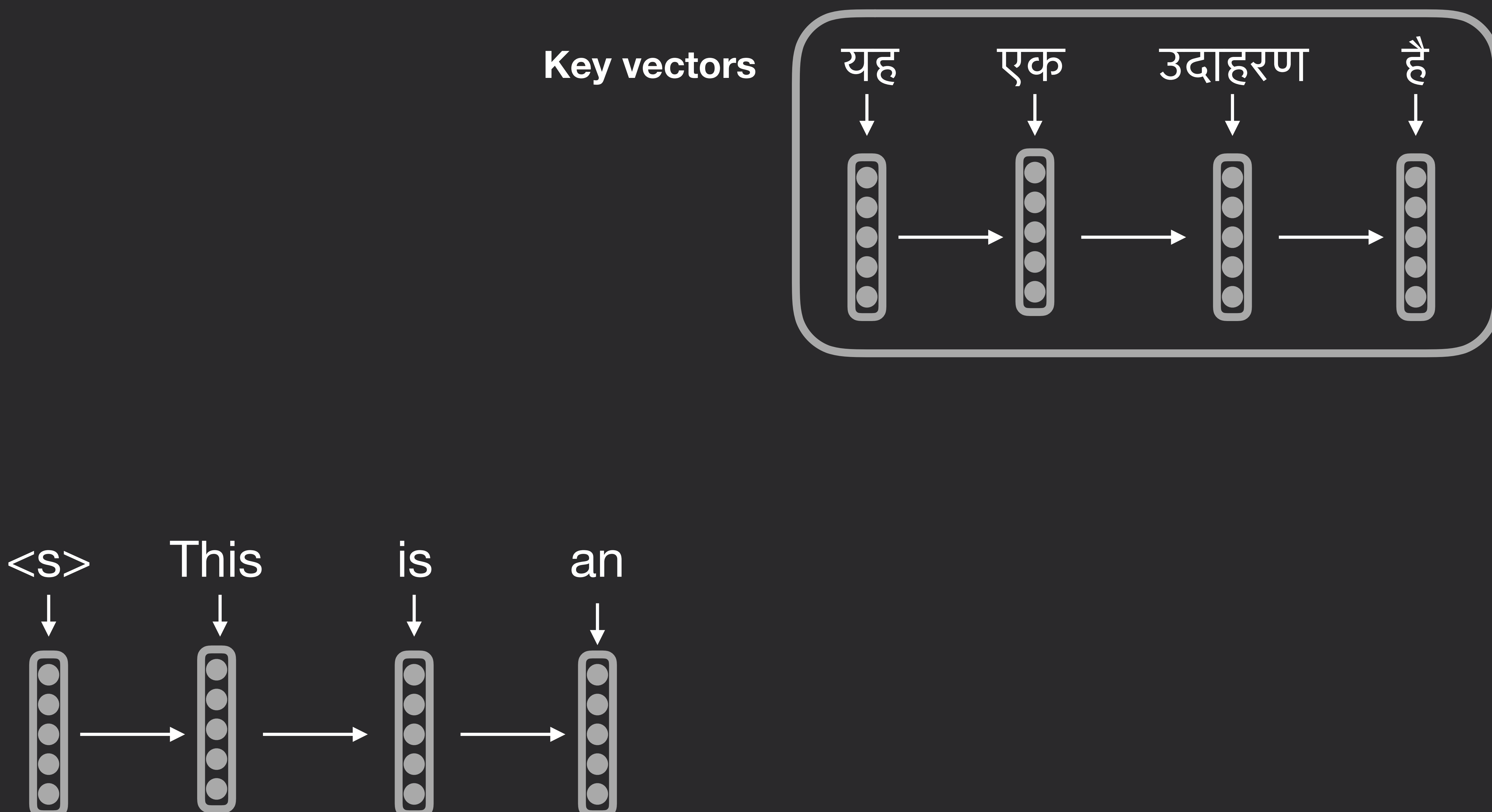
Attention



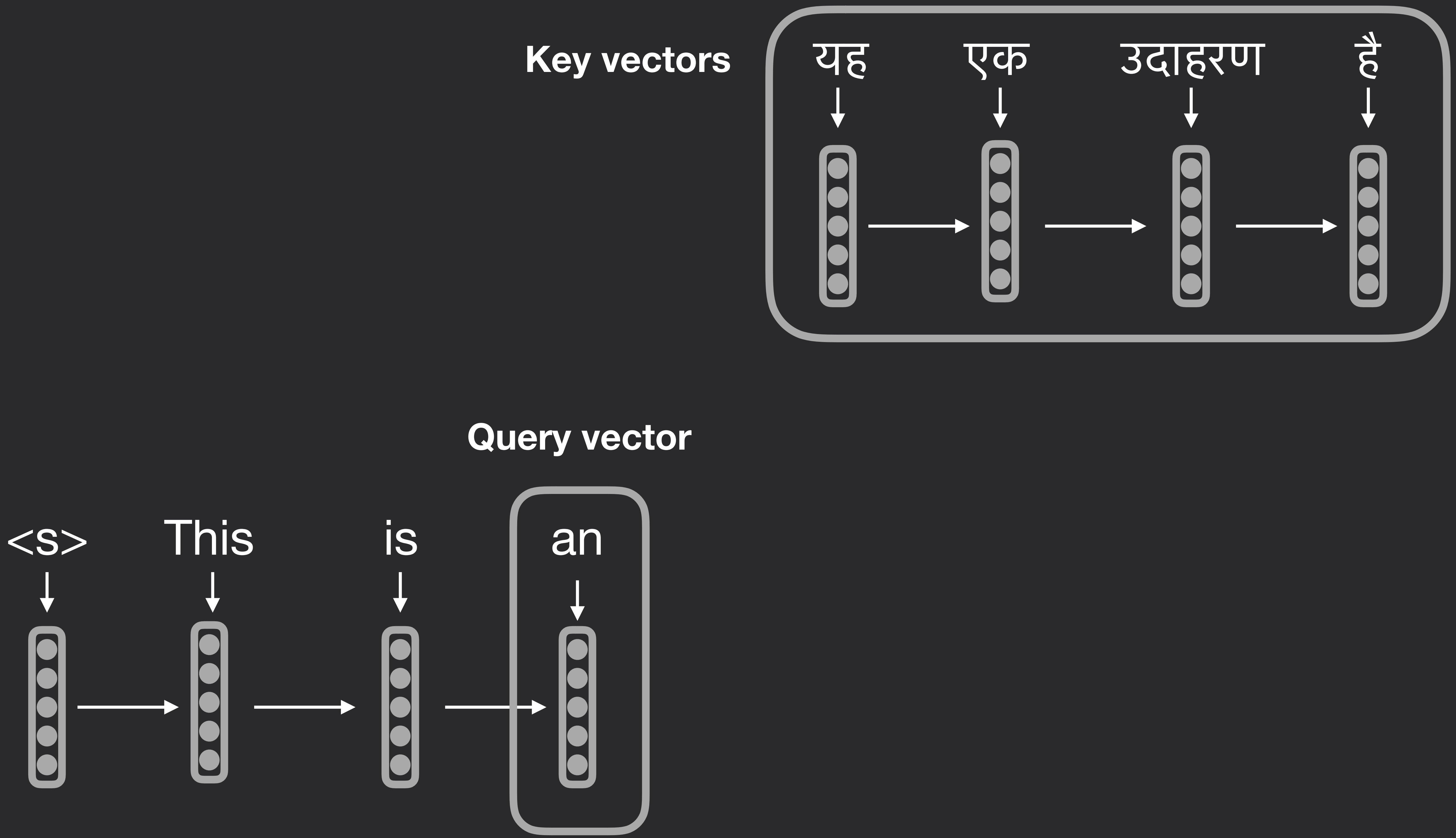
Attention



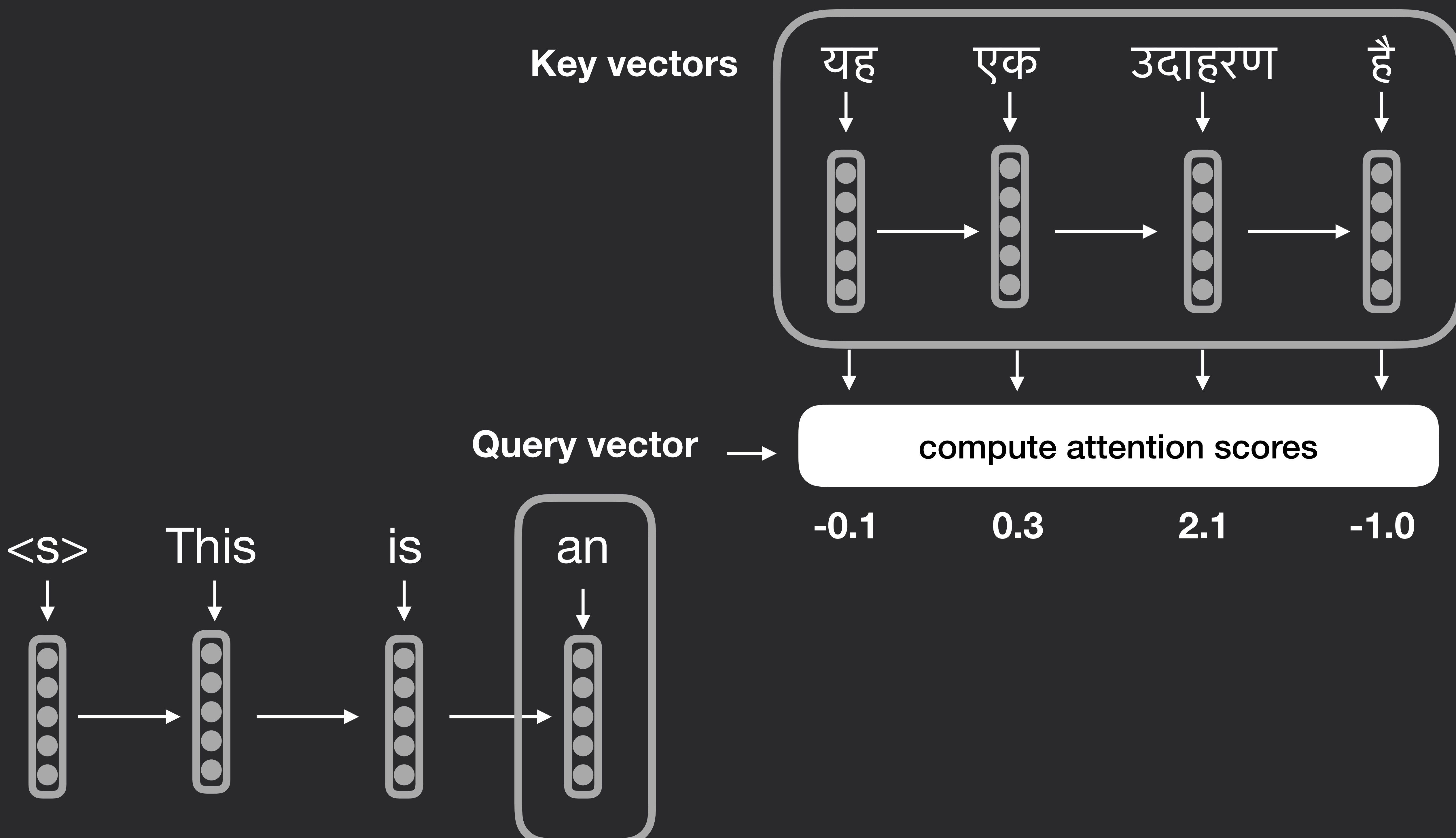
Attention



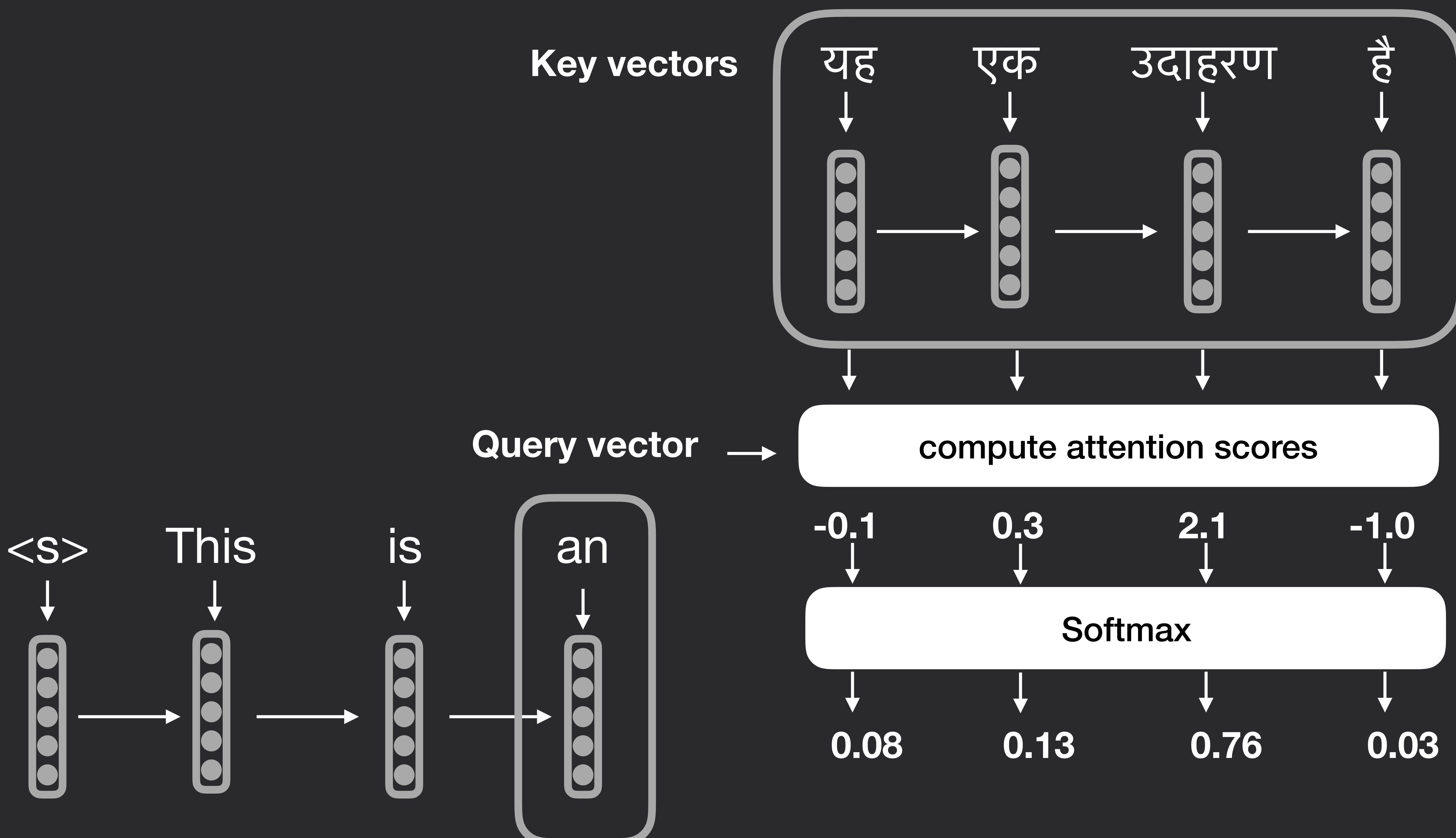
Attention



Attention

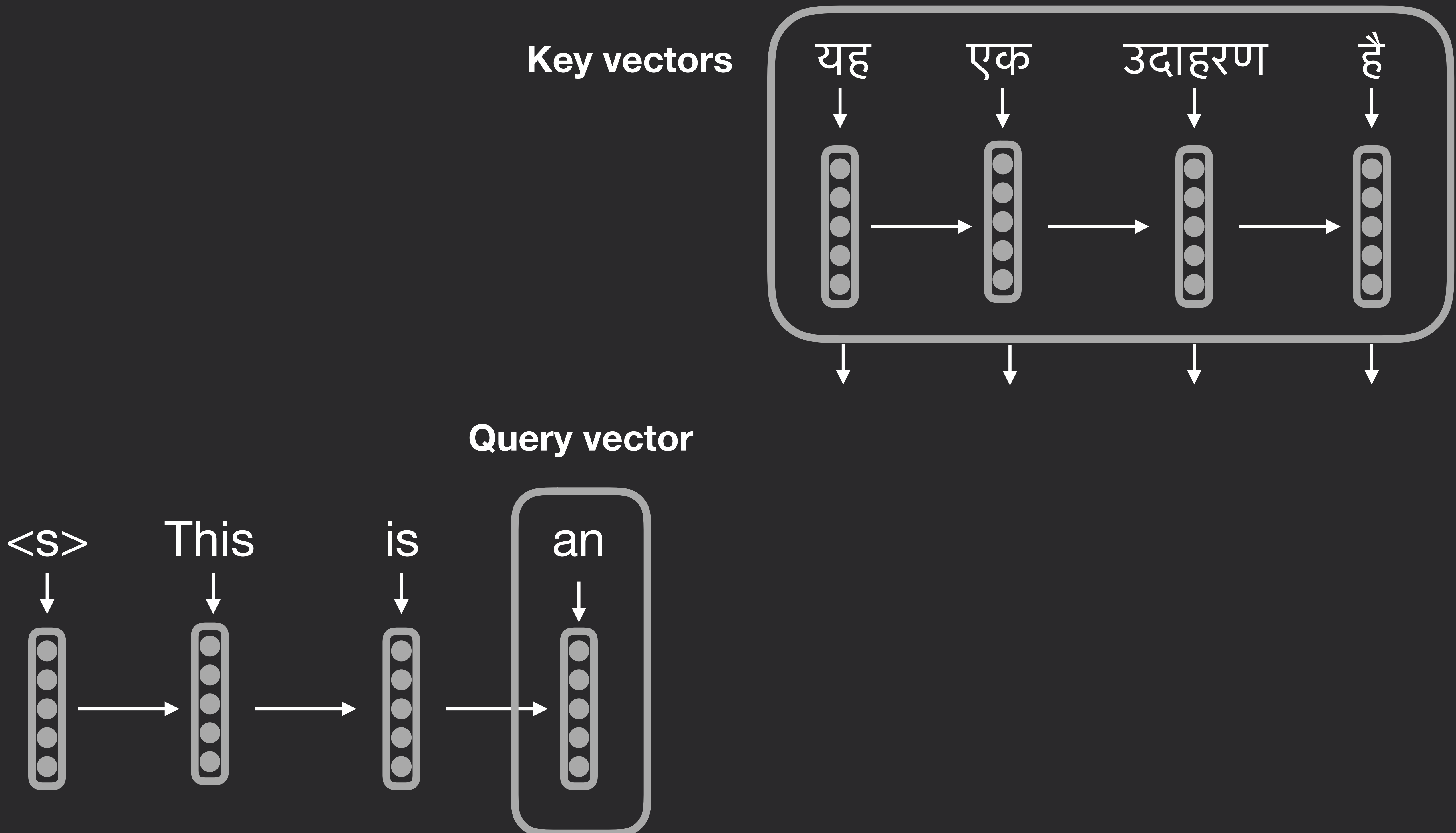

Key vectors
यह
एक
उदाहरण
है
Query vector
compute attention scores
-0.1
0.3
2.1
-1.0
<s>
This
is
an

Attention

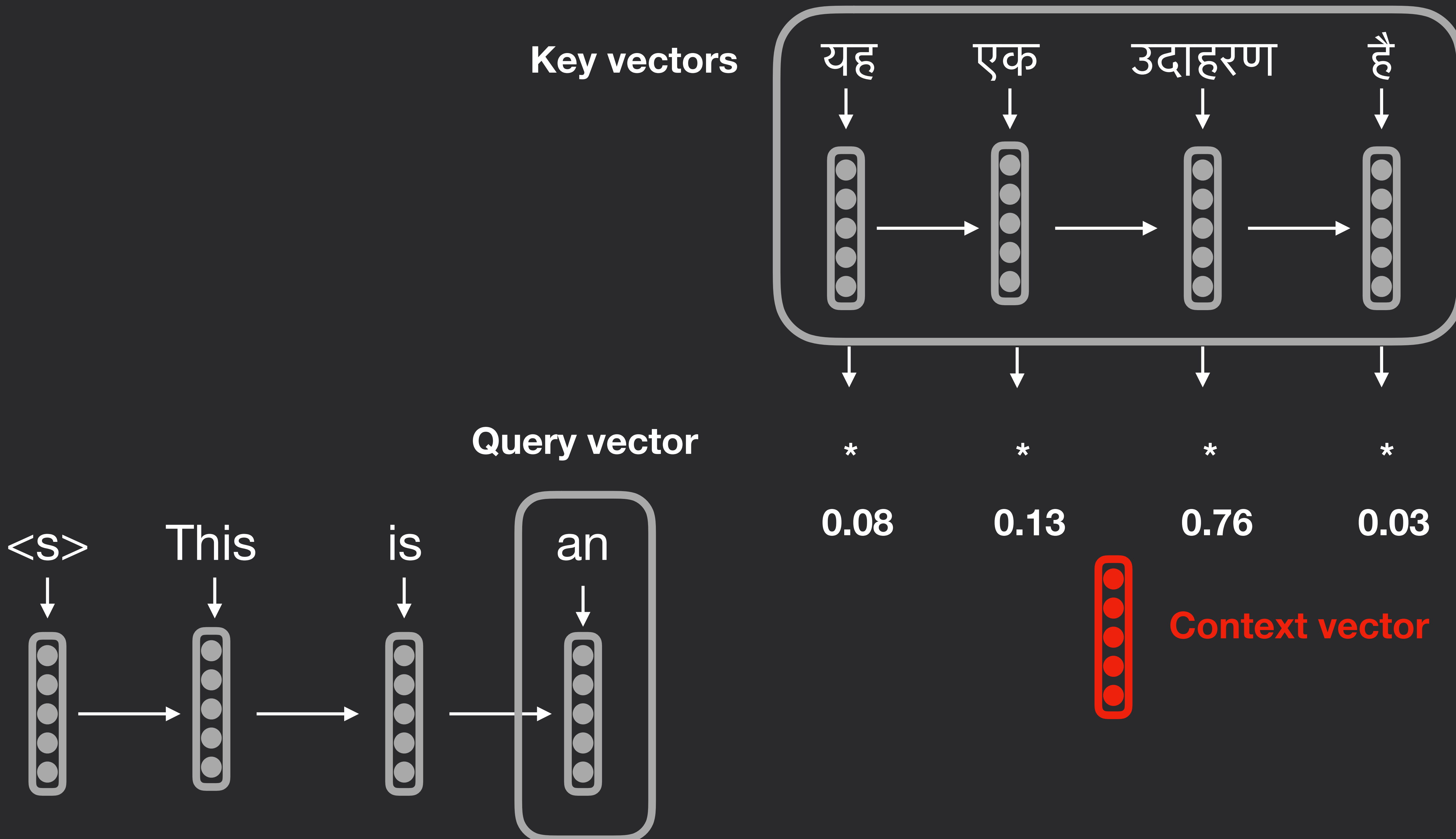


Key vectors
यह
एक
उदाहरण
है
Query vector
compute attention scores
-0.1
0.3
2.1
-1.0
Softmax
0.08
0.13
0.76
0.03
<s>
This
is
an

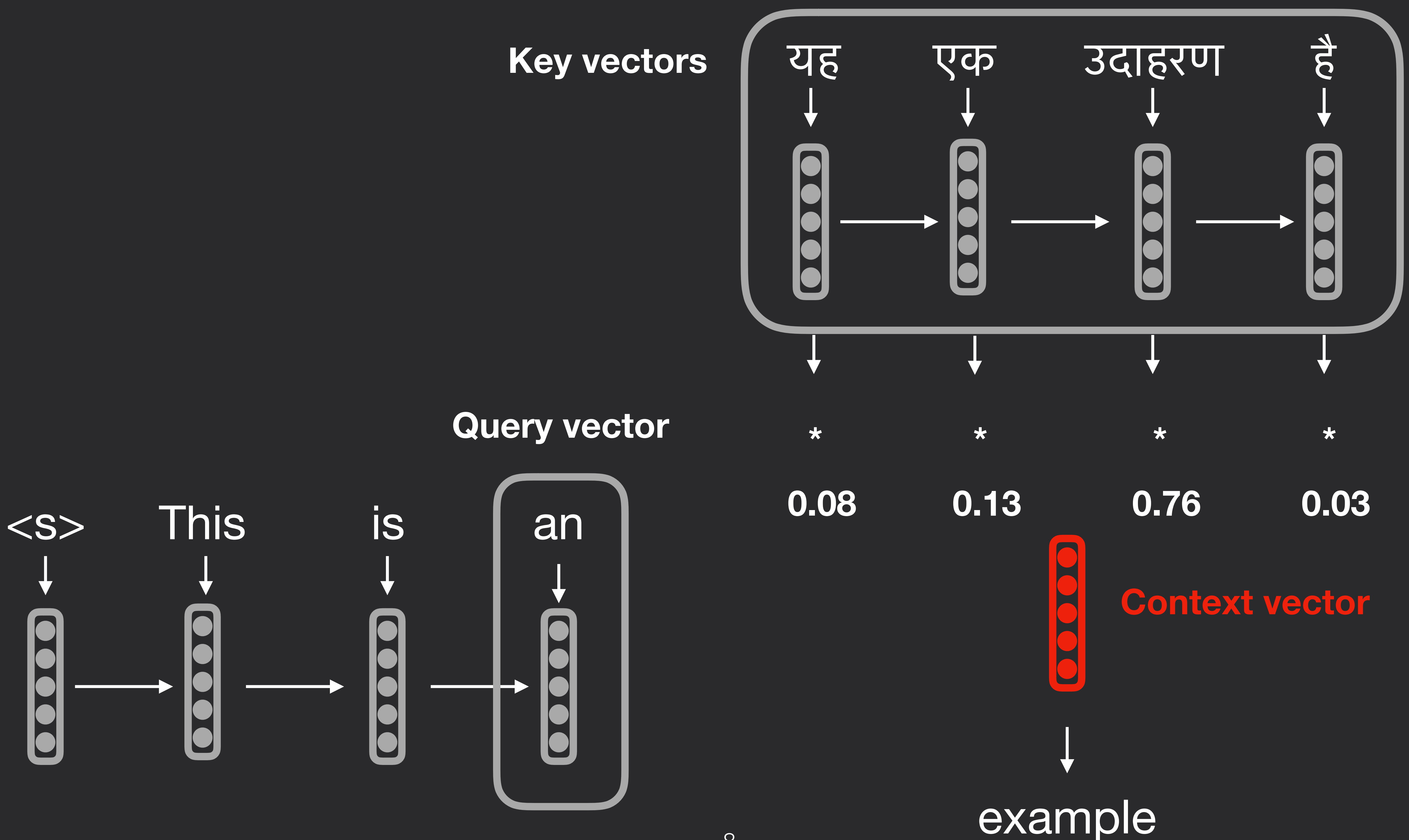
Attention



Attention


Key vectors
यह
एक
उदाहरण
है
Query vector
*
*
*
*
0.08
0.13
0.76
0.03
<s>
This
is
an
Context vector

Attention



Key vectors
यह
एक
उदाहरण
है
Query vector
*
*
*
*
0.08
0.13
0.76
0.03
<s>
This
is
an
Context vector
example

Score Functions

- Dot-Product attention

$$\textbf{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \boldsymbol{s}_t^\top \boldsymbol{h}_i$$

- Bi-linear attention

$$\textbf{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \boldsymbol{s}_t^\top \mathbf{W}_a \boldsymbol{h}_i$$

- MLP attention

$$\textbf{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \mathbf{v}_a^\top \tanh(\mathbf{W}_a[\boldsymbol{s}_t; \boldsymbol{h}_i])$$

- Scaled dot-product attention

$$\textbf{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \frac{\boldsymbol{s}_t^\top \boldsymbol{h}_i}{\sqrt{n}}$$

Score Functions

- Dot-Product attention $\textbf{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \boldsymbol{s}_t^\top \boldsymbol{h}_i$

- Bi-linear attention $\textbf{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \boldsymbol{s}_t^\top \mathbf{W}_a \boldsymbol{h}_i$

- MLP attention $\textbf{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \mathbf{v}_a^\top \tanh(\mathbf{W}_a [\boldsymbol{s}_t; \boldsymbol{h}_i])$

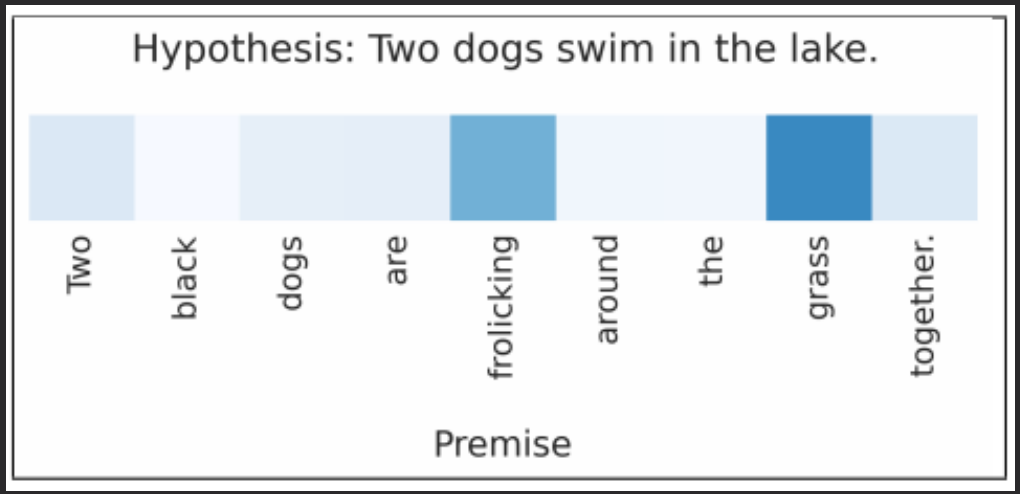
- Scaled dot-product attention $\textbf{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \dfrac{\boldsymbol{s}_t^\top \boldsymbol{h}_i}{\sqrt{n}}$

Outline

1. What is attention mechanism?
2. Attention-as-explanations
3. Manipulating attention weights
4. Results and discussion
5. Conclusion

Attention as explanation

- Used by model-developers to explain models' predictions

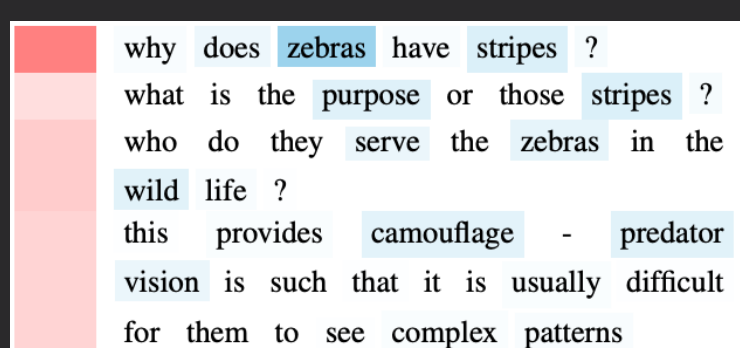


Entailment
Rocktäschel et al, 2015

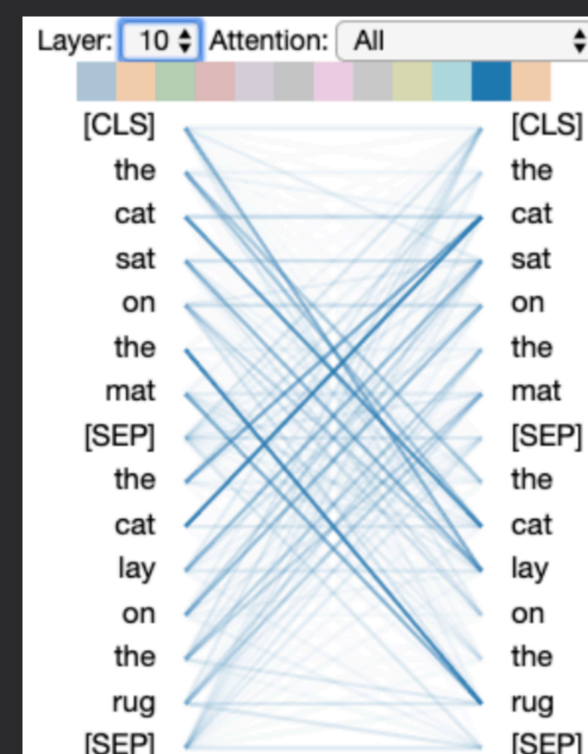
Image captioning
Xu et al, 2015

Attention as explanation

- Used by model-developers to explain models' predictions



Document classification
Yang et al, 2016



BERTViz
Vig et al, 2019

and many others…

Attention as explanation

- Used by model-developers to explain models' predictions

"By inspecting the network’s attention, for instance by visually highlighting attention weights, one could attempt to investigate and understand the outcome of neural networks. Hence, weight visualization is now common practice."

Galassi et al., 2019

Attention as explanation

- Used by model-developers to explain models' predictions
- Used by practitioners to audit models for bias, fairness, accountability, etc

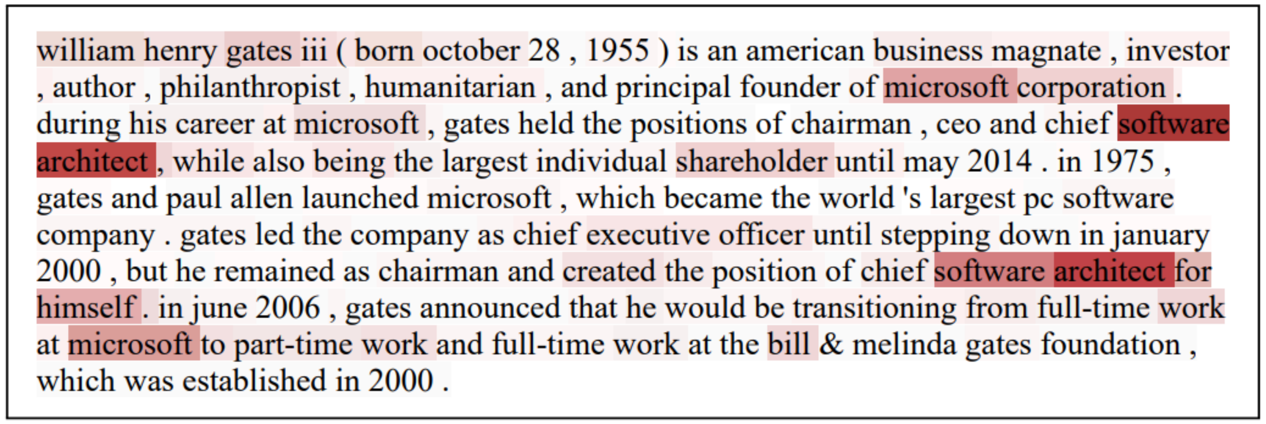


Figure 7: Visualization of the DNN's per-token attention weights. Predicted label (i.e., occupation): *software engineer.*

Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting, De-Arteaga, et al, 2019

Attention-as-explanation in FAT* contexts

** Fairness, accountability and transparency*

- Use attention mechanism to identify gender bias in occupation prediction models used as a part of high-stakes job recommendation models

De-Arteaga et al., 2019

Attention-as-explanation in FAT* contexts

** Fairness, accountability and transparency*

- Use attention mechanism to identify gender bias in occupation prediction models used as a part of high-stakes job recommendation models

"The attention weights indicate which tokens are the most predictive"

De-Arteaga et al., 2019

Attention-as-explanation in FAT* contexts

** Fairness, accountability and transparency*

- Use attention mechanism to identify gender bias in occupation prediction models used as a part of high-stakes job recommendation models

"The attention weights indicate which tokens are the most predictive"

We question this assumption: does attention *necessarily* indicate most predictive tokens?

De-Arteaga et al., 2019

Outline

1. What Is attention mechanism?
2. Attention-as-explanations in the FAT* context
   ** Fairness, accountability and transparency*
3. **Manipulating attention weights**
4. Results and discussion
5. Conclusion

Setup

- Setup tasks such that we know certain features *a-priori* to be useful for prediction
- Measure *“attention mass”* on these tokens
- Examine if the models can be manipulated
  - What is the price to pay?

Classification Tasks

Classification Tasks

Task	Input Example

Classification Tasks

Task	Input Example
Occupation Prediction (*Physician vs Surgeon*)	Ms. X practices medicine in Memphis, TN. Ms. X speaks English and Spanish.

Classification Tasks

Task	Input Example
Occupation Prediction (*Physician vs Surgeon*)	Ms. X practices medicine in Memphis, TN. Ms. X speaks English and Spanish.
Gender Identification	After that, Austen was educated at home until she went to boarding school early in 1785

Classification Tasks

Task	Input Example
Occupation Prediction (*Physician vs Surgeon*)	Ms. X practices medicine in Memphis, TN. Ms. X speaks English and Spanish.
Gender Identification	After that, Austen was educated at home until she went to boarding school early in 1785
Sentiment Analysis *(SST + Wikipedia)*	Good acting, good dialogue, good cinematography. Helen Reddy is an Australian singer and activist.

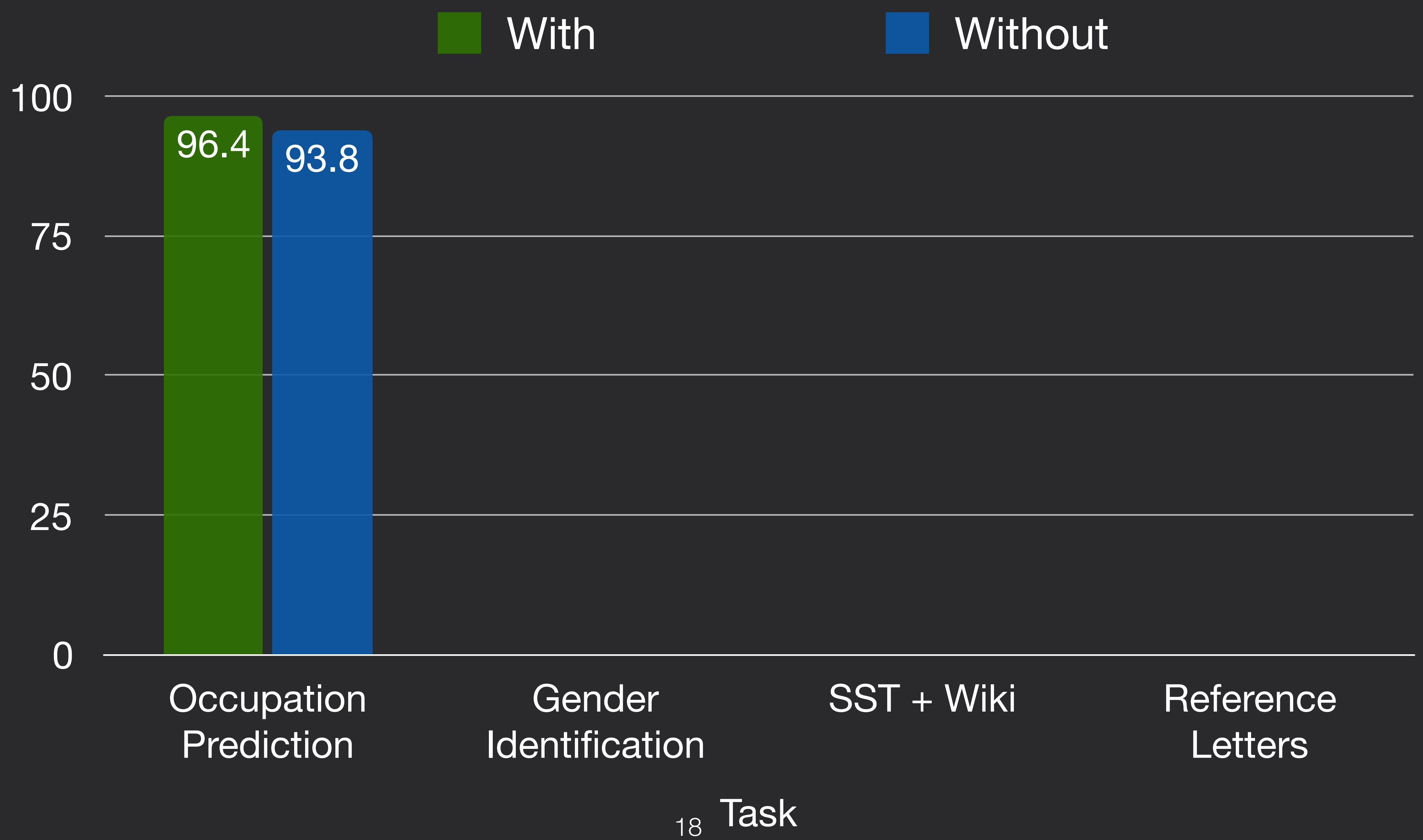
Classification Tasks

Task	Input Example
Occupation Prediction (*Physician vs Surgeon*)	Ms. X practices medicine in Memphis, TN. Ms. X speaks English and Spanish.
Gender Identification	After that, Austen was educated at home until she went to boarding school early in 1785
Sentiment Analysis *(SST + Wikipedia)*	Good acting, good dialogue, good cinematography. Helen Reddy is an Australian singer and activist.
Acceptance Prediction *(Reference Letters)*	It is with pleasure that I am writing this letter...I highly recommend her for your institution. Percentile:99.0 Rank:Extraordinary.

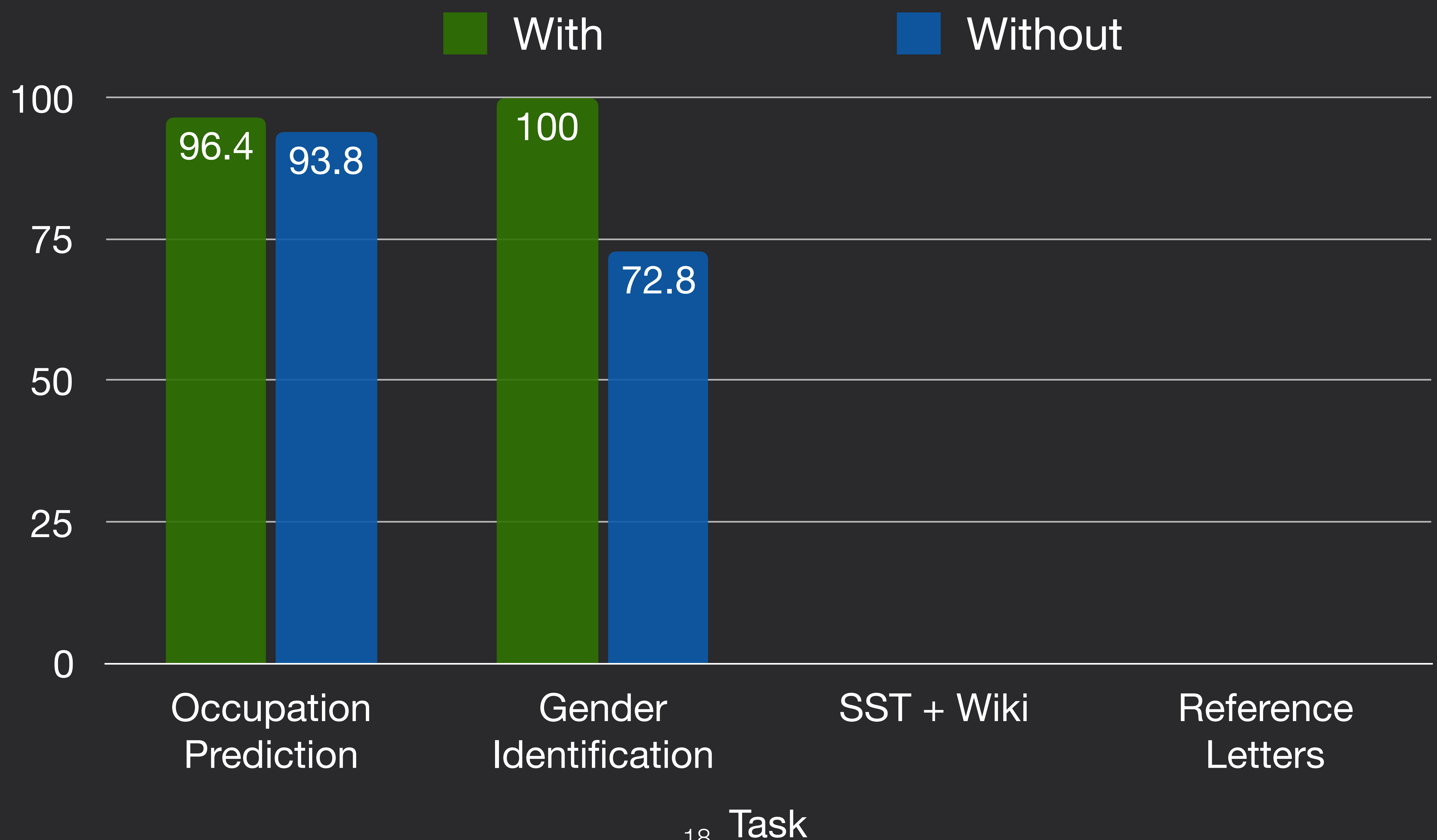
Need for impermissible tokens

Need for impermissible tokens

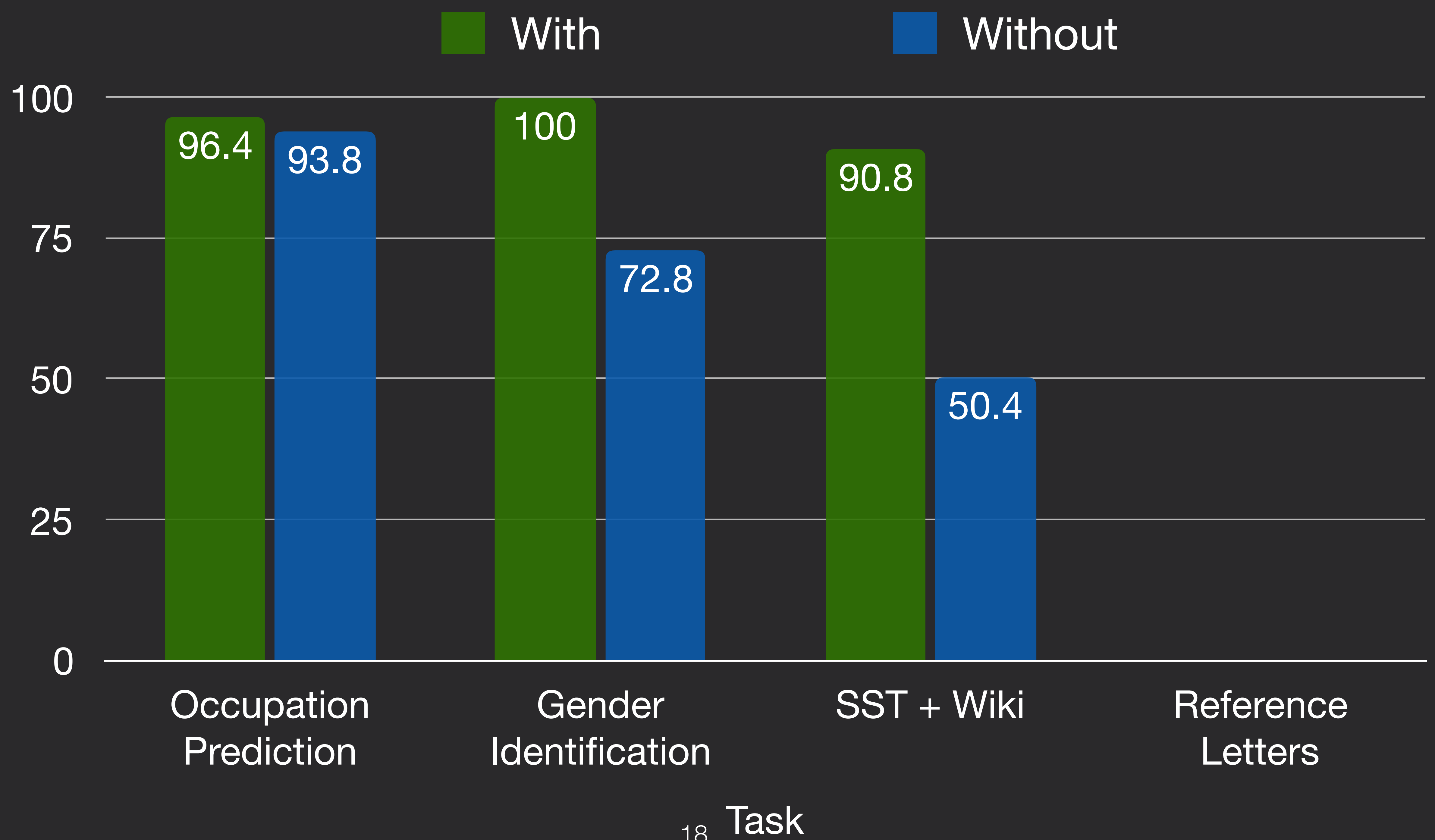
Need for impermissible tokens



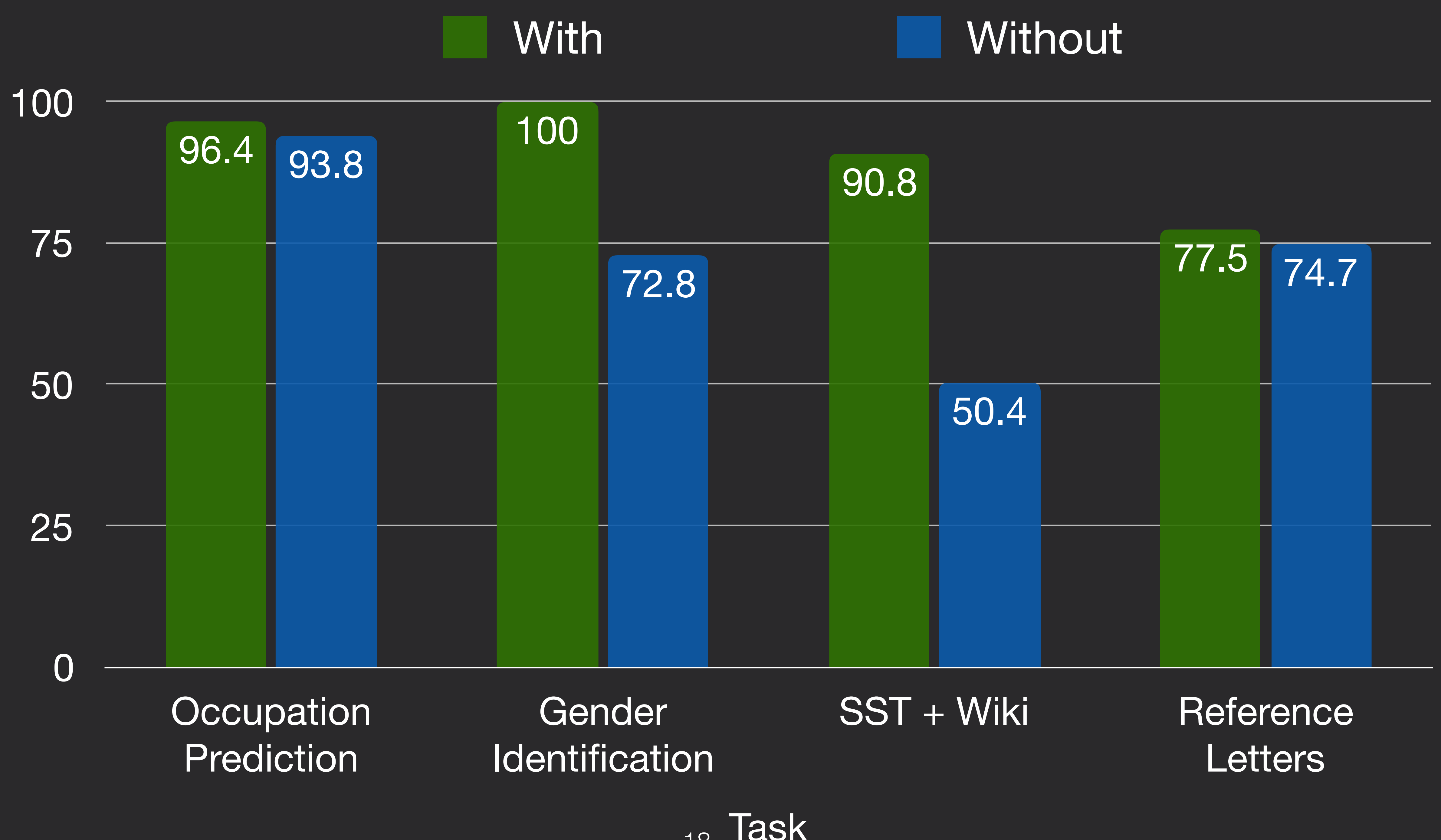
Need for impermissible tokens



Need for impermissible tokens



Need for impermissible tokens



Sequence-to-sequence Tasks

Sequence-to-sequence Tasks

Task	Example

Sequence-to-sequence Tasks

Task	Example
Bigram Flipping	$\{w_1, w_2 \ldots w_{2n-1}, w_{2n}\} \rightarrow \{w_2, w_1, \ldots w_{2n}, w_{2n-1}\}$

Sequence-to-sequence Tasks

Task	Example
Bigram Flipping	$\{w_1, w_2 \ldots w_{2n-1}, w_{2n}\} \rightarrow \{w_2, w_1, \ldots w_{2n}, w_{2n-1}\}$
Sequence Copying	$\{w_1, w_2, \ldots w_{n-1}, w_n\} \rightarrow \{w_1, w_2, \ldots w_n, w_{n-1}\}$

Sequence-to-sequence Tasks

Task	Example
Bigram Flipping	$\{w_1, w_2 \ldots w_{2n-1}, w_{2n}\} \rightarrow \{w_2, w_1, \ldots w_{2n}, w_{2n-1}\}$
Sequence Copying	$\{w_1, w_2, \ldots w_{n-1}, w_n\} \rightarrow \{w_1, w_2, \ldots w_n, w_{n-1}\}$
Sequence Reversal	$\{w_1, w_2, \ldots w_{n-1}, w_n\} \rightarrow \{w_n, w_{n-1}, \ldots w_2, w_1\}$

Sequence-to-sequence Tasks

Task	Example
Bigram Flipping	$\{w_1, w_2 \ldots w_{2n-1}, w_{2n}\} \rightarrow \{w_2, w_1, \ldots w_{2n}, w_{2n-1}\}$
Sequence Copying	$\{w_1, w_2, \ldots w_{n-1}, w_n\} \rightarrow \{w_1, w_2, \ldots w_n, w_{n-1}\}$
Sequence Reversal	$\{w_1, w_2, \ldots w_{n-1}, w_n\} \rightarrow \{w_n, w_{n-1}, \ldots w_2, w_1\}$
English - German MT	This is an example. → Dieser ist ein Beispiel.

Manipulating Attention

Manipulating Attention

- Let $\mathcal{I}$ be the impermissible tokens, ***m*** is the mask

$$\mathbf{m}_i = \begin{cases} 1, & \text{if } w_i \in \mathcal{I} \\ 0 & \text{otherwise} \end{cases}$$

Manipulating Attention

- Let $\mathcal{I}$ be the impermissible tokens, ***m*** is the mask

$$\mathbf{m}_i = \begin{cases} 1, & \text{if } w_i \in \mathcal{I} \\ 0 & \text{otherwise} \end{cases}$$

- For any task-specific loss function, a penalty term is added

$$\mathcal{L}' = \mathcal{L} + \mathcal{R}$$

Manipulating Attention

- Let $\mathcal{I}$ be the impermissible tokens, ***m*** is the mask

$$\mathbf{m}_i = \begin{cases} 1, & \text{if } w_i \in \mathcal{I} \\ 0 & \text{otherwise} \end{cases}$$

- For any task-specific loss function, a penalty term is added

$$\mathcal{L}' = \mathcal{L} + \mathcal{R}$$

- The penalty term penalizes the model for allocating attention to impermissible tokens

$$\mathcal{R} = -\lambda \log\left(1 - \boldsymbol{\alpha}^T \mathbf{m}\right)$$

Manipulating Attention

$$\mathcal{R} = -\lambda \log\left(1 - \boldsymbol{\alpha}^T \mathbf{m}\right)$$

Manipulating Attention

$$\mathcal{R} = -\lambda \log\left(\boxed{1 - \boldsymbol{\alpha}^T \mathbf{m}}\right)$$

Total attention mass
on all the "allowed" tokens

Manipulating Attention

Penalty coefficient that modulates attention on *impermissible* tokens

$$\mathcal{R} = \boxed{-\lambda} \log\left(\boxed{1 - \boldsymbol{\alpha}^T \mathbf{m}}\right)$$

Total attention mass on all the "allowed" tokens

Manipulating Attention

Penalty coefficient that modulates attention on *impermissible* tokens

$$\mathcal{R} = \boxed{-\lambda} \log\left(\boxed{1 - \boldsymbol{\alpha}^T \mathbf{m}}\right)$$

Total attention mass on all the "allowed" tokens

- Side note: In a parallel work, *Wiegreffe and Pinter (2019)* propose a different penalty term

$$\mathcal{R}' = -\lambda \mathrm{KL}(\boldsymbol{\alpha}_{\text{new}} \parallel \boldsymbol{\alpha}_{\text{old}})$$

Manipulating Attention

Manipulating Attention

- Multiple attention heads

Manipulating Attention

- Multiple attention heads
  - Optimizing the mean over a set of attention heads

$$\mathcal{R} = -\frac{\lambda}{|\mathcal{H}|}\sum_{h\in\mathcal{H}}\log\left(1-\boldsymbol{\alpha}_h^T\mathbf{m}\right)$$

Manipulating Attention

- Multiple attention heads
  - Optimizing the mean over a set of attention heads

$$\mathcal{R} = -\frac{\lambda}{|\mathcal{H}|} \sum_{h \in \mathcal{H}} \log\left(1 - \boldsymbol{\alpha}_h^T \mathbf{m}\right)$$

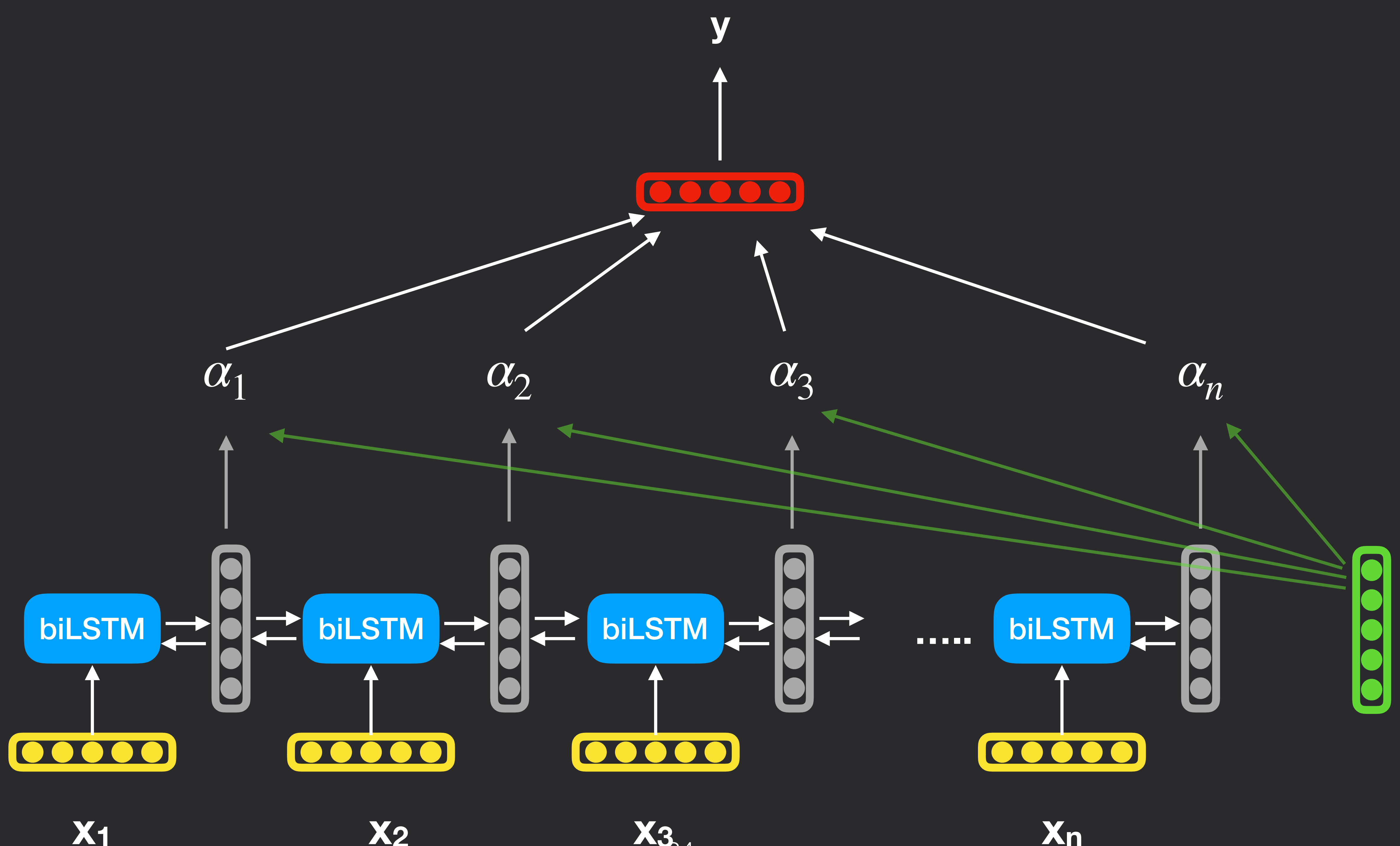
  - One of the attention heads can be assigned a large amount of attention to impermissible tokens

$$\mathcal{R} = -\lambda \cdot \min_{h \in \mathcal{H}} \log\left(1 - \boldsymbol{\alpha}_h^T \mathbf{m}\right)$$

Outline

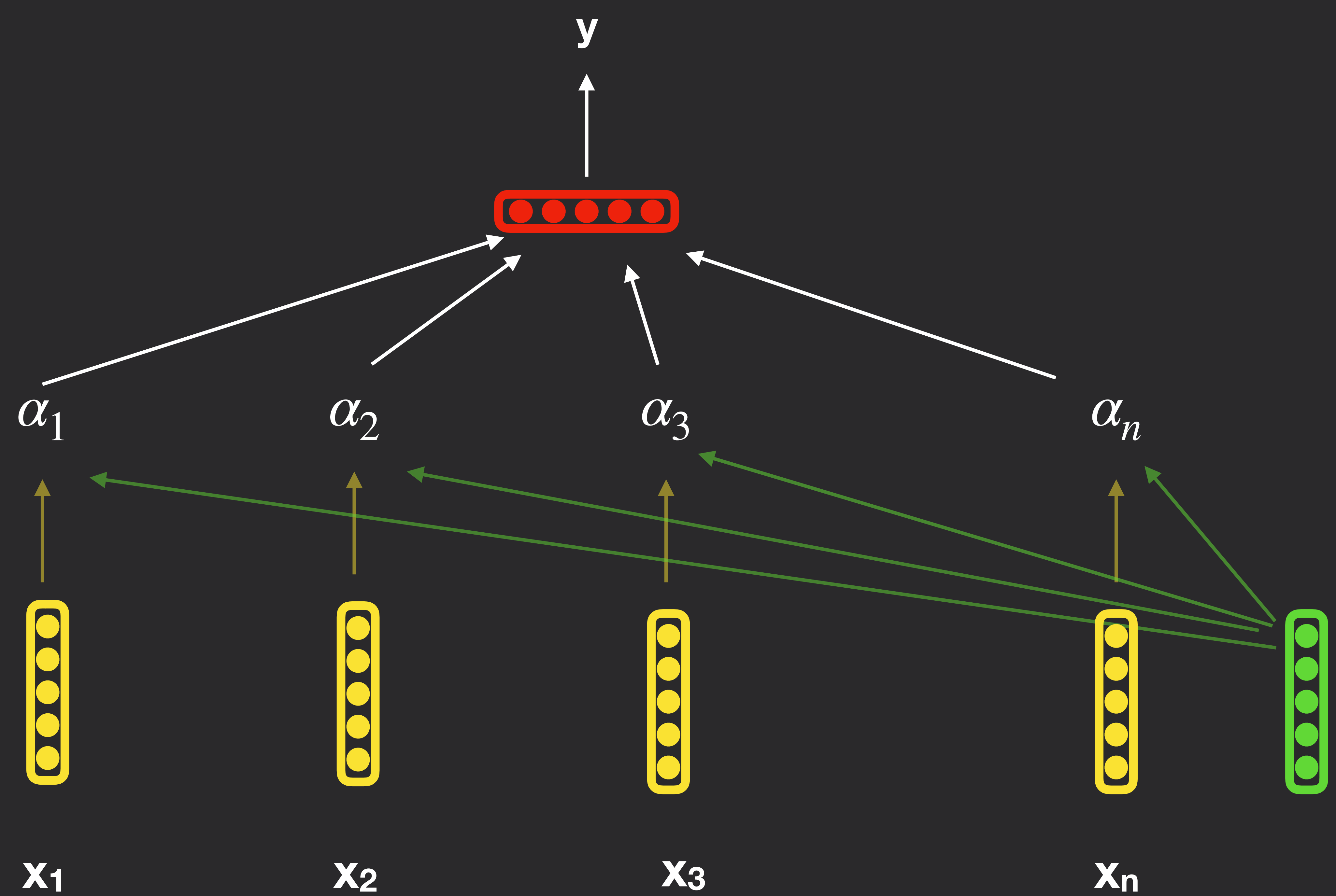
1. What Is attention mechanism?
2. Attention-as-explanations
3. Manipulating attention weights
4. **Results and discussion**
5. Conclusion

BiLSTM + Attention

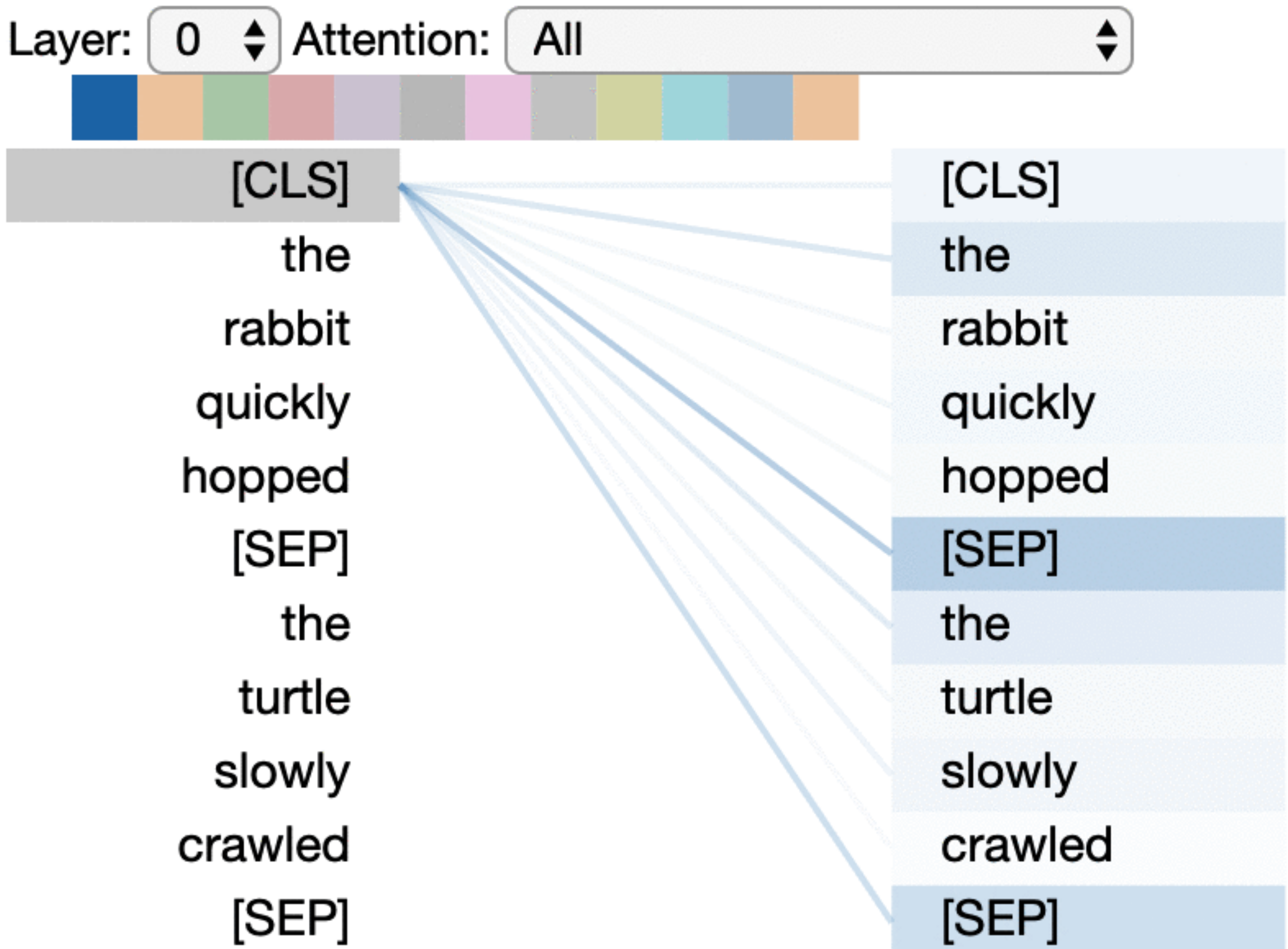


Embedding + Attention

(No recurrent connections)

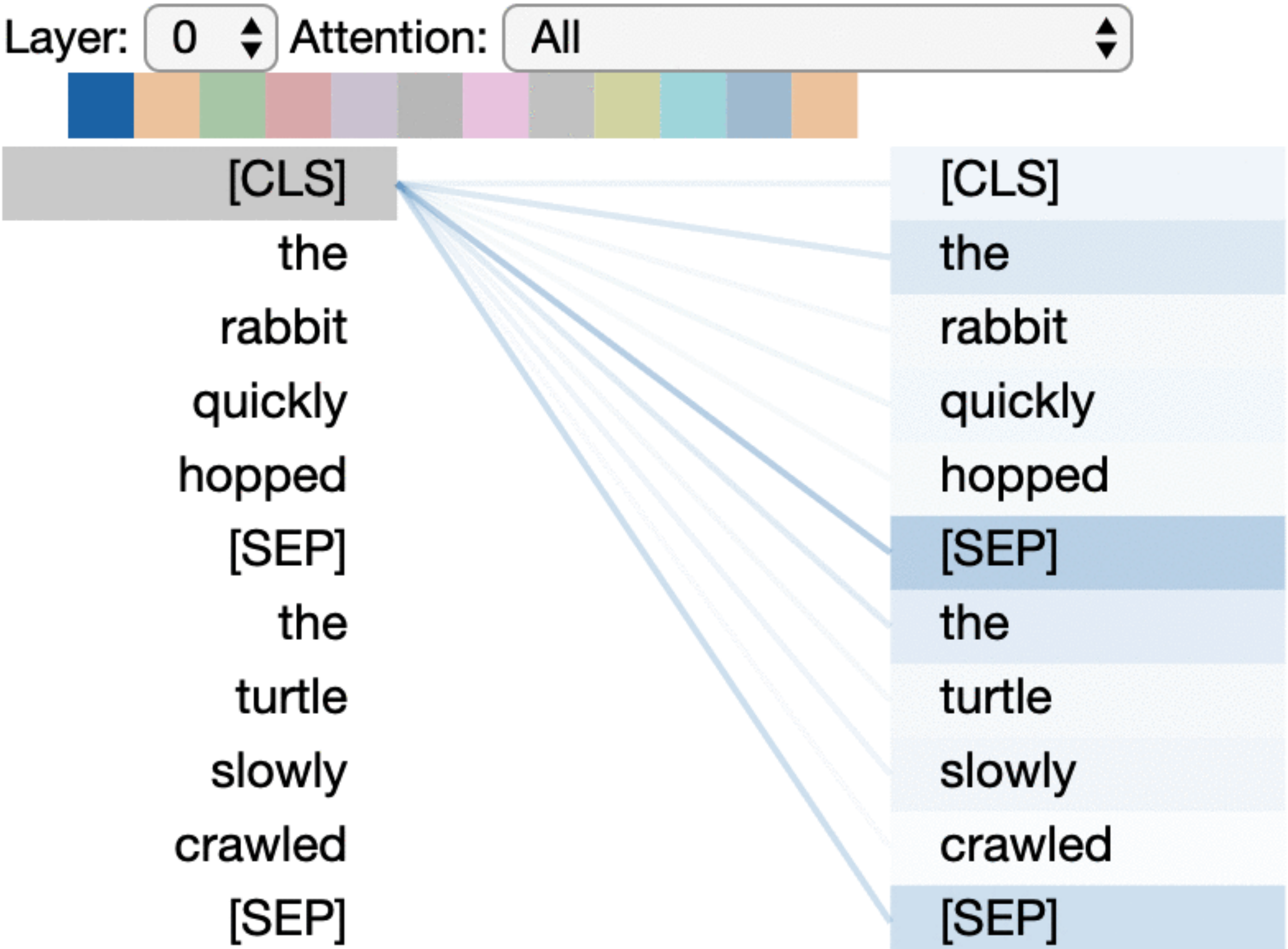


Transformer-based Model



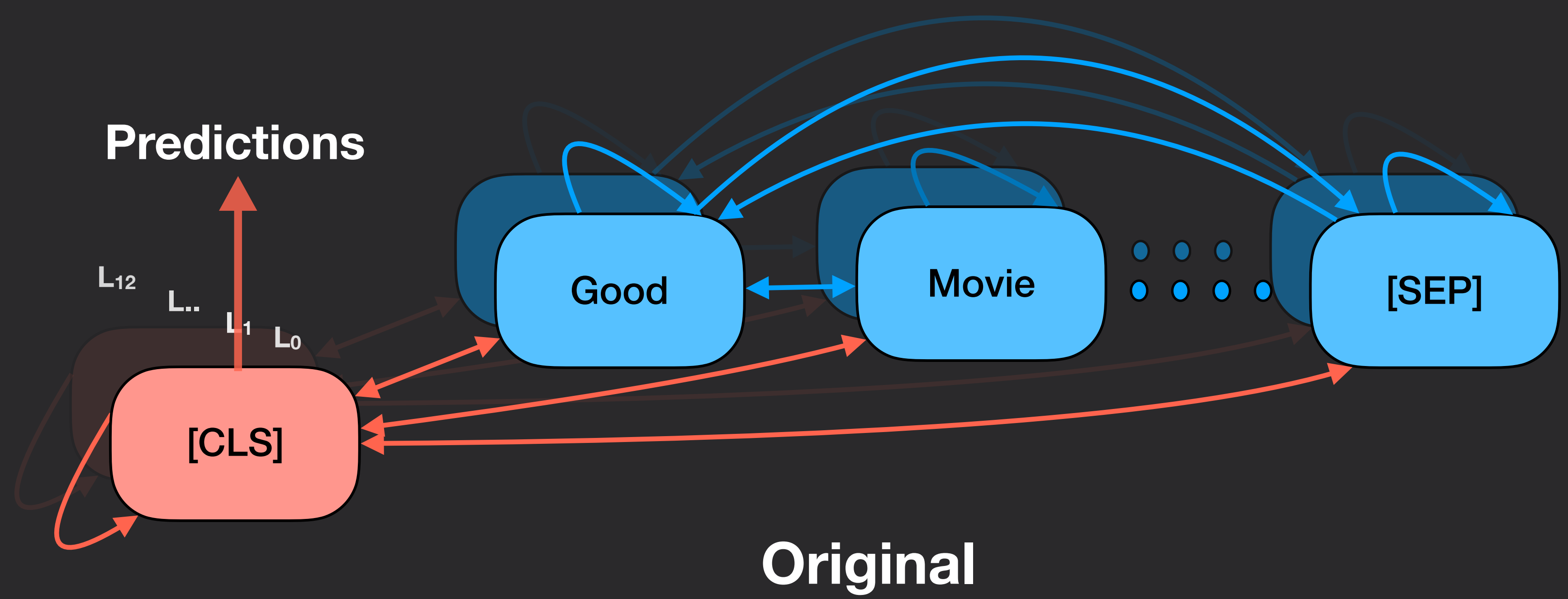
Devlin et. al

Transformer-based Model

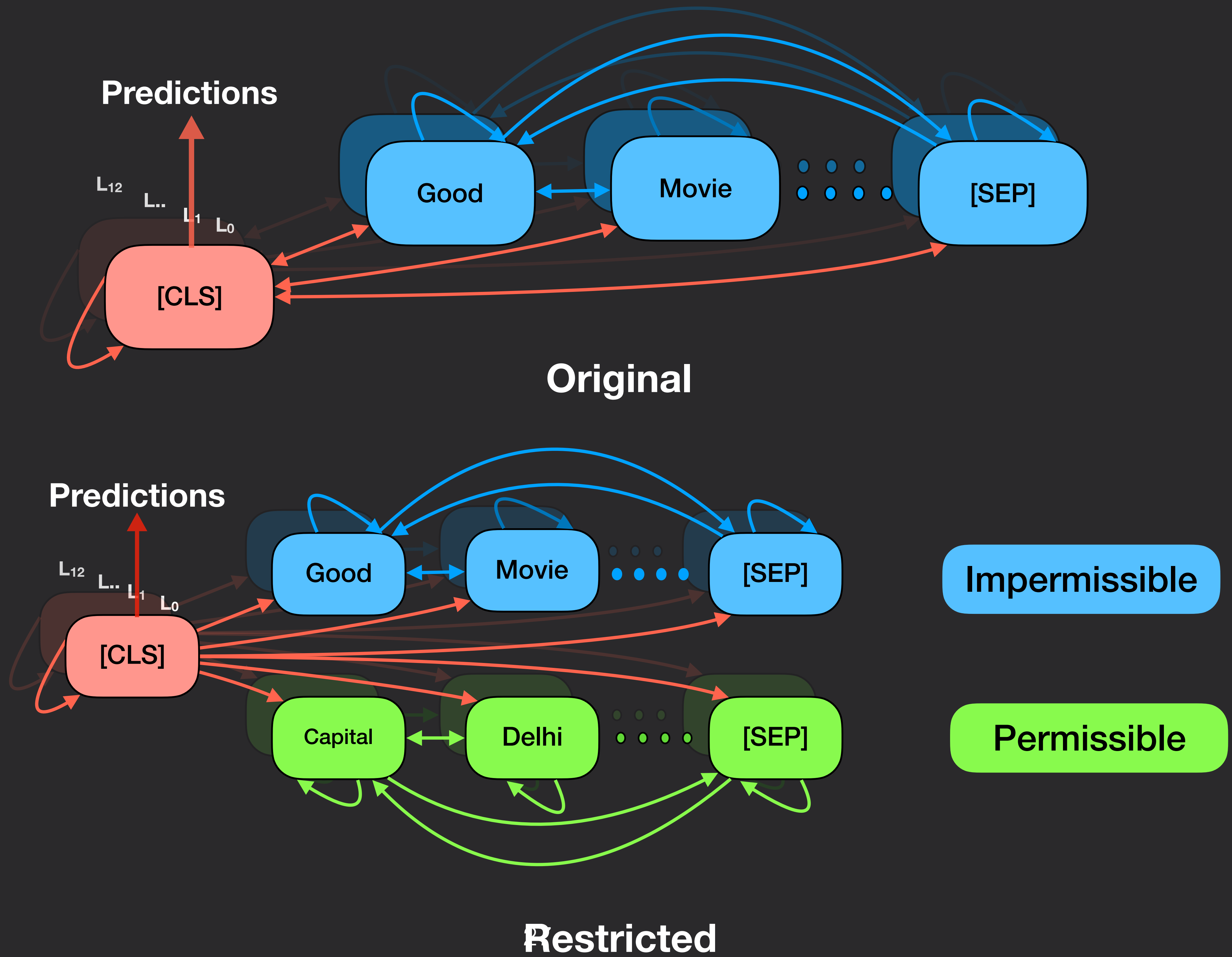


Devlin et. al

Restricted BERT

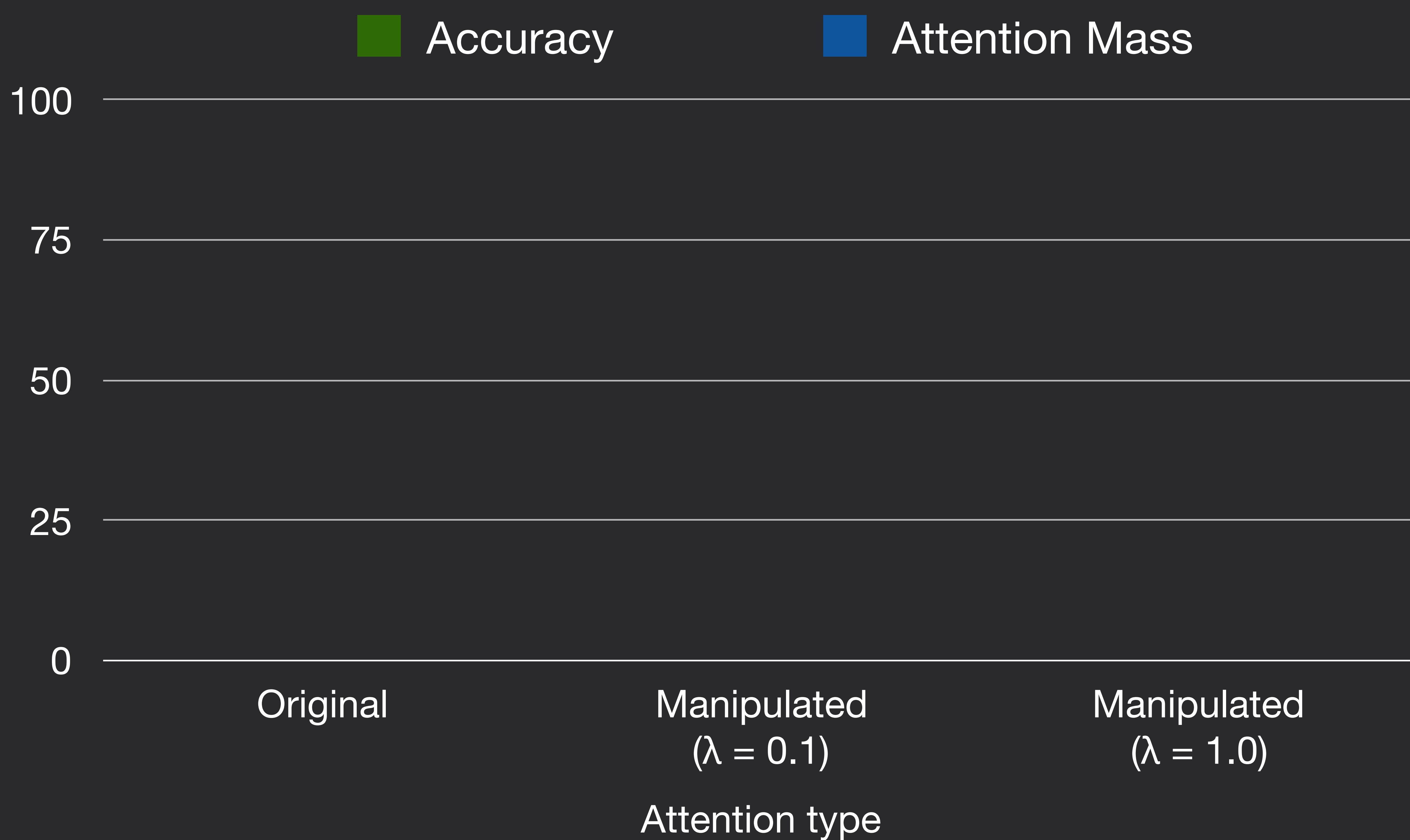


Restricted BERT

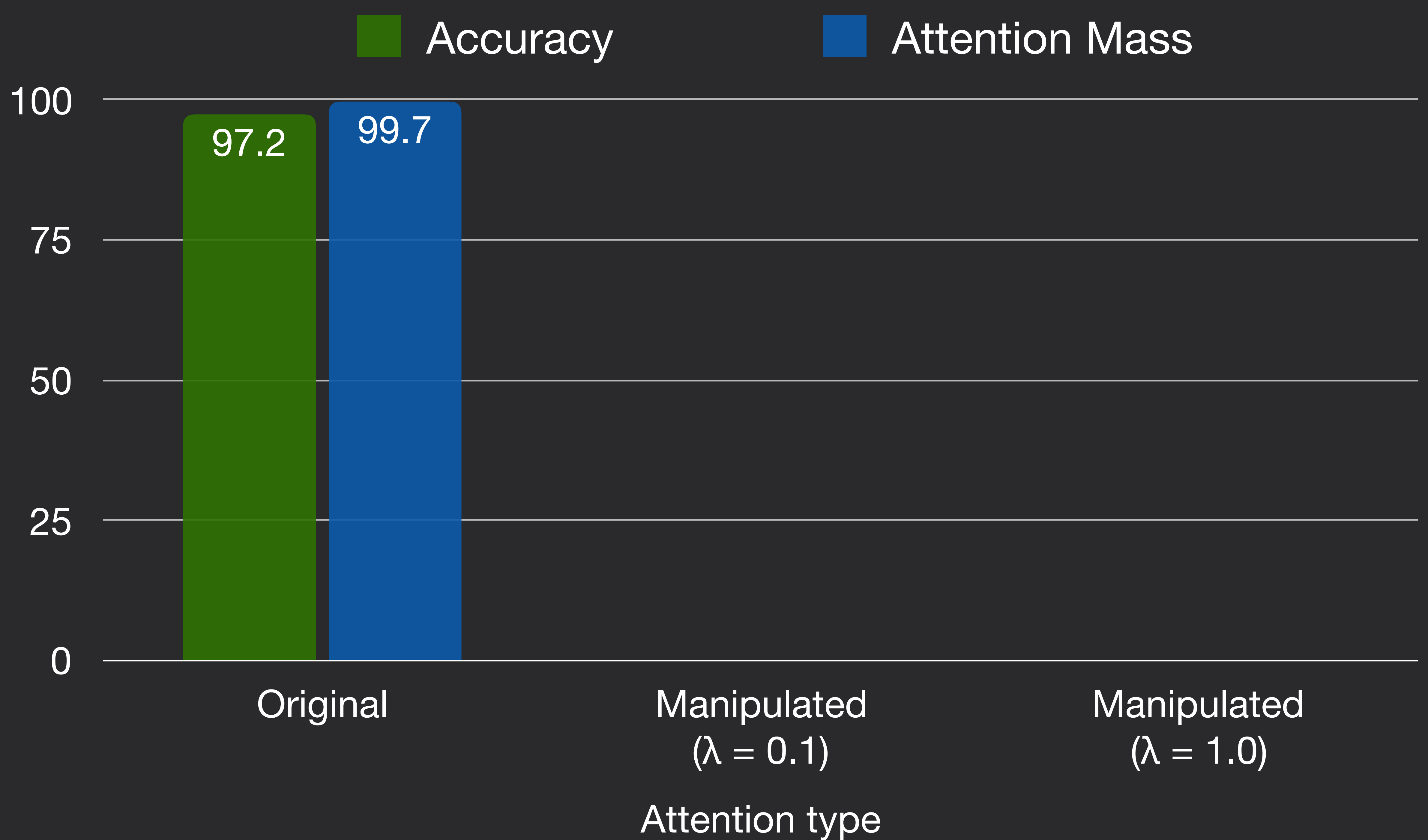


Occupation Prediction

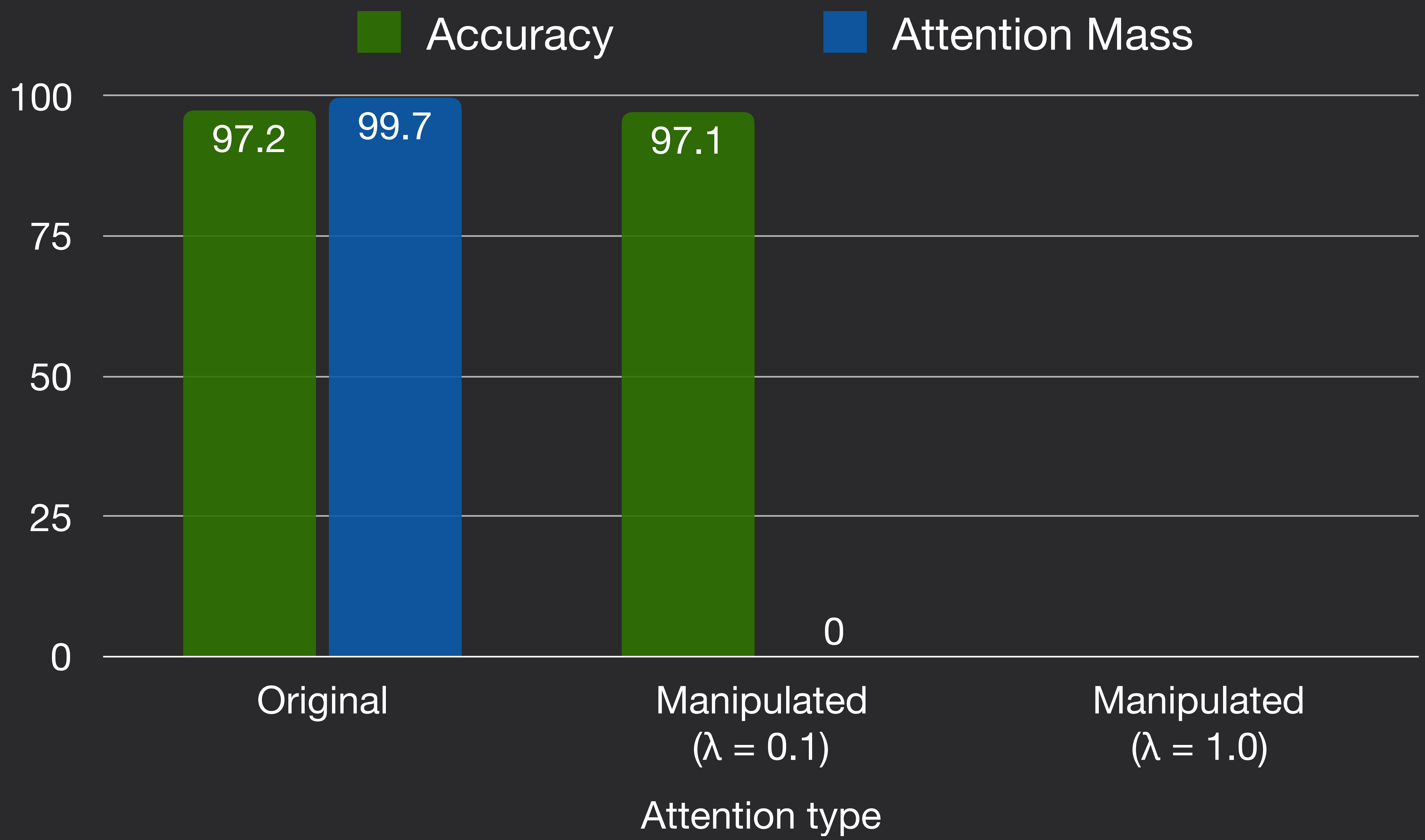
Occupation Prediction



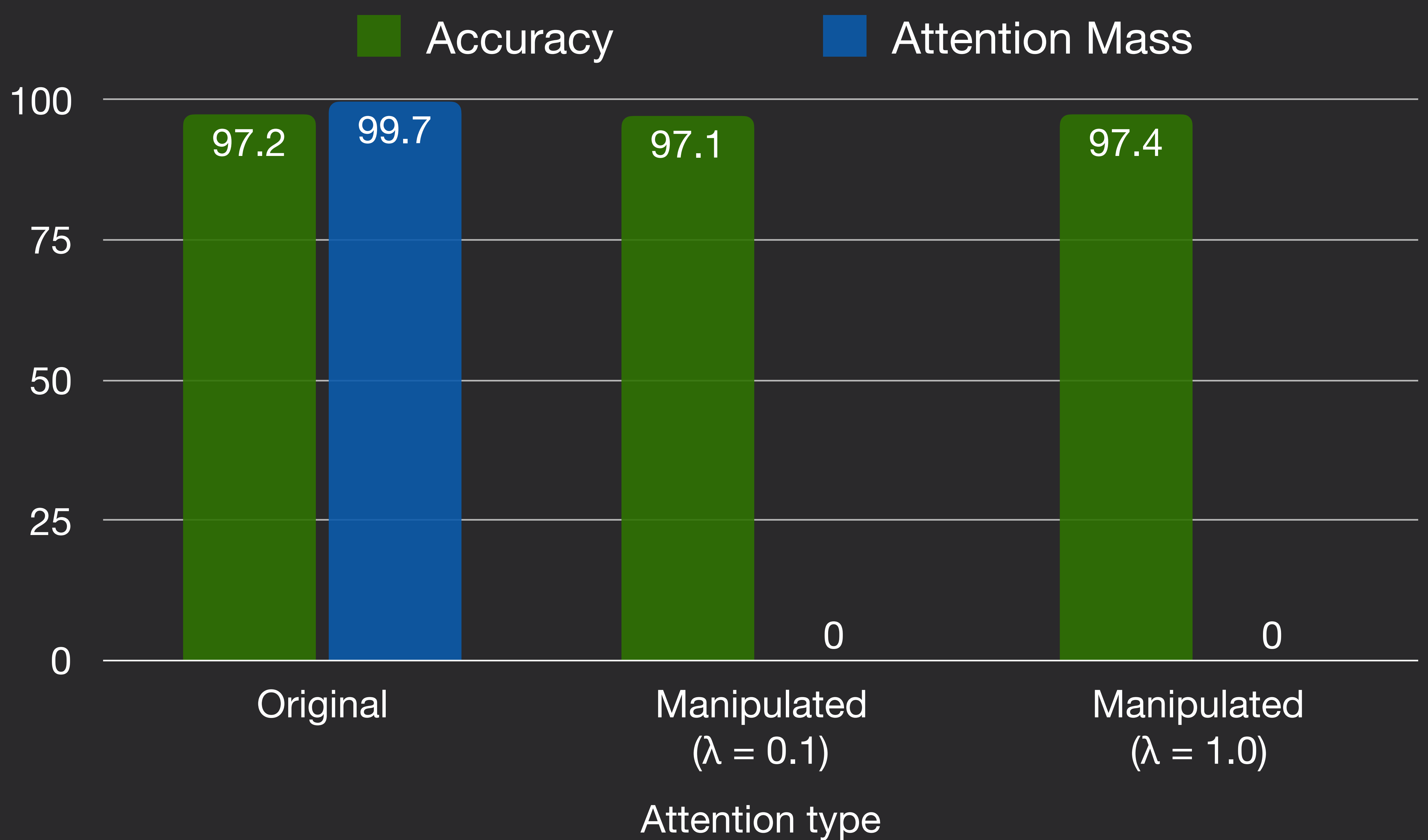
Occupation Prediction



Occupation Prediction



Occupation Prediction



Classification Tasks

Model	λ	$\mathcal{I}$	Occupation Pred.		Gender Identify		SST + Wiki		Ref. Letters	
			Acc.	A.M.	Acc.	A.M.	Acc.	A.M.	Acc.	A.M.
Embedding	0.0	✗	93.8	-	66.8	-	48.9	-	74.2	2.3
Embedding	0.0	✓	96.3	51.4	100	99.2	70.7	48.4	77.5	2.3
Embedding	0.1	✓	96.2	4.6	99.4	3.4	67.9	36.4	76.8	0.5
Embedding	1.0	✓	96.2	1.3	99.2	0.8	48.4	8.7	76.9	0.1
BiLSTM	0.0	✗	93.3	-	63.3	-	49.1	-	74.7	-
BiLSTM	0.0	✓	96.4	50.3	100	96.8	76.9	77.7	77.5	4.9
BiLSTM	0.1	✓	96.4	0.08	100	$< 10^{-6}$	60.6	0.04	76.9	3.9
BiLSTM	1.0	✓	96.7	$< 10^{-2}$	100	$< 10^{-6}$	61.0	0.07	74.2	$< 10^{-2}$
BERT	0.0	✗	95.0	-	72.8	-	50.4	-	68.2	
BERT (mean)	0.0	✓	97.2	13.9	100	80.8	90.8	59.0	74.7	2.6
BERT (mean)	0.1	✓	97.2	0.01	99.9	$< 10^{-3}$	90.9	$< 10^{-2}$	76.2	$< 10^{-1}$
BERT (mean)	1.0	✓	97.2	$< 10^{-3}$	99.9	$< 10^{-3}$	90.6	$< 10^{-3}$	75.2	$< 10^{-2}$
BERT	0.0	✗	95.0	-	72.8	-	50.4	-	68.2	
BERT (max)	0.0	✓	97.2	99.7	100	99.7	90.8	96.2	74.7	28.9
BERT (max)	0.1	✓	97.1	$< 10^{-3}$	99.9	$< 10^{-3}$	90.7	$< 10^{-2}$	76.7	0.6
BERT (max)	1.0	✓	97.4	$< 10^{-3}$	99.8	$< 10^{-4}$	90.2	$< 10^{-3}$	75.9	$< 10^{-2}$

Classification Tasks

Model	λ	$\mathcal{I}$	Occupation Pred.		Gender Identify		SST + Wiki		Ref. Letters	
			Acc.	A.M.	Acc.	A.M.	Acc.	A.M.	Acc.	A.M.
Embedding	0.0	✗	93.8	-	66.8	-	48.9	-	74.2	2.3
Embedding	0.0	✓	96.3	51.4	100	99.2	70.7	48.4	77.5	2.3
Embedding	0.1	✓	96.2	4.6	99.4	3.4	67.9	36.4	76.8	0.5
Embedding	1.0	✓	96.2	1.3	99.2	0.8	48.4	8.7	76.9	0.1
BiLSTM	0.0	✗	93.3	-	63.3	-	49.1	-	74.7	-
BiLSTM	0.0	✓	96.4	50.3	100	96.8	76.9	77.7	77.5	4.9
BiLSTM	0.1	✓	96.4	0.08	100	$< 10^{-6}$	60.6	0.04	76.9	3.9
BiLSTM	1.0	✓	96.7	$< 10^{-2}$	100	$< 10^{-6}$	61.0	0.07	74.2	$< 10^{-2}$
BERT	0.0	✗	95.0	-	72.8	-	50.4	-	68.2	
BERT (mean)	0.0	✓	97.2	13.9	100	80.8	90.8	59.0	74.7	2.6
BERT (mean)	0.1	✓	97.2	0.01	99.9	$< 10^{-3}$	90.9	$< 10^{-2}$	76.2	$< 10^{-1}$
BERT (mean)	1.0	✓	97.2	$< 10^{-3}$	99.9	$< 10^{-3}$	90.6	$< 10^{-3}$	75.2	$< 10^{-2}$
BERT	0.0	✗	95.0	-	72.8	-	50.4	-	68.2	
BERT (max)	0.0	✓	97.2	99.7	100	99.7	90.8	96.2	74.7	28.9
BERT (max)	0.1	✓	97.1	$< 10^{-3}$	99.9	$< 10^{-3}$	90.7	$< 10^{-2}$	76.7	0.6
BERT (max)	1.0	✓	97.4	$< 10^{-3}$	99.8	$< 10^{-4}$	90.2	$< 10^{-3}$	75.9	$< 10^{-2}$

Classification Tasks

Model	λ	$\mathcal{I}$	Occupation Pred.		Gender Identify		SST + Wiki		Ref. Letters	
			Acc.	A.M.	Acc.	A.M.	Acc.	A.M.	Acc.	A.M.
Embedding	0.0	✗	93.8	-	66.8	-	48.9	-	74.2	2.3
Embedding	0.0	✓	96.3	51.4	100	99.2	70.7	48.4	77.5	2.3
Embedding	0.1	✓	96.2	4.6	99.4	3.4	67.9	36.4	76.8	0.5
Embedding	1.0	✓	96.2	1.3	99.2	0.8	48.4	8.7	76.9	0.1
BiLSTM	0.0	✗	93.3	-	63.3	-	49.1	-	74.7	-
BiLSTM	0.0	✓	96.4	50.3	100	96.8	76.9	77.7	77.5	4.9
BiLSTM	0.1	✓	96.4	0.08	100	$< 10^{-6}$	60.6	0.04	76.9	3.9
BiLSTM	1.0	✓	96.7	$< 10^{-2}$	100	$< 10^{-6}$	61.0	0.07	74.2	$< 10^{-2}$
BERT	0.0	✗	95.0	-	72.8	-	50.4	-	68.2	
BERT (mean)	0.0	✓	97.2	13.9	100	80.8	90.8	59.0	74.7	2.6
BERT (mean)	0.1	✓	97.2	0.01	99.9	$< 10^{-3}$	90.9	$< 10^{-2}$	76.2	$< 10^{-1}$
BERT (mean)	1.0	✓	97.2	$< 10^{-3}$	99.9	$< 10^{-3}$	90.6	$< 10^{-3}$	75.2	$< 10^{-2}$
BERT	0.0	✗	95.0	-	72.8	-	50.4	-	68.2	
BERT (max)	0.0	✓	97.2	99.7	100	99.7	90.8	96.2	74.7	28.9
BERT (max)	0.1	✓	97.1	$< 10^{-3}$	99.9	$< 10^{-3}$	90.7	$< 10^{-2}$	76.7	0.6
BERT (max)	1.0	✓	97.4	$< 10^{-3}$	99.8	$< 10^{-4}$	90.2	$< 10^{-3}$	75.9	$< 10^{-2}$

Classification Tasks

Model	λ	$\mathcal{I}$	Occupation Pred.		Gender Identify		SST + Wiki		Ref. Letters	
			Acc.	A.M.	Acc.	A.M.	Acc.	A.M.	Acc.	A.M.
Embedding	0.0	✗	93.8	-	66.8	-	48.9	-	74.2	2.3
Embedding	0.0	✓	96.3	51.4	100	99.2	70.7	48.4	77.5	2.3
Embedding	0.1	✓	96.2	4.6	99.4	3.4	67.9	36.4	76.8	0.5
Embedding	1.0	✓	96.2	1.3	99.2	0.8	48.4	8.7	76.9	0.1
BiLSTM	0.0	✗	93.3	-	63.3	-	49.1	-	74.7	-
BiLSTM	0.0	✓	96.4	50.3	100	96.8	76.9	77.7	77.5	4.9
BiLSTM	0.1	✓	96.4	0.08	100	$< 10^{-6}$	60.6	0.04	76.9	3.9
BiLSTM	1.0	✓	96.7	$< 10^{-2}$	100	$< 10^{-6}$	61.0	0.07	74.2	$< 10^{-2}$
BERT	0.0	✗	95.0	-	72.8	-	50.4	-	68.2	
BERT (mean)	0.0	✓	97.2	13.9	100	80.8	90.8	59.0	74.7	2.6
BERT (mean)	0.1	✓	97.2	0.01	99.9	$< 10^{-3}$	90.9	$< 10^{-2}$	76.2	$< 10^{-1}$
BERT (mean)	1.0	✓	97.2	$< 10^{-3}$	99.9	$< 10^{-3}$	90.6	$< 10^{-3}$	75.2	$< 10^{-2}$
BERT	0.0	✗	95.0	-	72.8	-	50.4	-	68.2	
BERT (max)	0.0	✓	97.2	99.7	100	99.7	90.8	96.2	74.7	28.9
BERT (max)	0.1	✓	97.1	$< 10^{-3}$	99.9	$< 10^{-3}$	90.7	$< 10^{-2}$	76.7	0.6
BERT (max)	1.0	✓	97.4	$< 10^{-3}$	99.8	$< 10^{-4}$	90.2	$< 10^{-3}$	75.9	$< 10^{-2}$

Classification Tasks

Model	λ	$\mathcal{I}$	Occupation Pred.		Gender Identify		SST + Wiki		Ref. Letters	
			Acc.	A.M.	Acc.	A.M.	Acc.	A.M.	Acc.	A.M.
Embedding	0.0	✗	93.8	-	66.8	-	48.9	-	74.2	2.3
Embedding	0.0	✓	96.3	51.4	100	99.2	70.7	48.4	77.5	2.3
Embedding	0.1	✓	96.2	4.6	99.4	3.4	67.9	36.4	76.8	0.5
Embedding	1.0	✓	96.2	1.3	99.2	0.8	48.4	8.7	76.9	0.1
BiLSTM	0.0	✗	93.3	-	63.3	-	49.1	-	74.7	-
BiLSTM	0.0	✓	96.4	50.3	100	96.8	76.9	77.7	77.5	4.9
BiLSTM	0.1	✓	96.4	0.08	100	$< 10^{-6}$	60.6	0.04	76.9	3.9
BiLSTM	1.0	✓	96.7	$< 10^{-2}$	100	$< 10^{-6}$	61.0	0.07	74.2	$< 10^{-2}$
BERT	0.0	✗	95.0	-	72.8	-	50.4	-	68.2	
BERT (mean)	0.0	✓	97.2	13.9	100	80.8	90.8	59.0	74.7	2.6
BERT (mean)	0.1	✓	97.2	0.01	99.9	$< 10^{-3}$	90.9	$< 10^{-2}$	76.2	$< 10^{-1}$
BERT (mean)	1.0	✓	97.2	$< 10^{-3}$	99.9	$< 10^{-3}$	90.6	$< 10^{-3}$	75.2	$< 10^{-2}$
BERT	0.0	✗	95.0	-	72.8	-	50.4	-	68.2	
BERT (max)	0.0	✓	97.2	99.7	100	99.7	90.8	96.2	74.7	28.9
BERT (max)	0.1	✓	97.1	$< 10^{-3}$	99.9	$< 10^{-3}$	90.7	$< 10^{-2}$	76.7	0.6
BERT (max)	1.0	✓	97.4	$< 10^{-3}$	99.8	$< 10^{-4}$	90.2	$< 10^{-3}$	75.9	$< 10^{-2}$

Classification Tasks

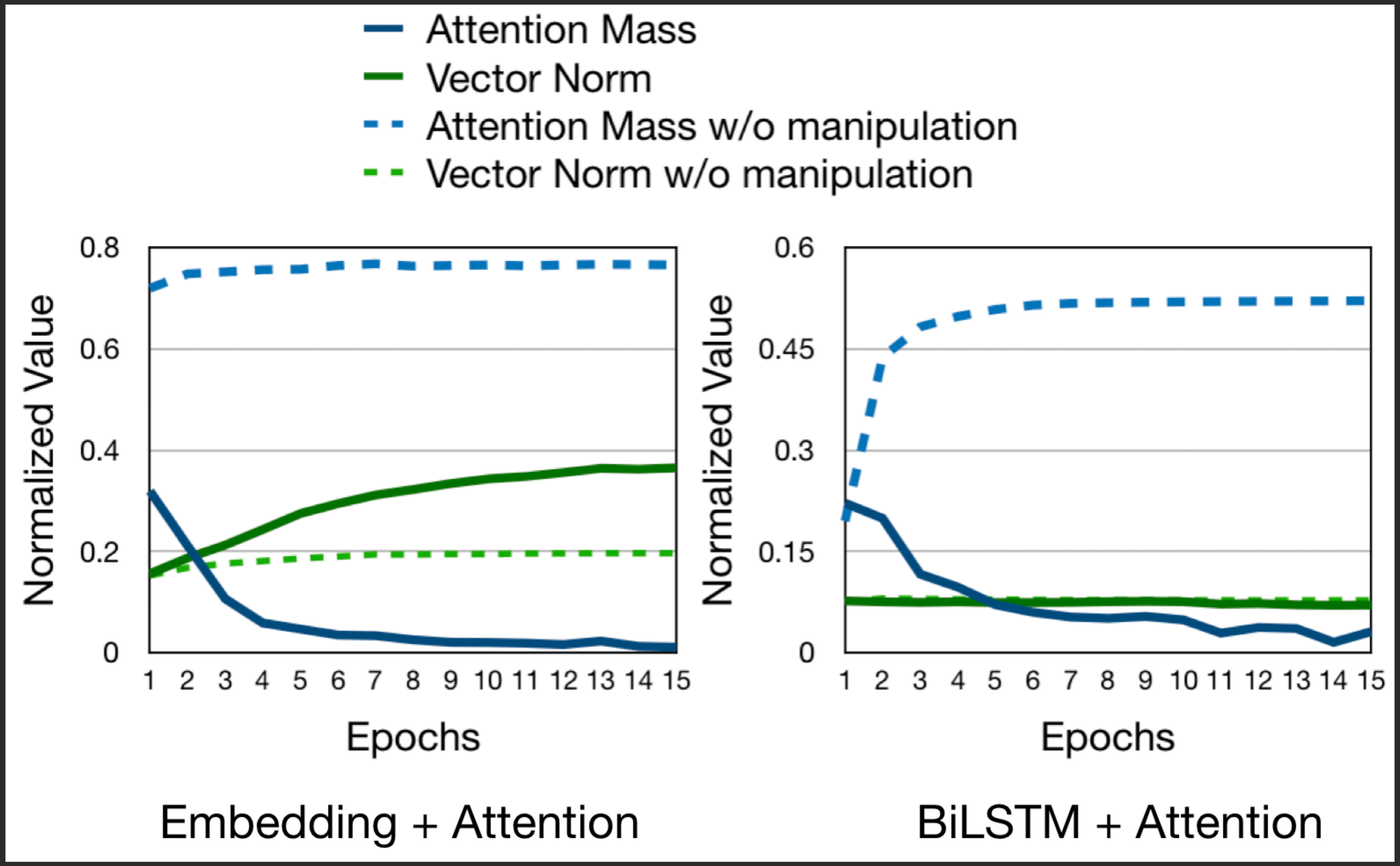
Model	λ	$\mathcal{I}$	Occupation Pred.		Gender Identify		SST + Wiki		Ref. Letters	
			Acc.	A.M.	Acc.	A.M.	Acc.	A.M.	Acc.	A.M.
Embedding	0.0	✗	93.8	-	66.8	-	48.9	-	74.2	2.3
Embedding	0.0	✓	96.3	51.4	100	99.2	70.7	48.4	77.5	2.3
Embedding	0.1	✓	96.2	4.6	99.4	3.4	67.9	36.4	76.8	0.5
Embedding	1.0	✓	96.2	1.3	99.2	0.8	48.4	8.7	76.9	0.1
BiLSTM	0.0	✗	93.3	-	63.3	-	49.1	-	74.7	-
BiLSTM	0.0	✓	96.4	50.3	100	96.8	76.9	77.7	77.5	4.9
BiLSTM	0.1	✓	96.4	0.08	100	$< 10^{-6}$	60.6	0.04	76.9	3.9
BiLSTM	1.0	✓	96.7	$< 10^{-2}$	100	$< 10^{-6}$	61.0	0.07	74.2	$< 10^{-2}$
BERT	0.0	✗	95.0	-	72.8	-	50.4	-	68.2	
BERT (mean)	0.0	✓	97.2	13.9	100	80.8	90.8	59.0	74.7	2.6
BERT (mean)	0.1	✓	97.2	0.01	99.9	$< 10^{-3}$	90.9	$< 10^{-2}$	76.2	$< 10^{-1}$
BERT (mean)	1.0	✓	97.2	$< 10^{-3}$	99.9	$< 10^{-3}$	90.6	$< 10^{-3}$	75.2	$< 10^{-2}$
BERT	0.0	✗	95.0	-	72.8	-	50.4	-	68.2	
BERT (max)	0.0	✓	97.2	99.7	100	99.7	90.8	96.2	74.7	28.9
BERT (max)	0.1	✓	97.1	$< 10^{-3}$	99.9	$< 10^{-3}$	90.7	$< 10^{-2}$	76.7	0.6
BERT (max)	1.0	✓	97.4	$< 10^{-3}$	99.8	$< 10^{-4}$	90.2	$< 10^{-3}$	75.9	$< 10^{-2}$

Classification Tasks

Model	λ	$\mathcal{I}$	Occupation Pred.		Gender Identify		SST + Wiki		Ref. Letters	
			Acc.	A.M.	Acc.	A.M.	Acc.	A.M.	Acc.	A.M.
Embedding	0.0	✗	93.8	-	66.8	-	48.9	-	74.2	2.3
Embedding	0.0	✓	96.3	51.4	100	99.2	70.7	48.4	77.5	2.3
Embedding	0.1	✓	96.2	4.6	99.4	3.4	67.9	36.4	76.8	0.5
Embedding	1.0	✓	96.2	1.3	99.2	0.8	48.4	8.7	76.9	0.1
BiLSTM	0.0	✗	93.3	-	63.3	-	49.1	-	74.7	-
BiLSTM	0.0	✓	96.4	50.3	100	96.8	76.9	77.7	77.5	4.9
BiLSTM	0.1	✓	96.4	0.08	100	$< 10^{-6}$	60.6	0.04	76.9	3.9
BiLSTM	1.0	✓	96.7	$< 10^{-2}$	100	$< 10^{-6}$	61.0	0.07	74.2	$< 10^{-2}$
BERT	0.0	✗	95.0	-	72.8	-	50.4	-	68.2	
BERT (mean)	0.0	✓	97.2	13.9	100	80.8	90.8	59.0	74.7	2.6
BERT (mean)	0.1	✓	97.2	0.01	99.9	$< 10^{-3}$	90.9	$< 10^{-2}$	76.2	$< 10^{-1}$
BERT (mean)	1.0	✓	97.2	$< 10^{-3}$	99.9	$< 10^{-3}$	90.6	$< 10^{-3}$	75.2	$< 10^{-2}$
BERT	0.0	✗	95.0	-	72.8	-	50.4	-	68.2	
BERT (max)	0.0	✓	97.2	99.7	100	99.7	90.8	96.2	74.7	28.9
BERT (max)	0.1	✓	97.1	$< 10^{-3}$	99.9	$< 10^{-3}$	90.7	$< 10^{-2}$	76.7	0.6
BERT (max)	1.0	✓	97.4	$< 10^{-3}$	99.8	$< 10^{-4}$	90.2	$< 10^{-3}$	75.9	$< 10^{-2}$

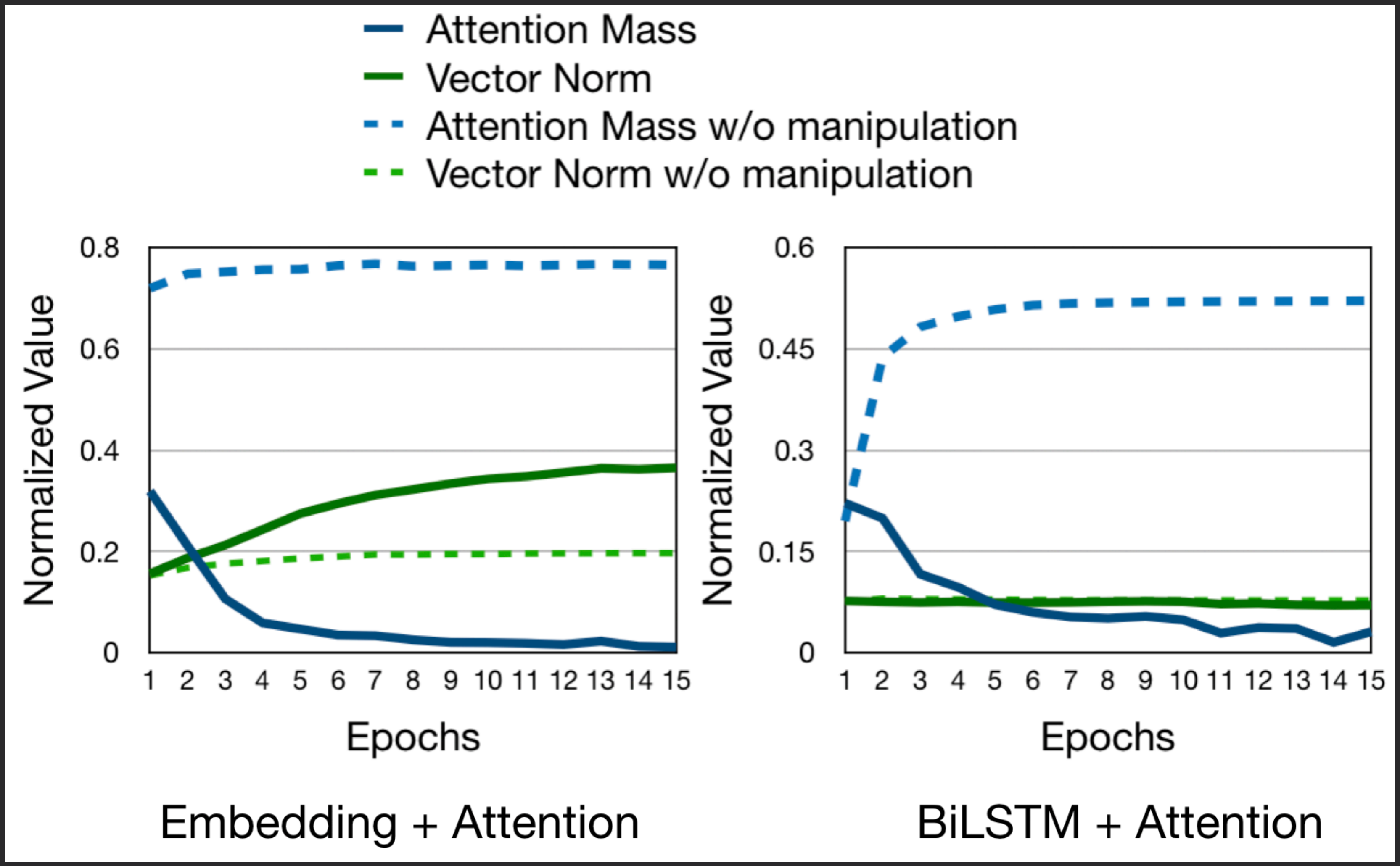
Alternate mechanisms

Gender-Identification



Alternate mechanisms

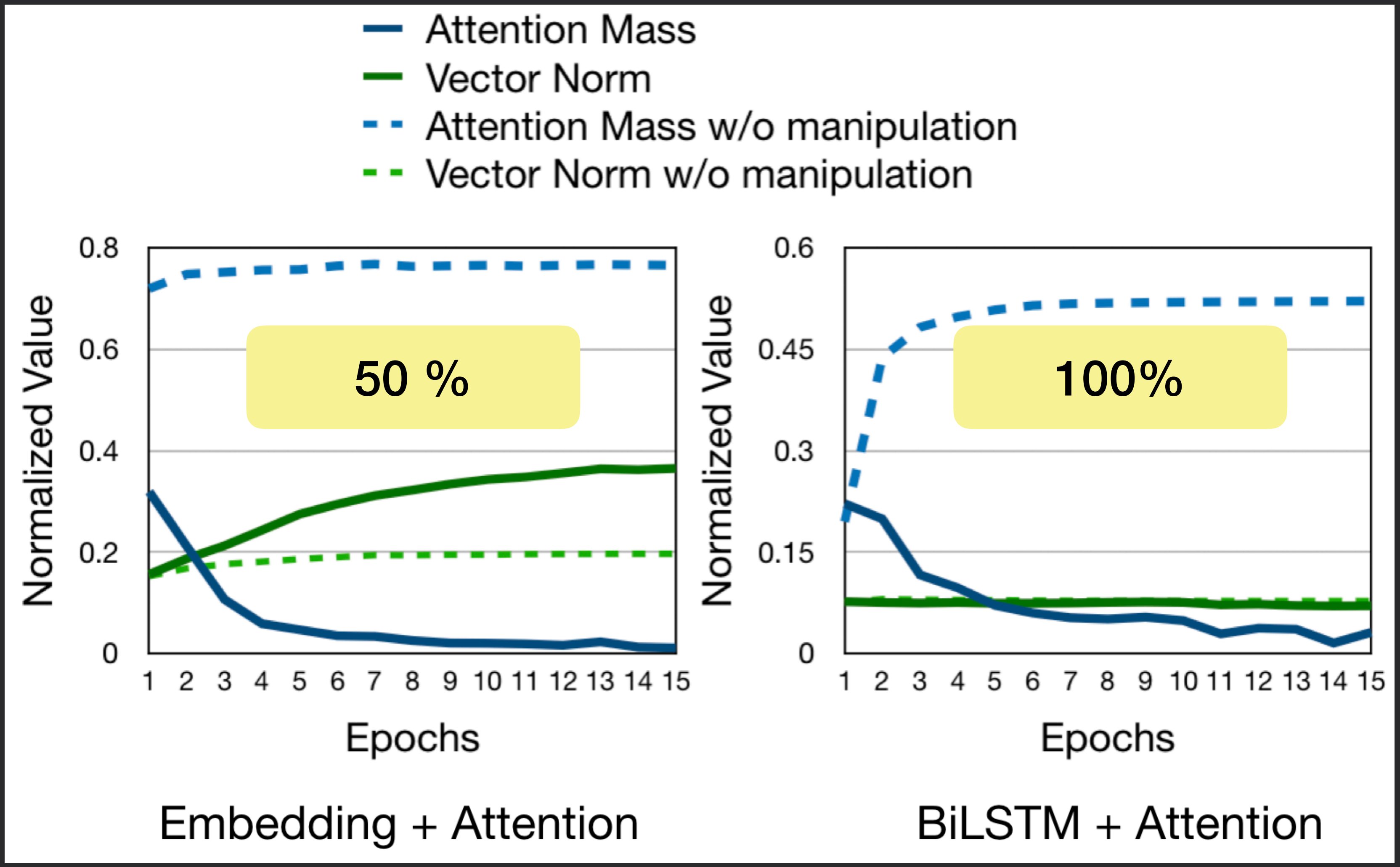
Gender-Identification



At inference time, what if we hard set the corresponding attention mass to ZERO?

Alternate mechanisms

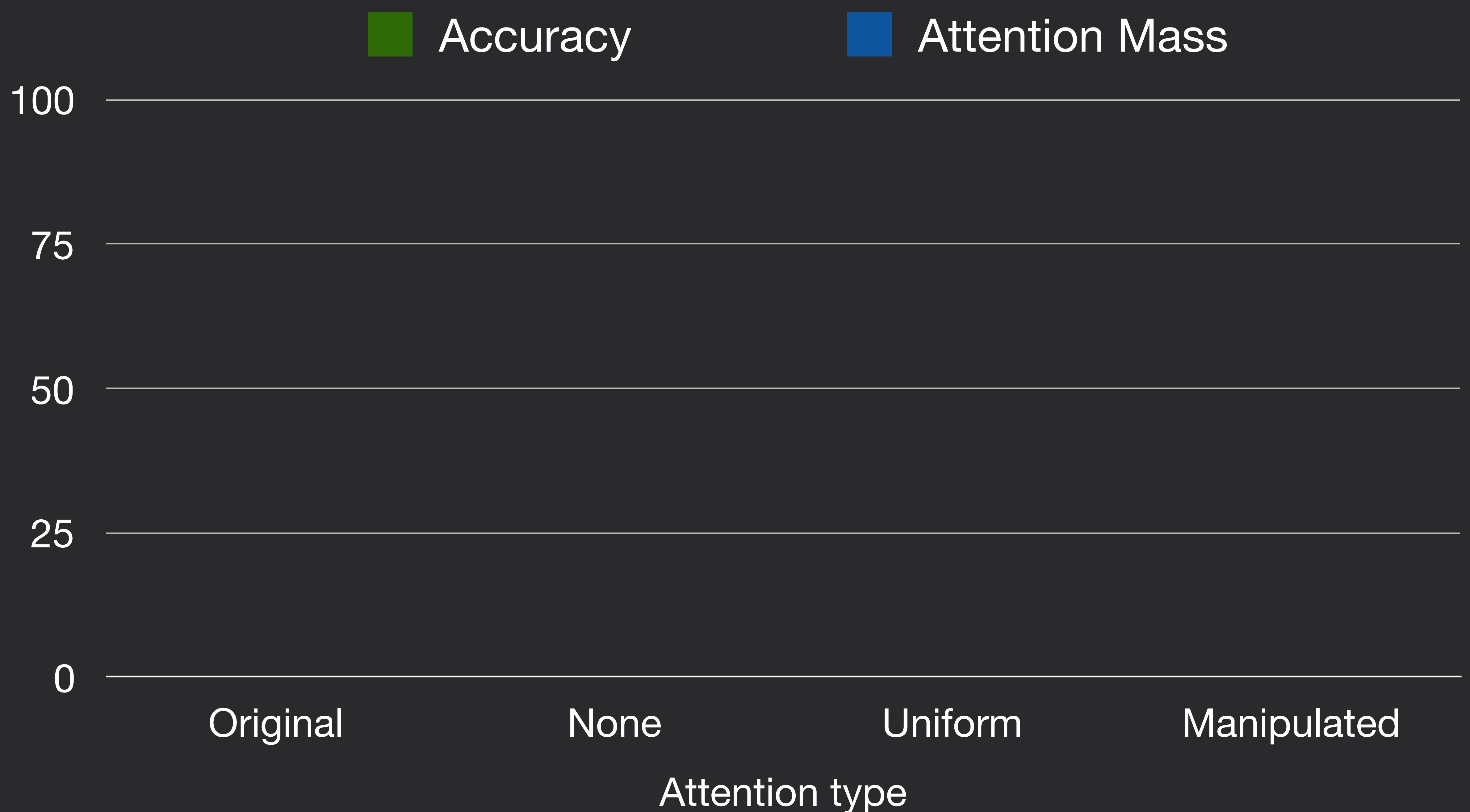
Gender-Identification



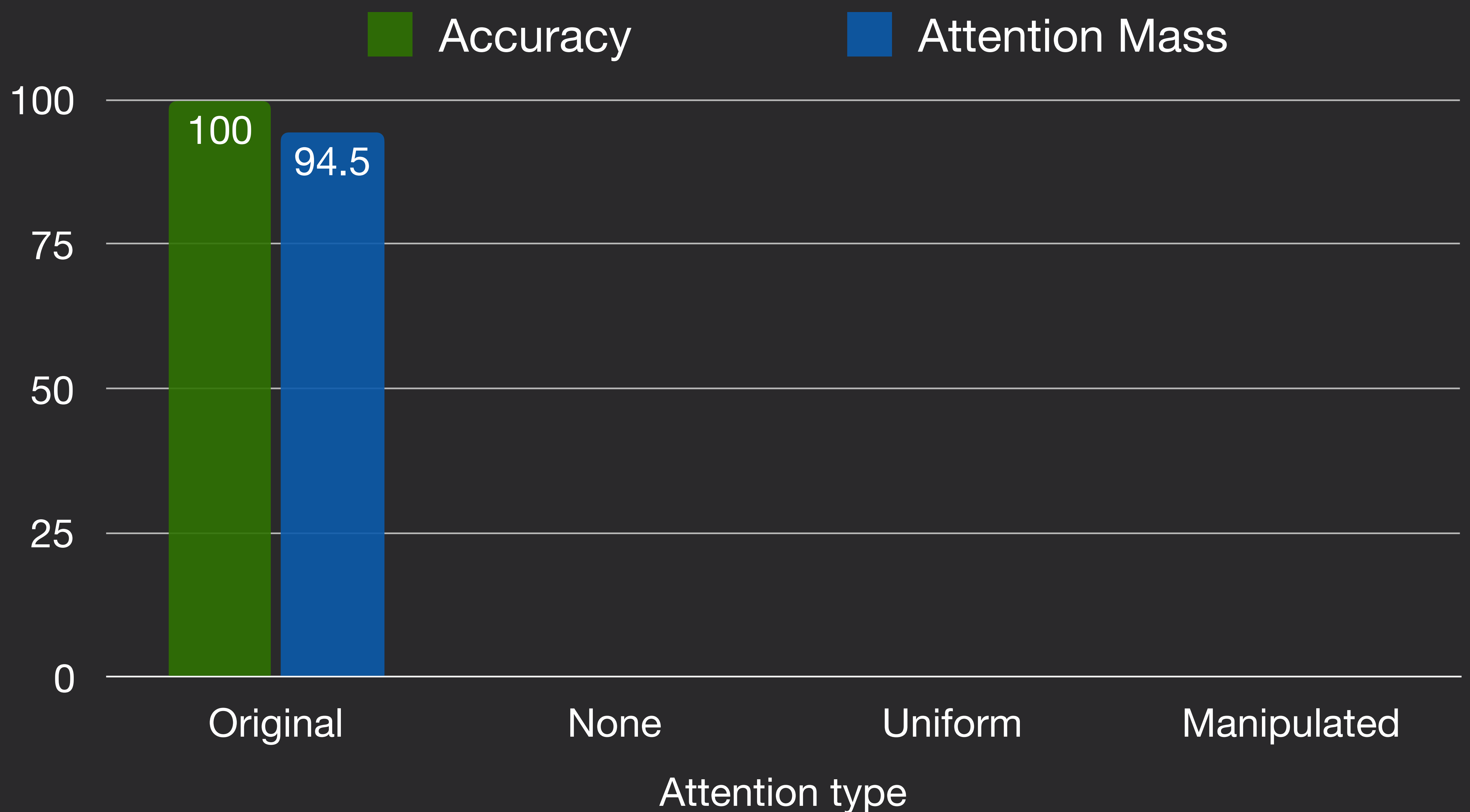
At inference time, what if we hard set the corresponding attention mass to ZERO?

Bigram Flip

Bigram Flip

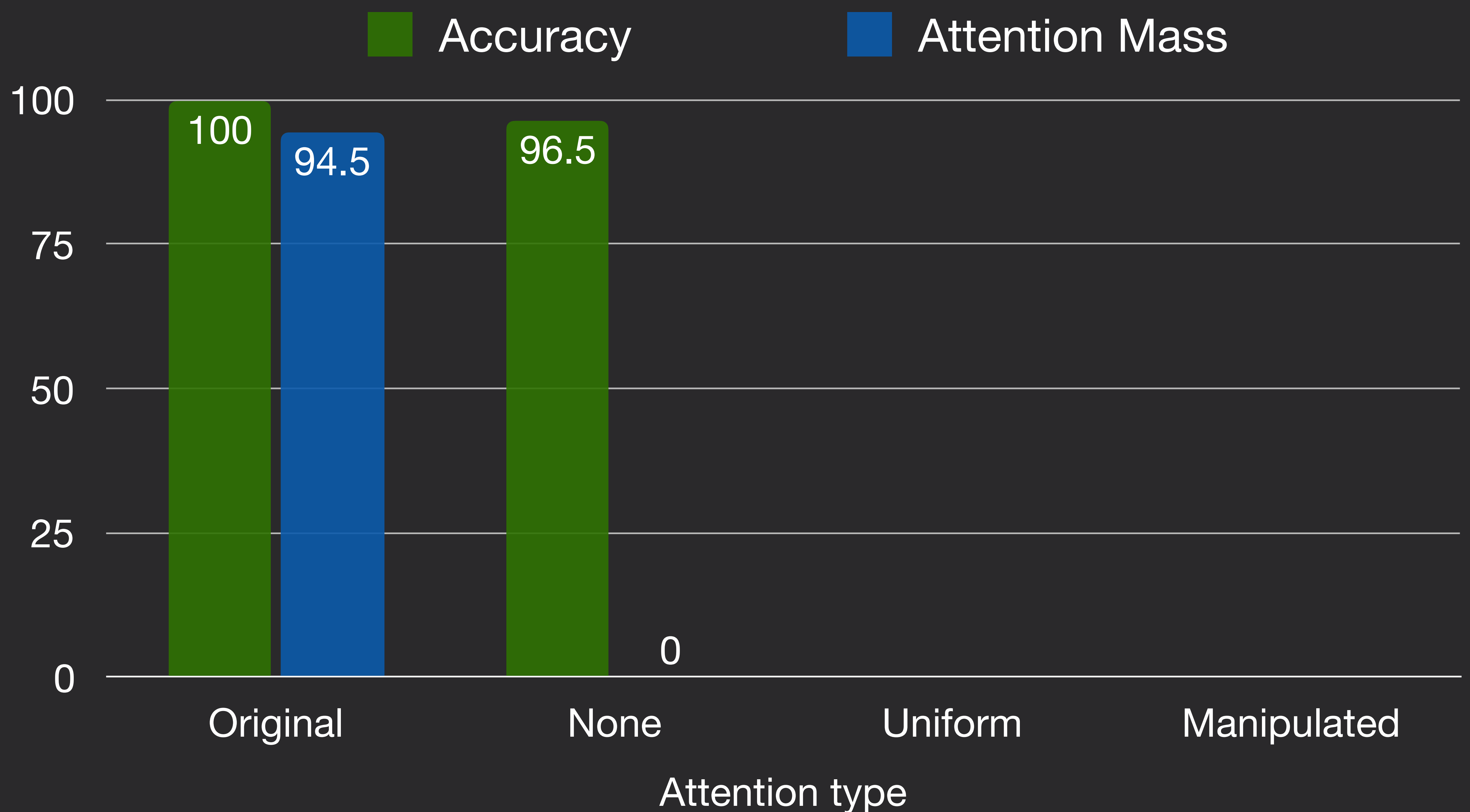

Accuracy
Attention Mass
100
75
50
25
0
Original
None
Uniform
Manipulated
Attention type

Bigram Flip



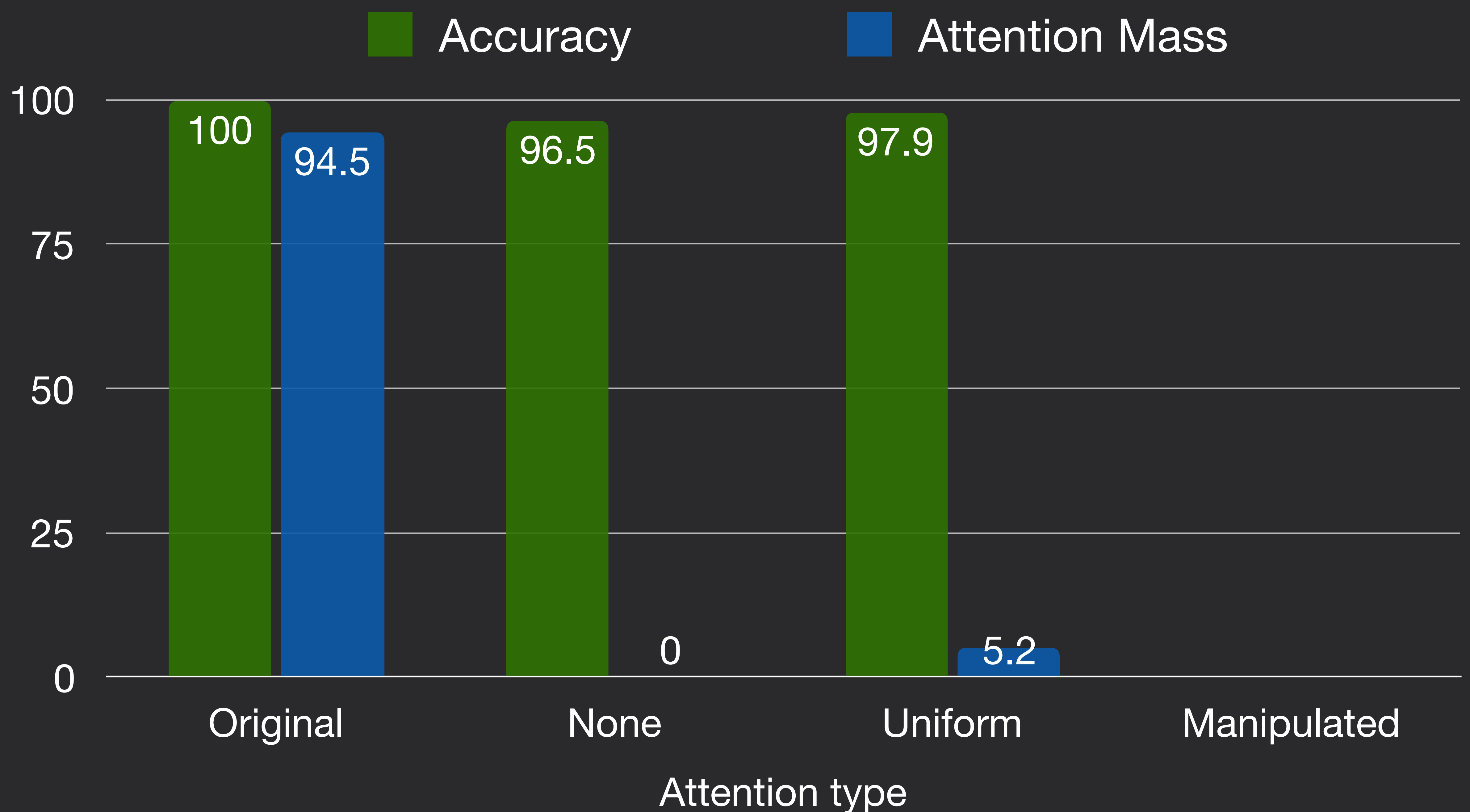
Accuracy
Attention Mass
100
75
50
25
0
100
94.5
Original
None
Uniform
Manipulated
Attention type

Bigram Flip

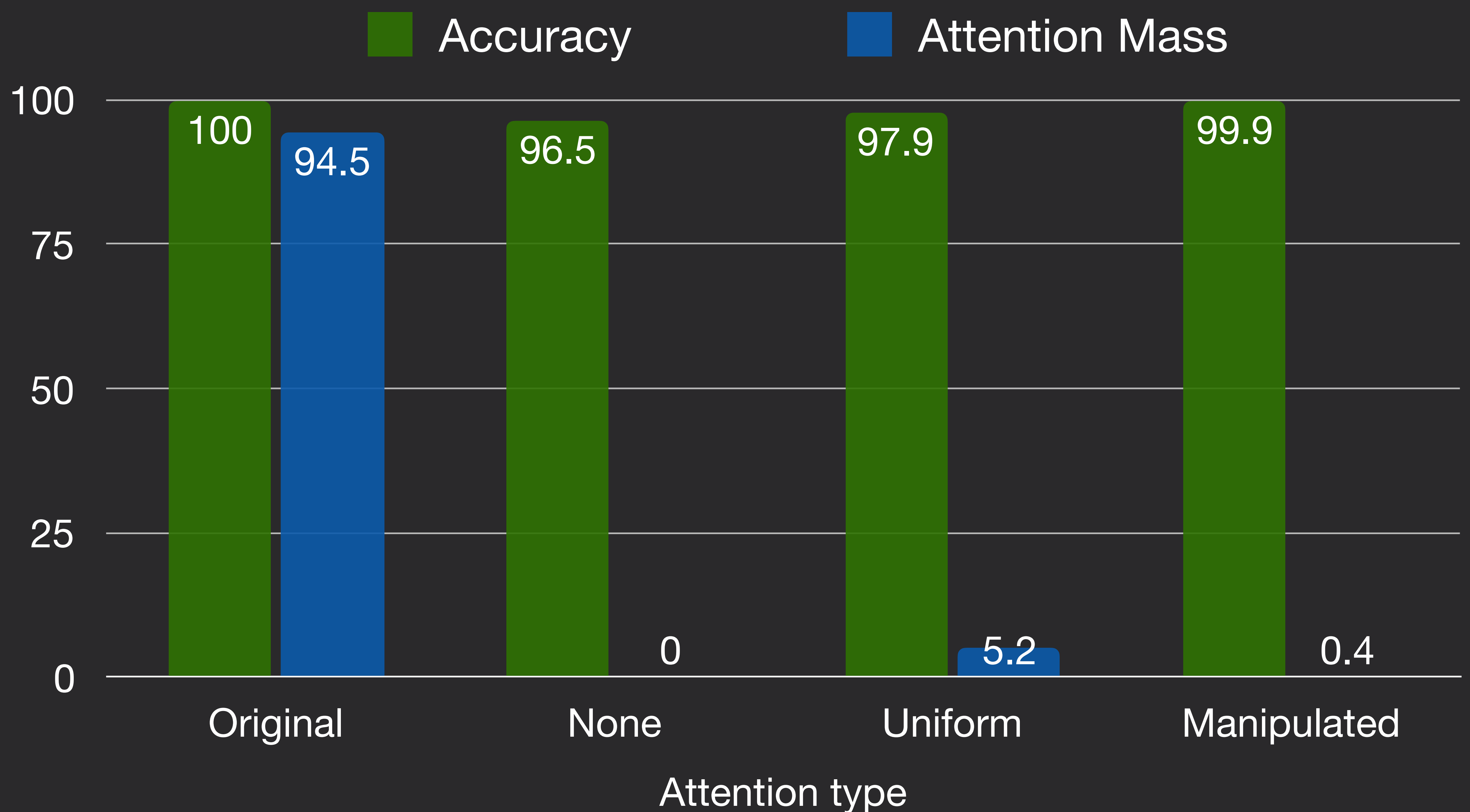


Accuracy
Attention Mass
100
75
50
25
0
100
94.5
96.5
0
Original
None
Uniform
Manipulated
Attention type

Bigram Flip

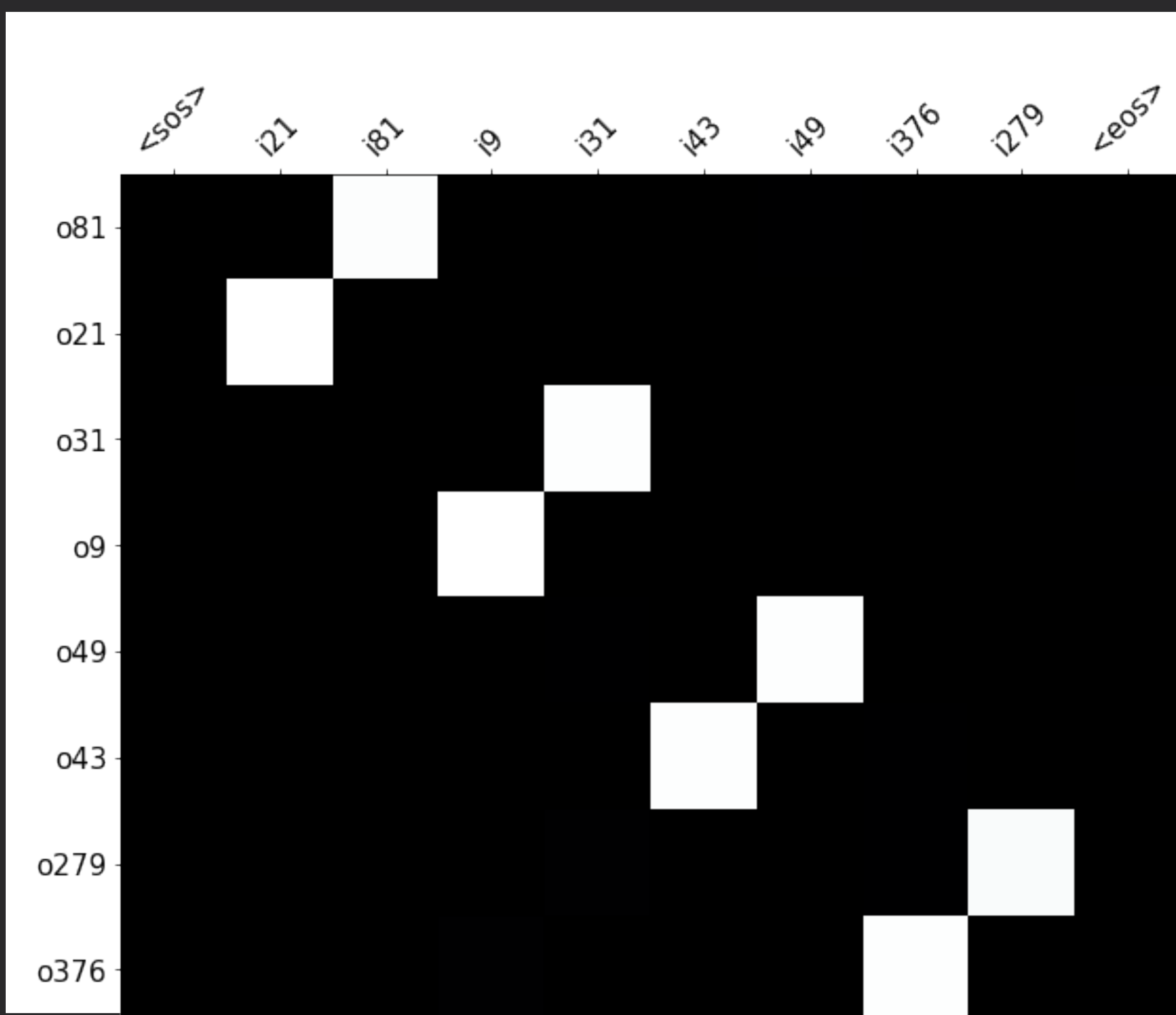

Accuracy
Attention Mass
100
75
50
25
0
100
94.5
96.5
0
97.9
5.2
Original
None
Uniform
Manipulated
Attention type

Bigram Flip



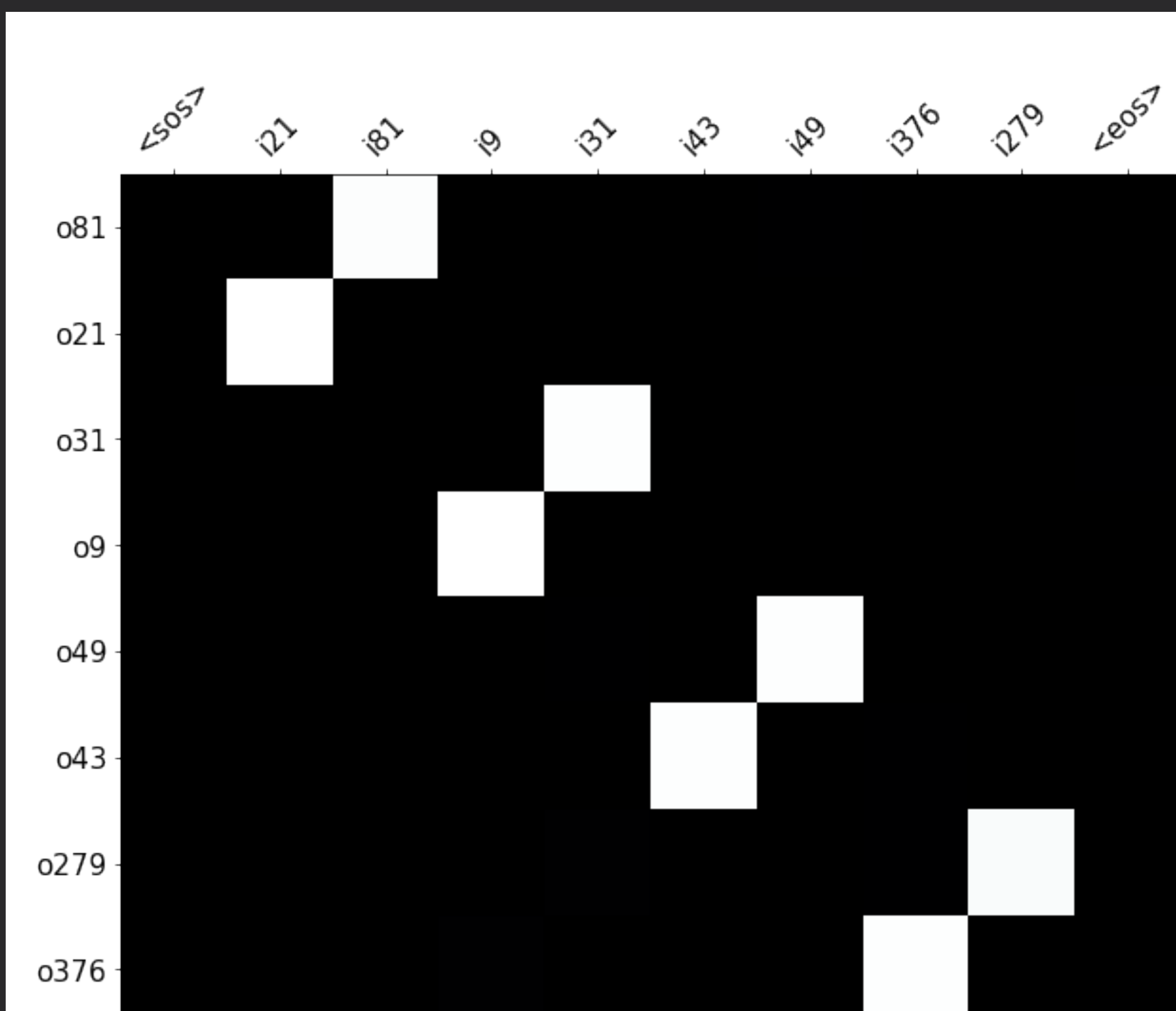
Accuracy
Attention Mass
100
75
50
25
0
100
94.5
96.5
0
97.9
5.2
99.9
0.4
Original
None
Uniform
Manipulated
Attention type

Bigram Flip

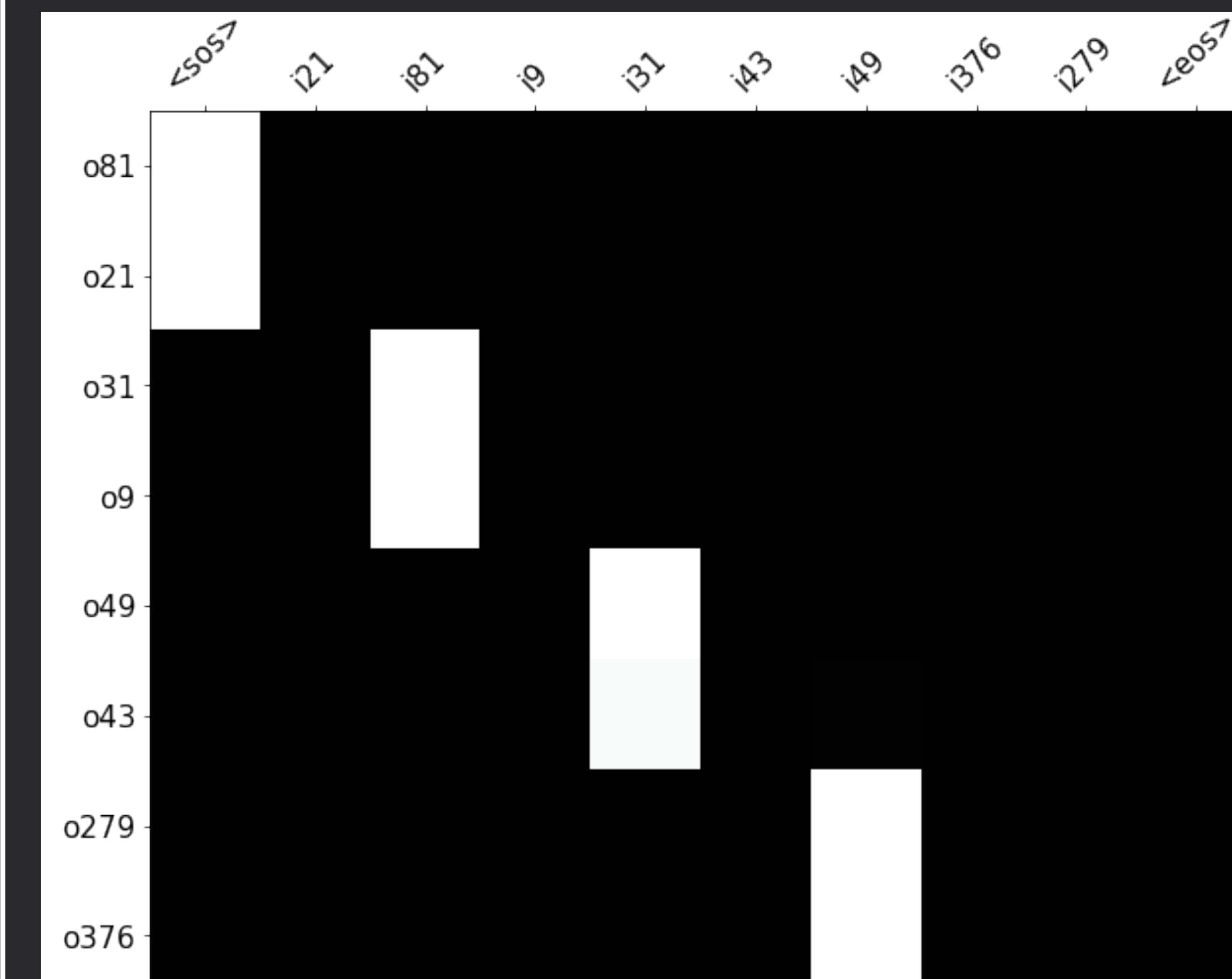


Original

Bigram Flip



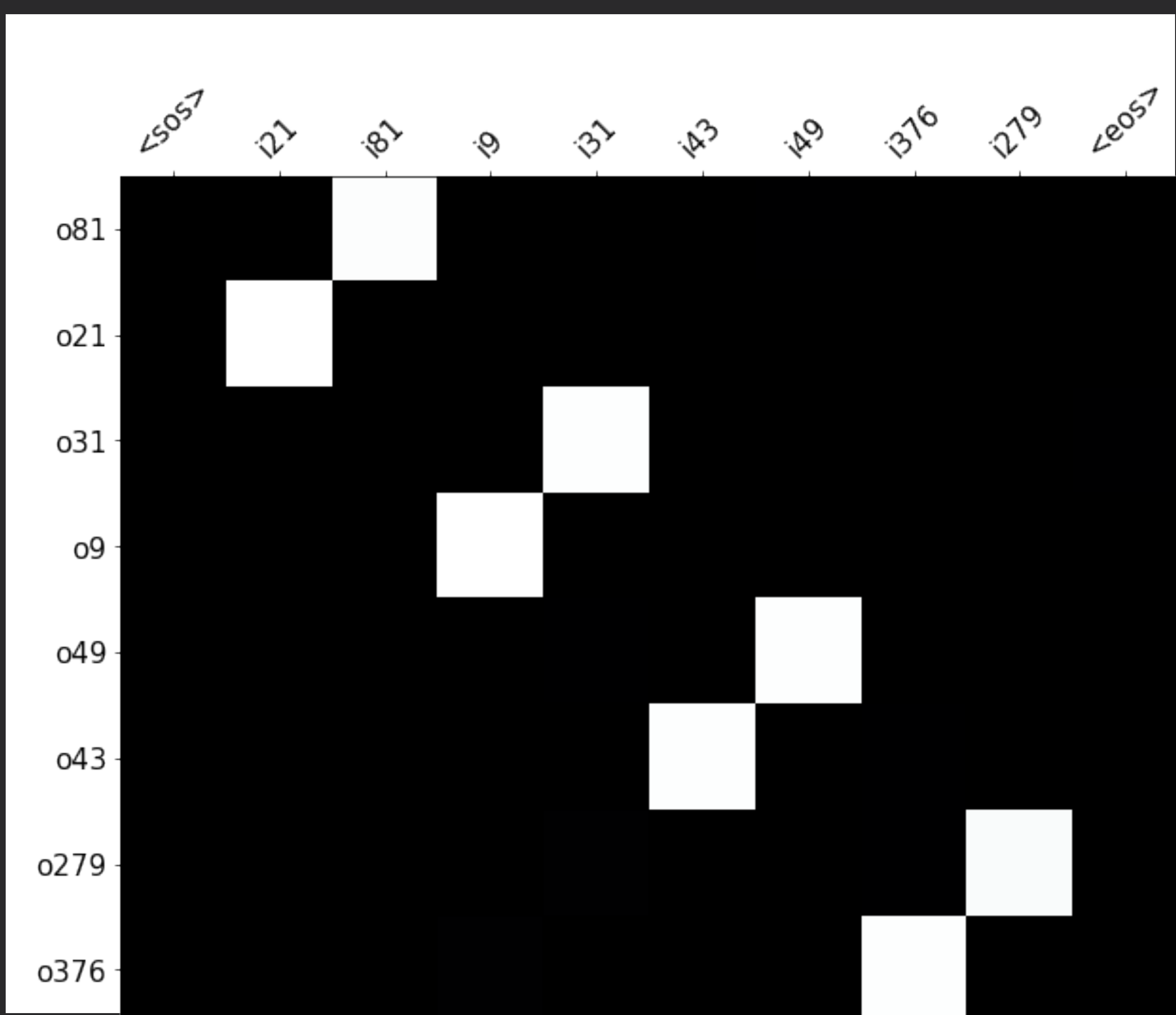
Original



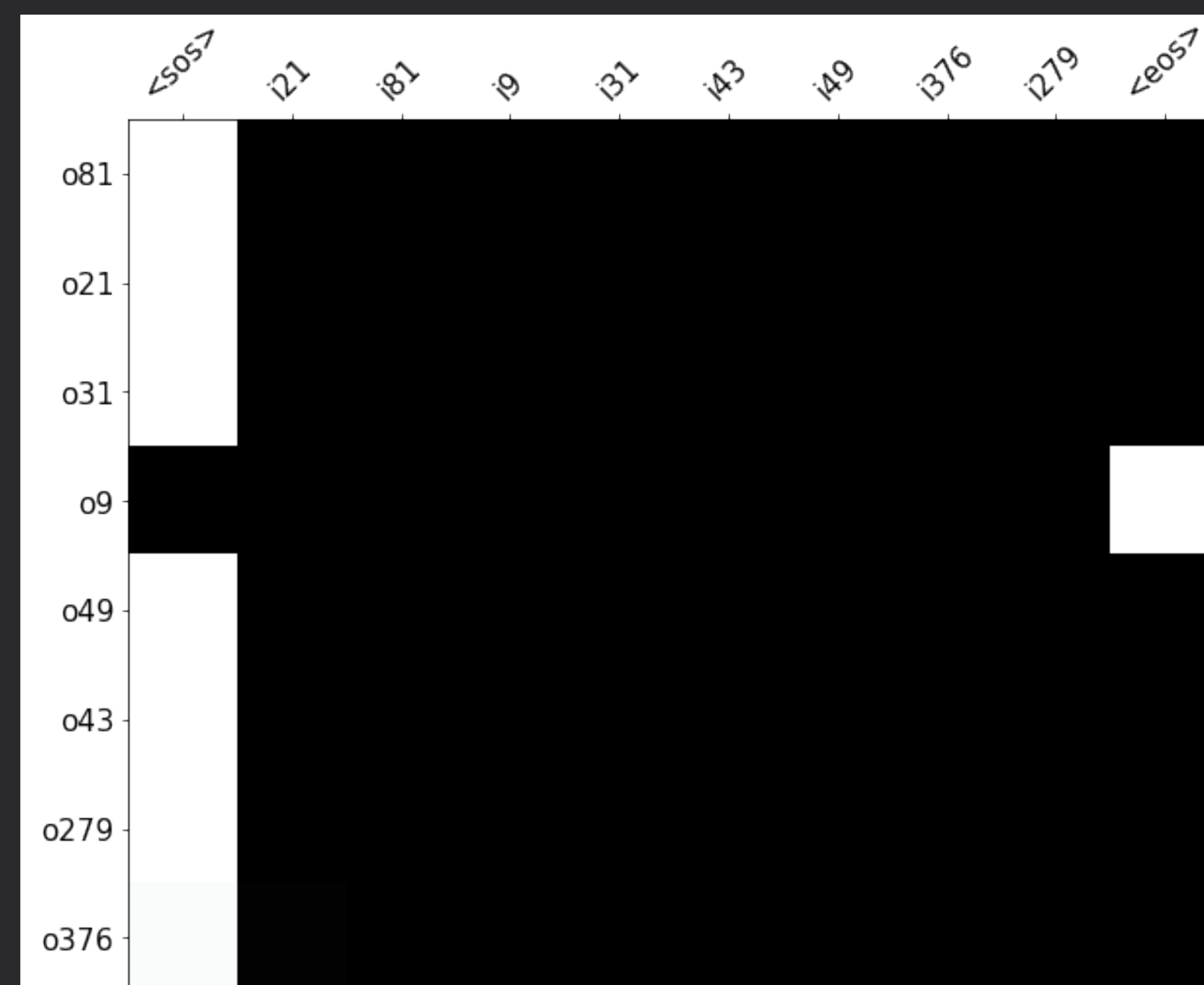
Manipulated

Bigram Flip

A different seed



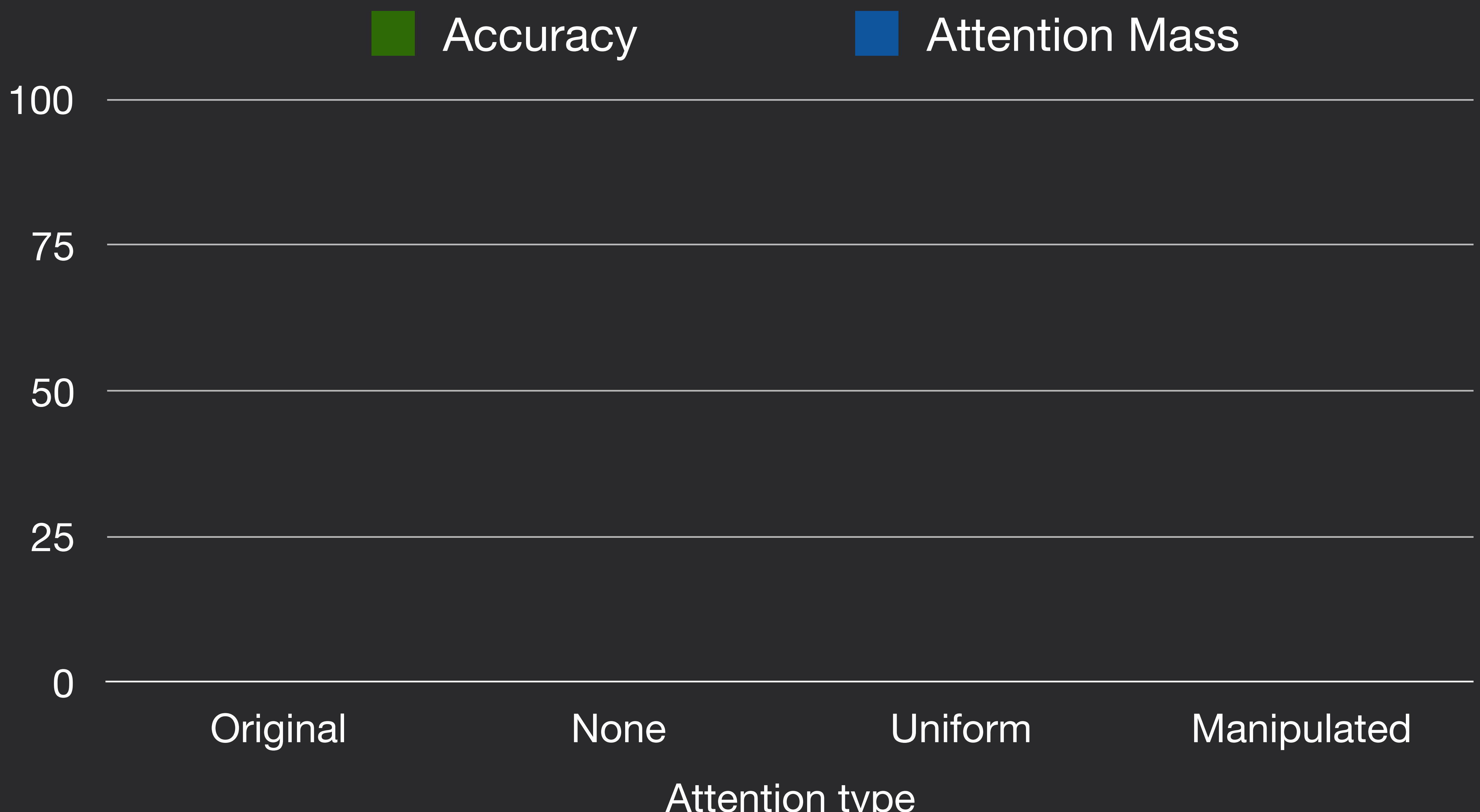
Original



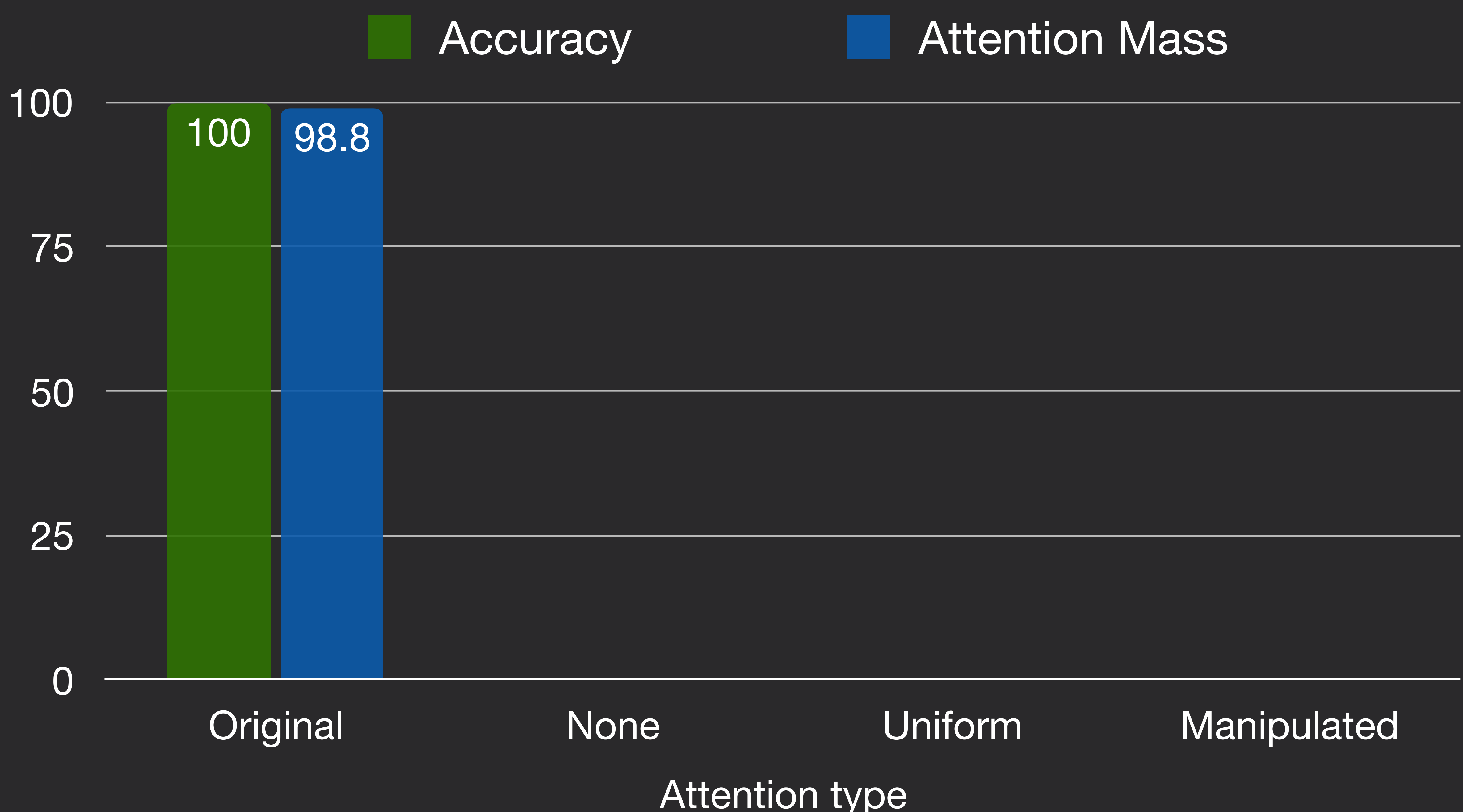
Manipulated

Sequence Copy

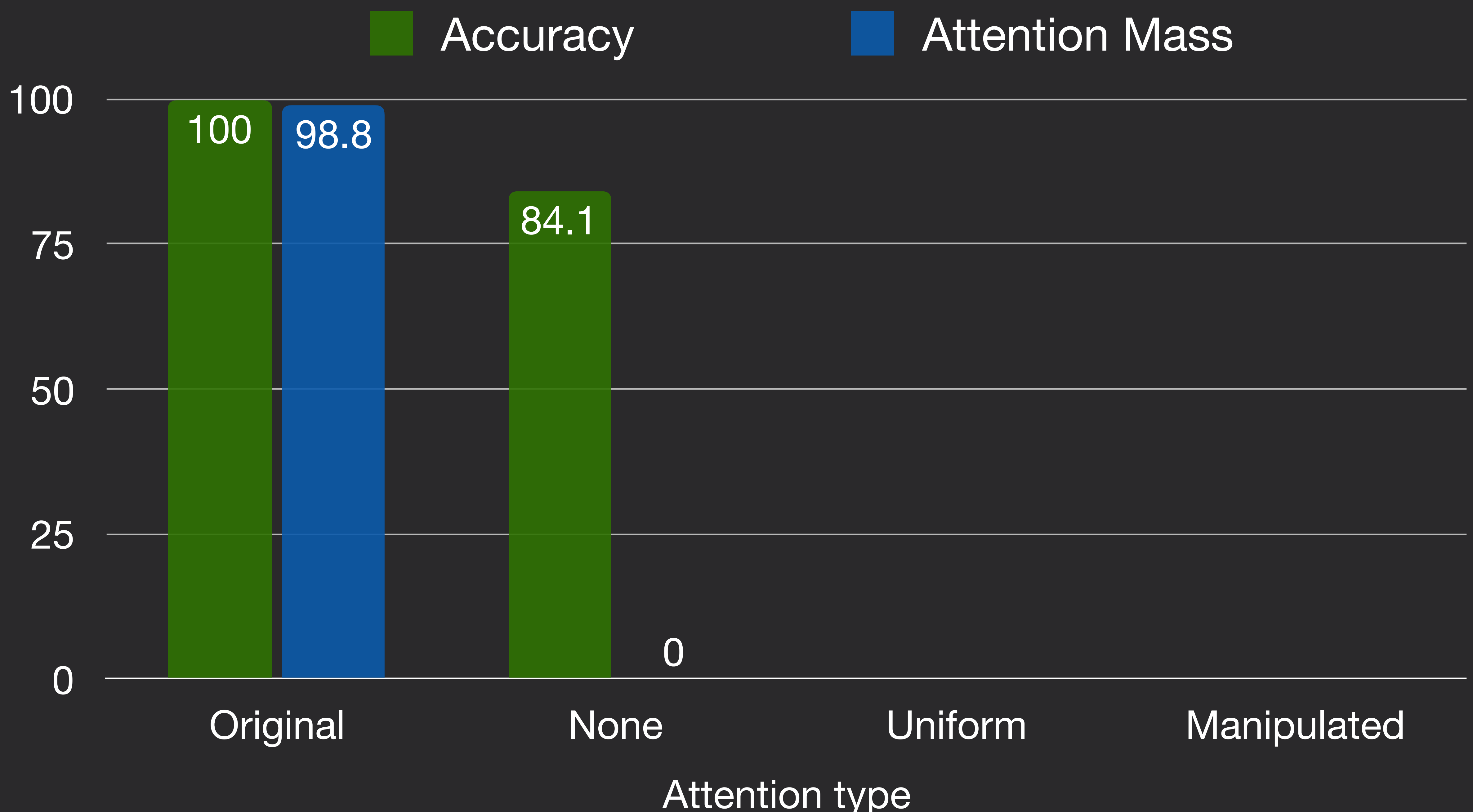
Sequence Copy



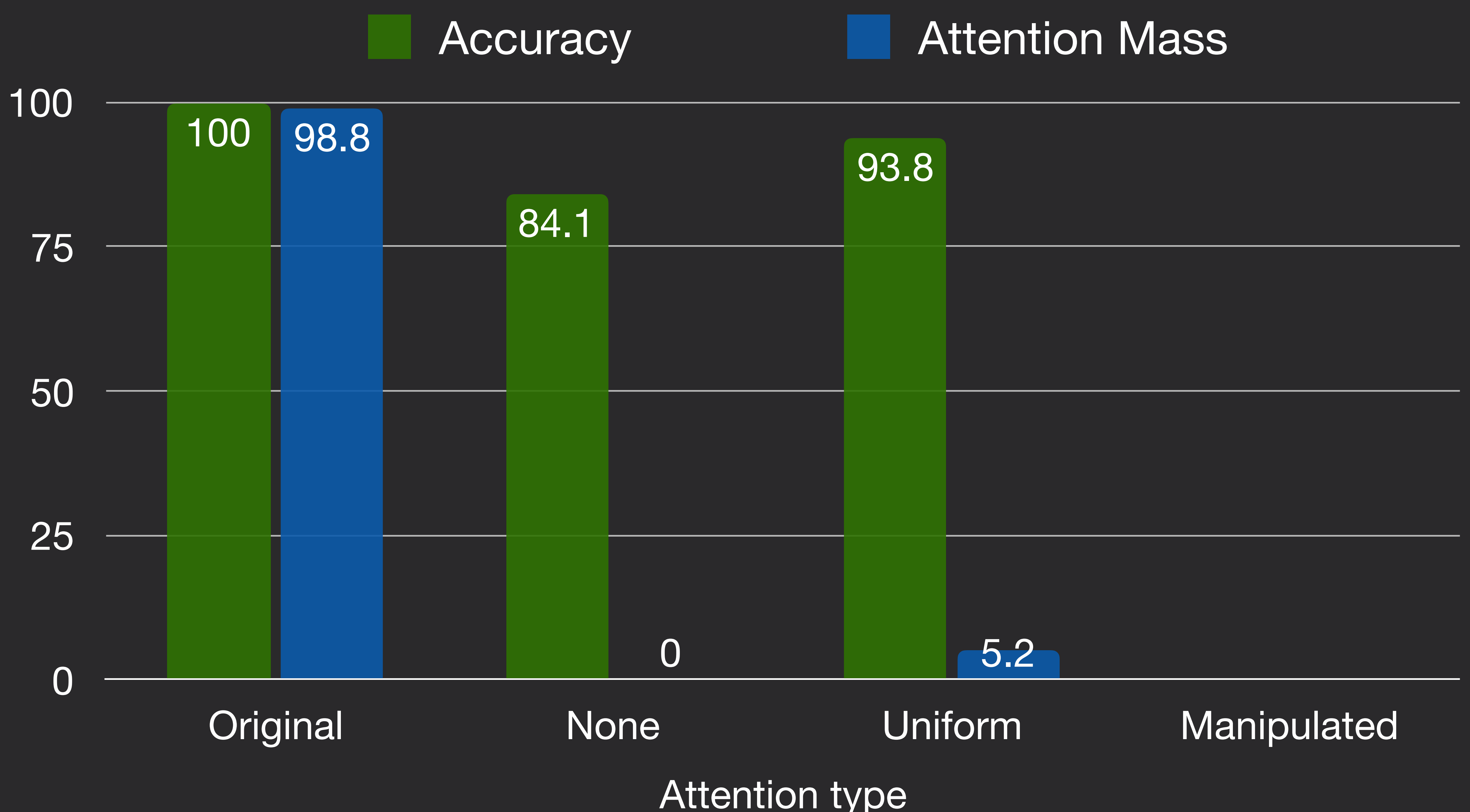
Sequence Copy



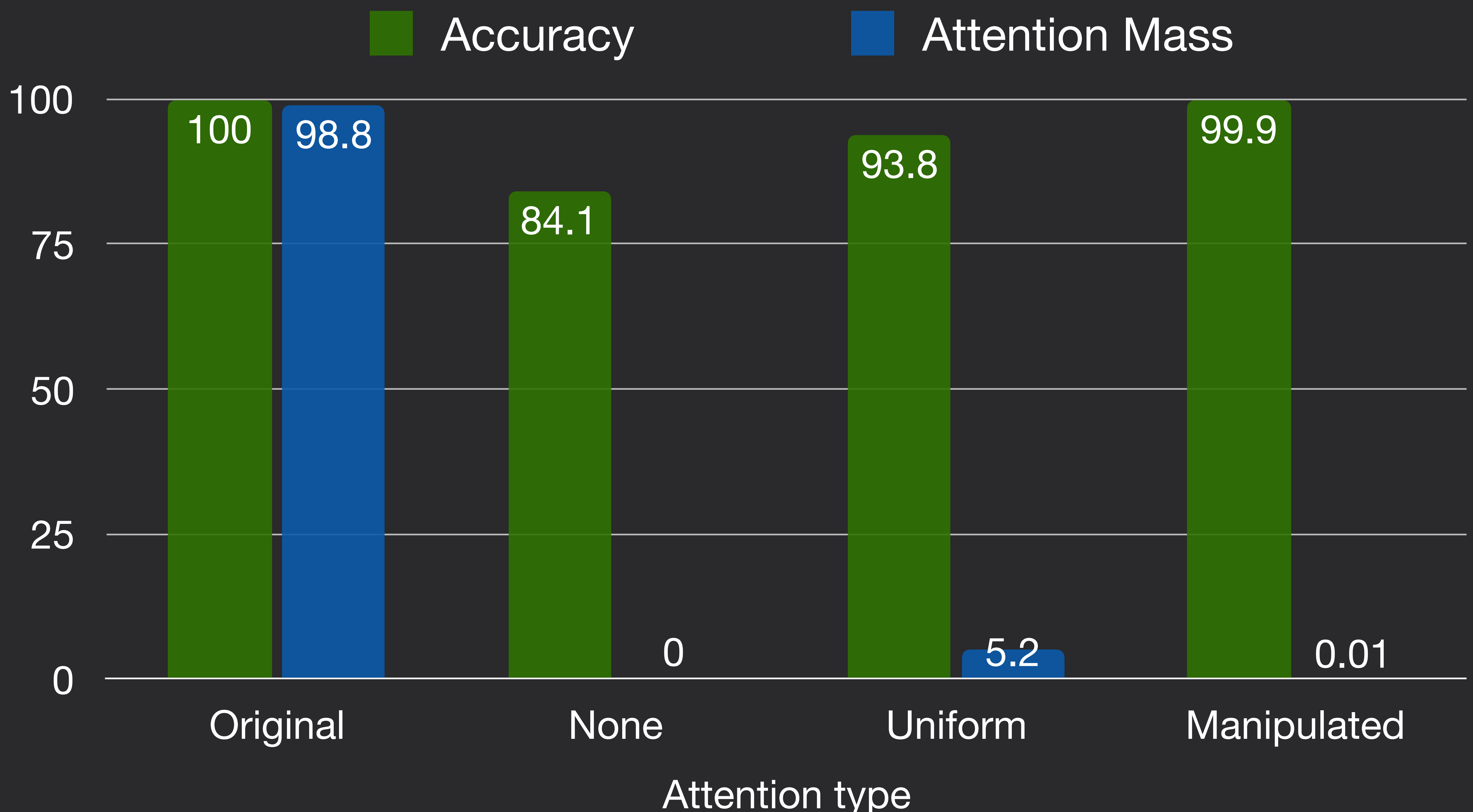
Sequence Copy



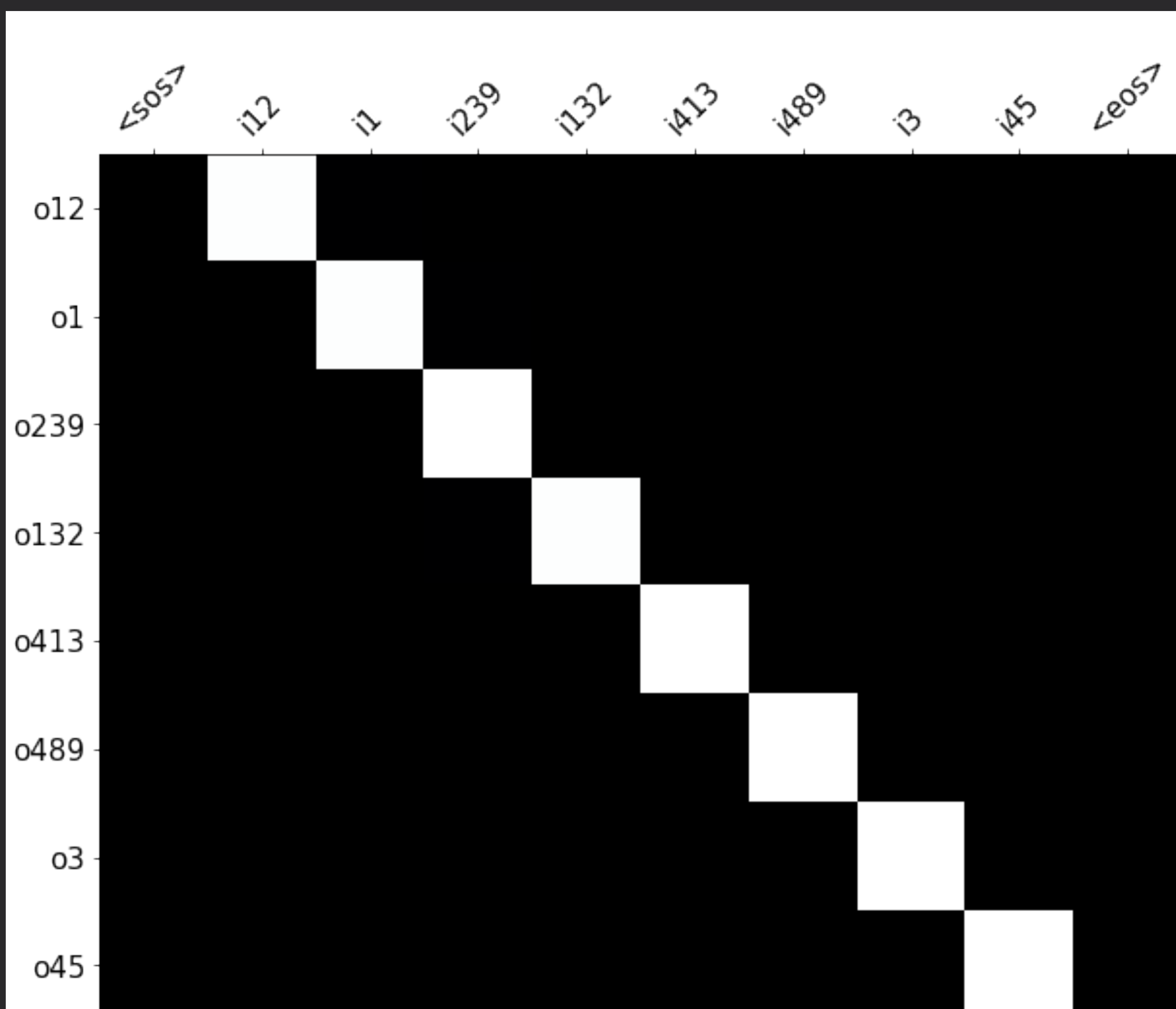
Sequence Copy



Sequence Copy

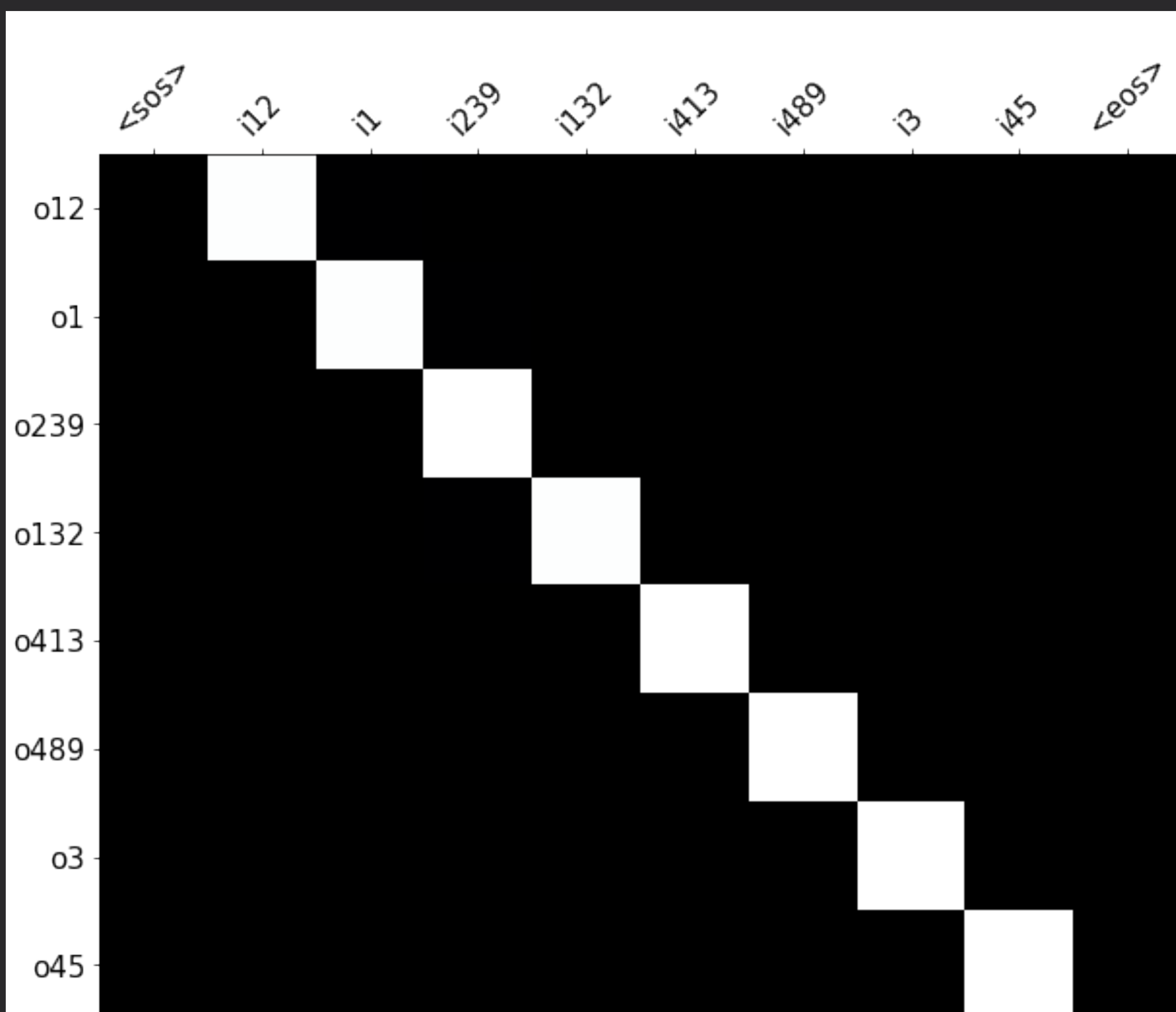


Sequence Copy

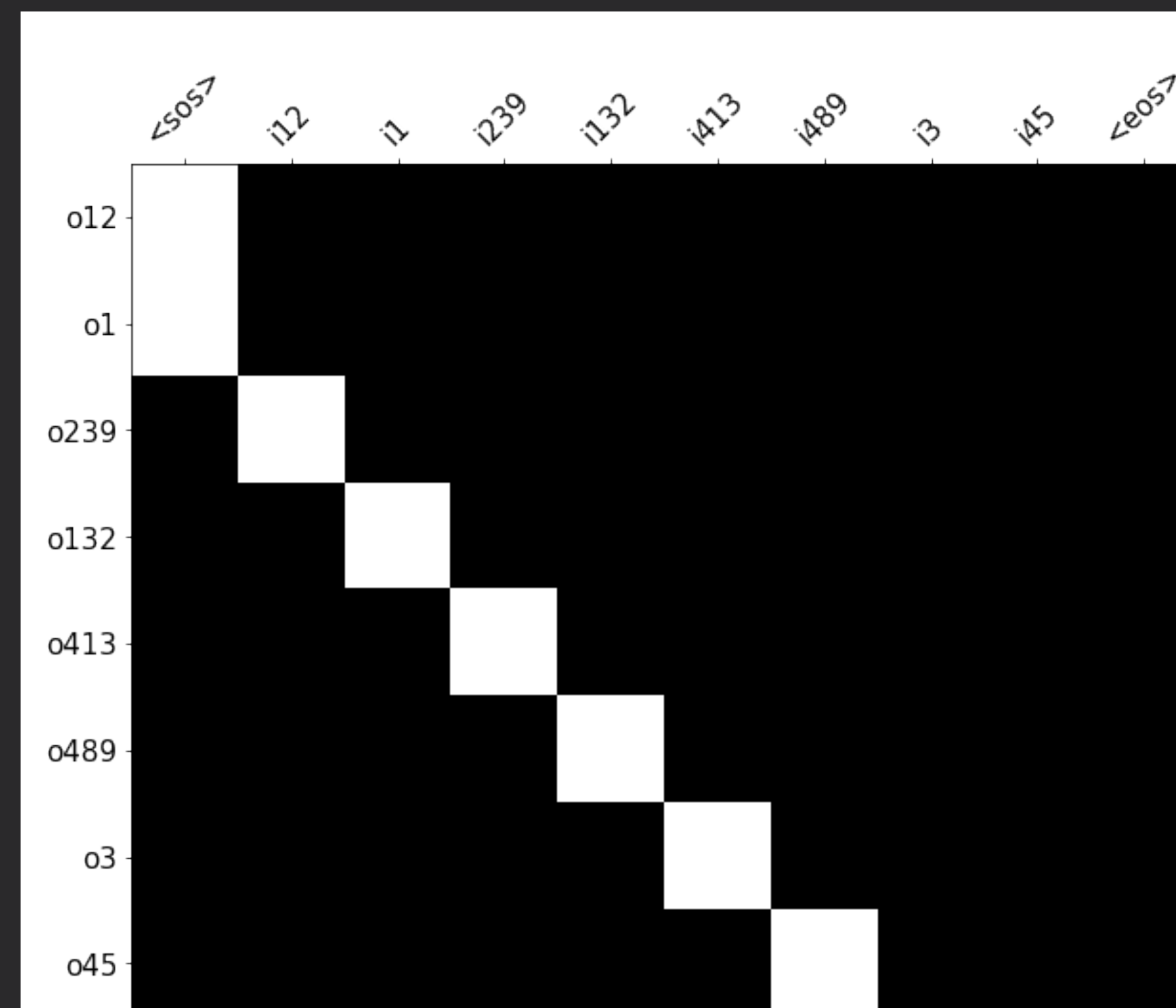


Original

Sequence Copy

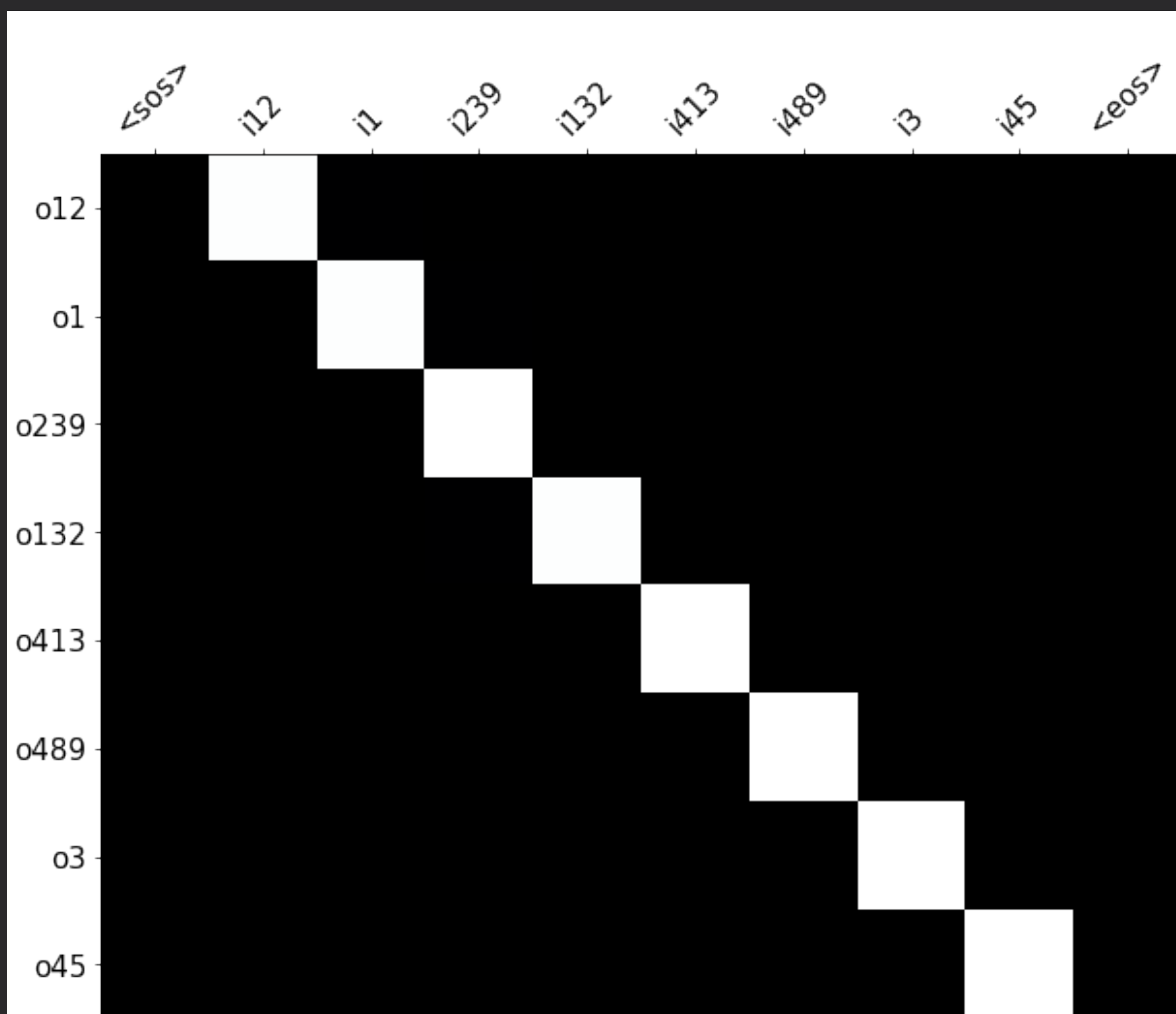


Original



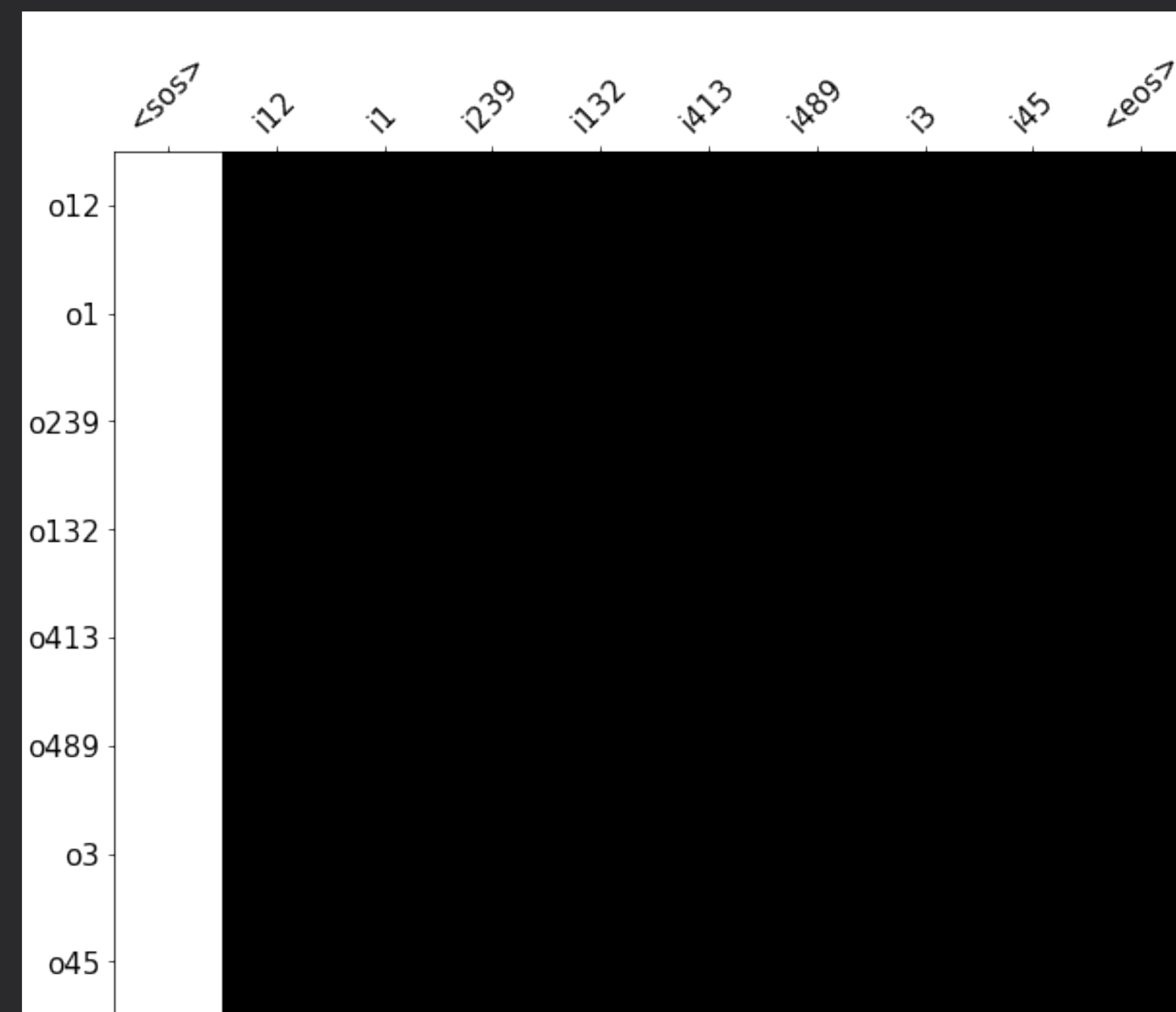
Manipulated

Sequence Copy



Original

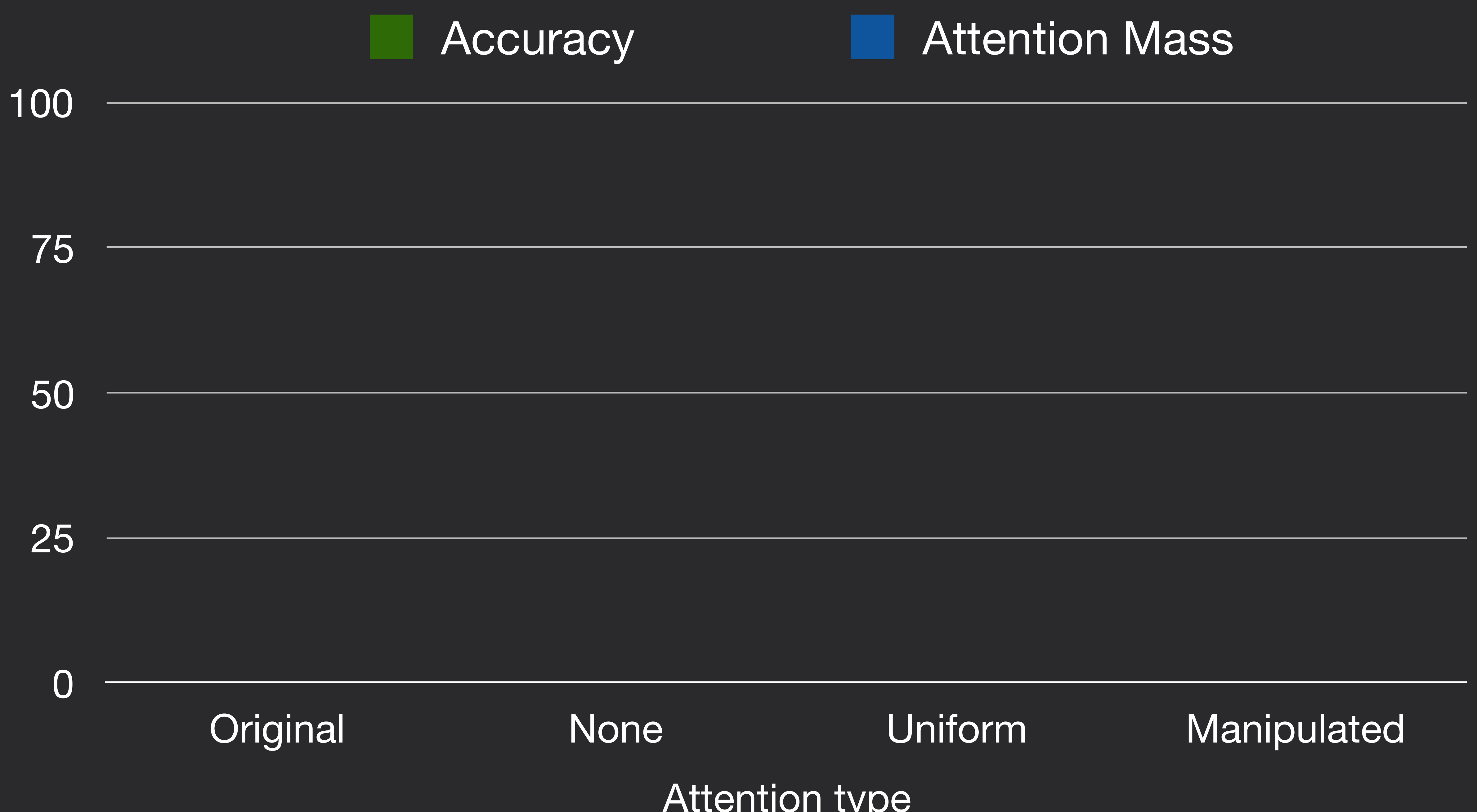
A different seed



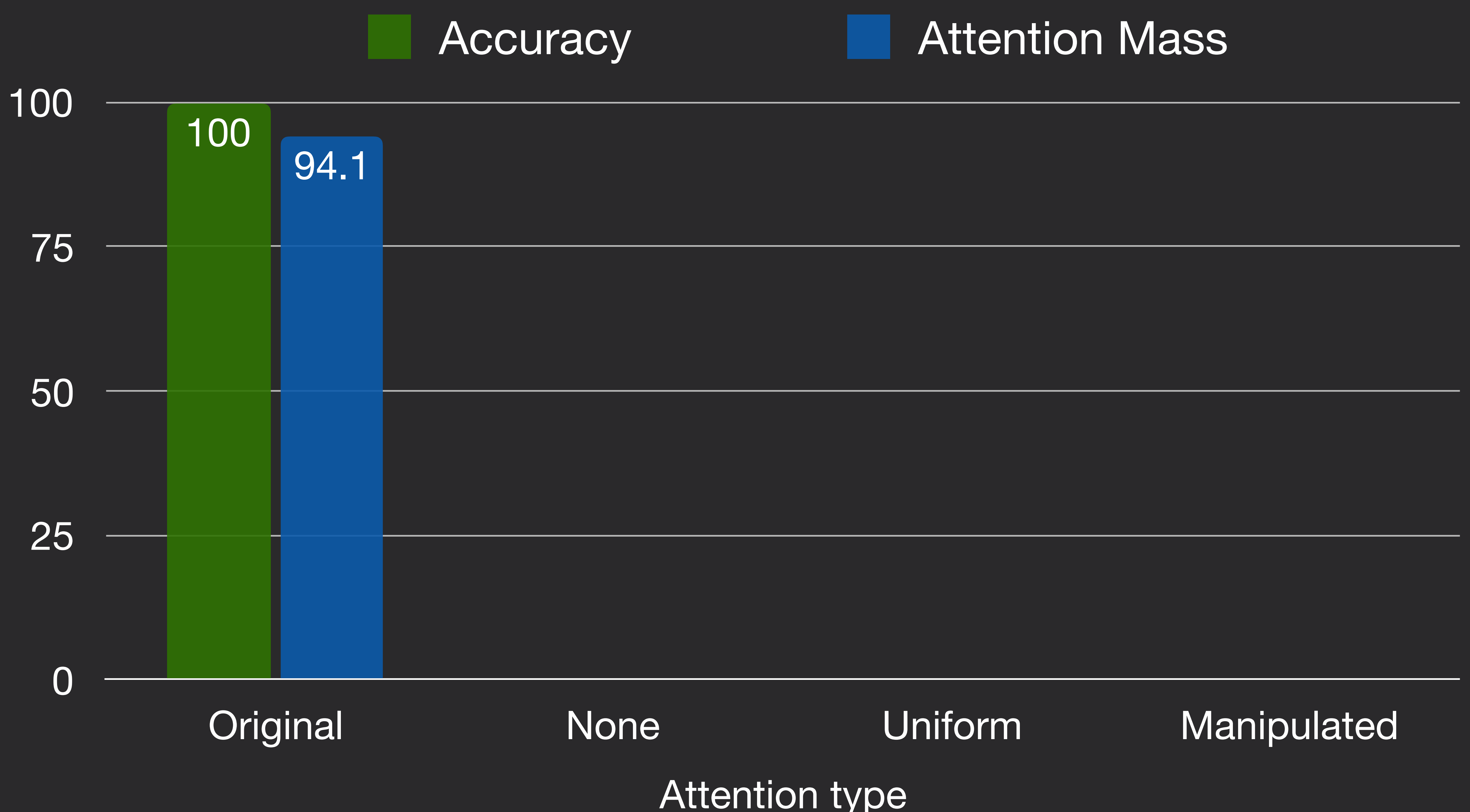
Manipulated

Sequence Reverse

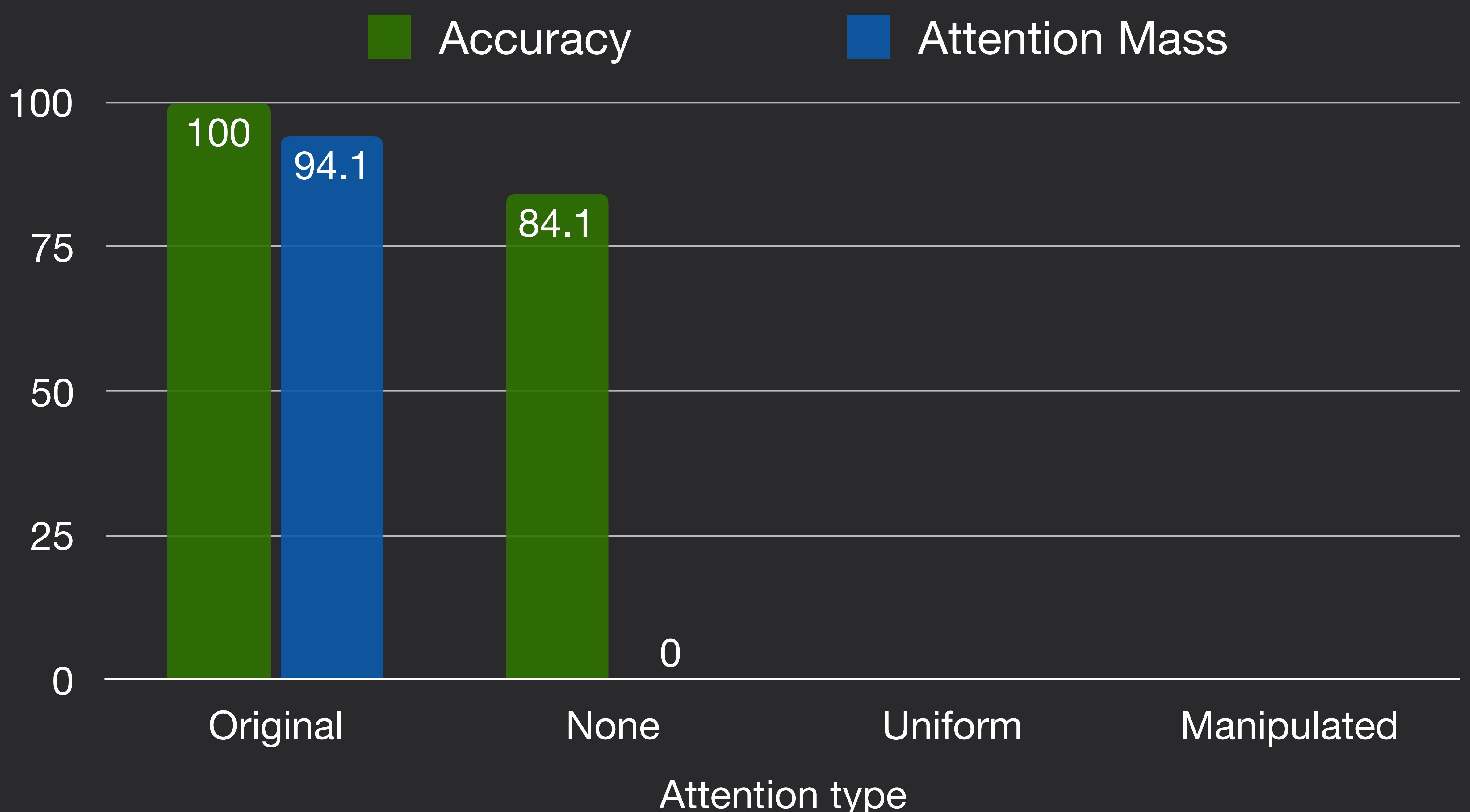
Sequence Reverse



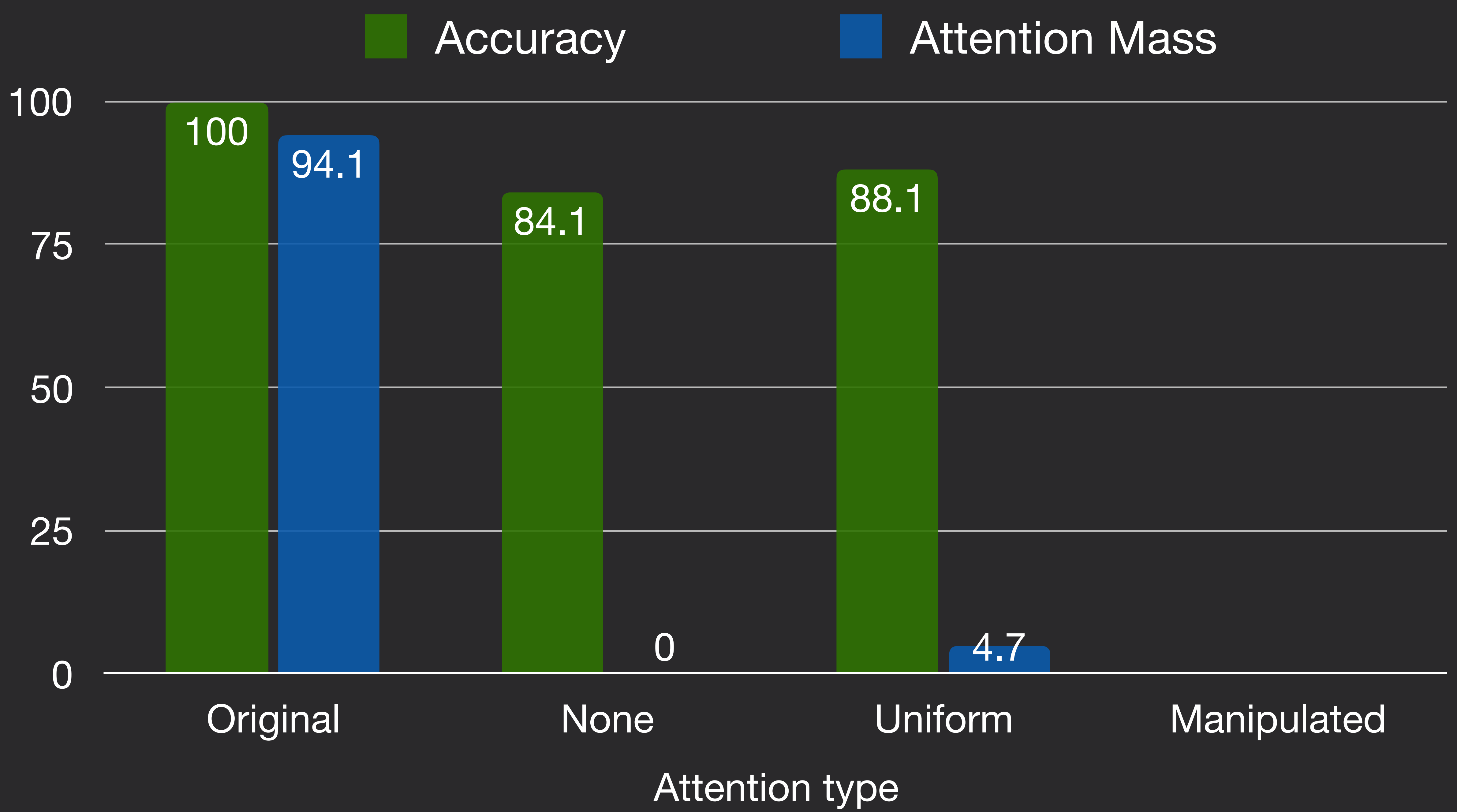
Sequence Reverse



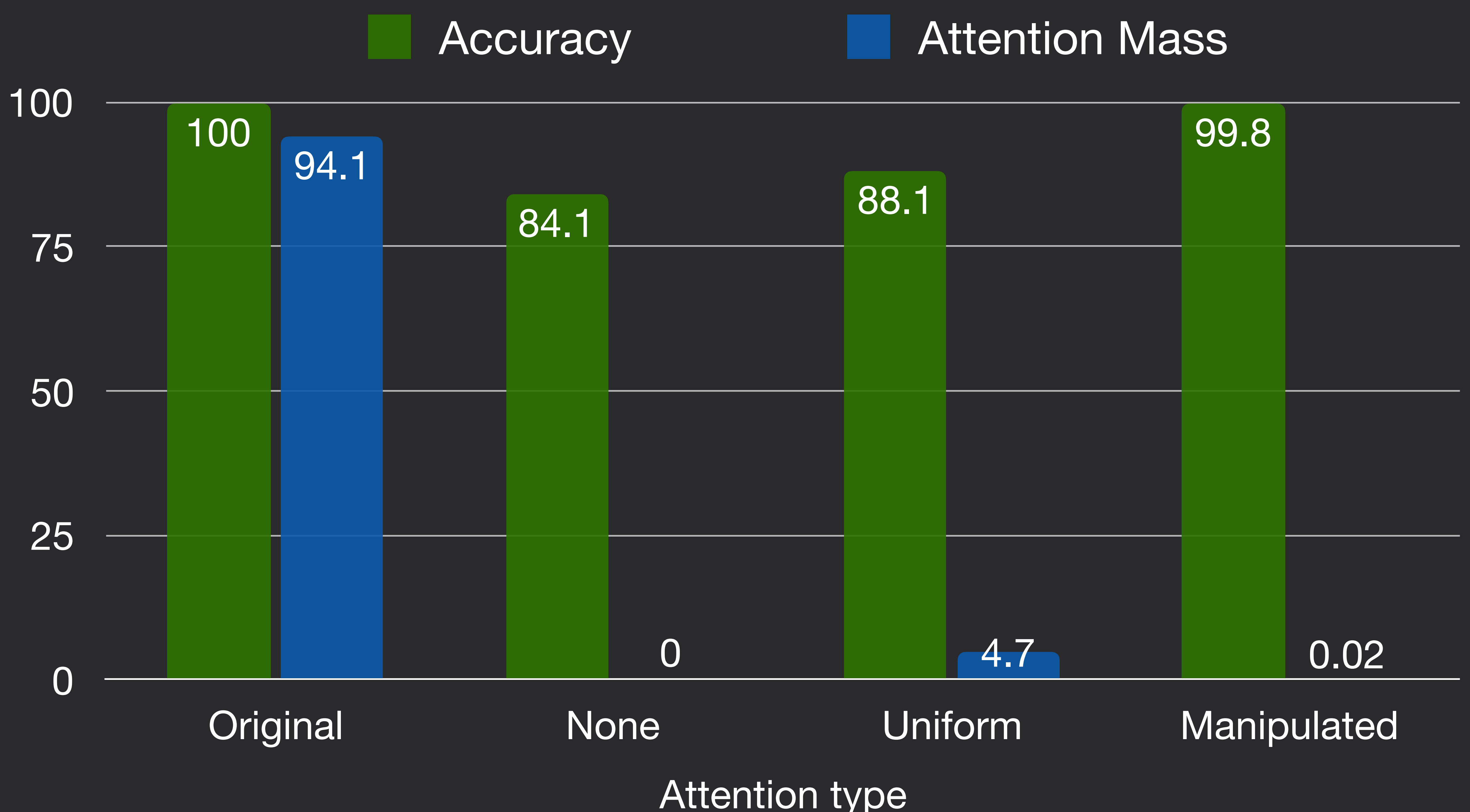
Sequence Reverse



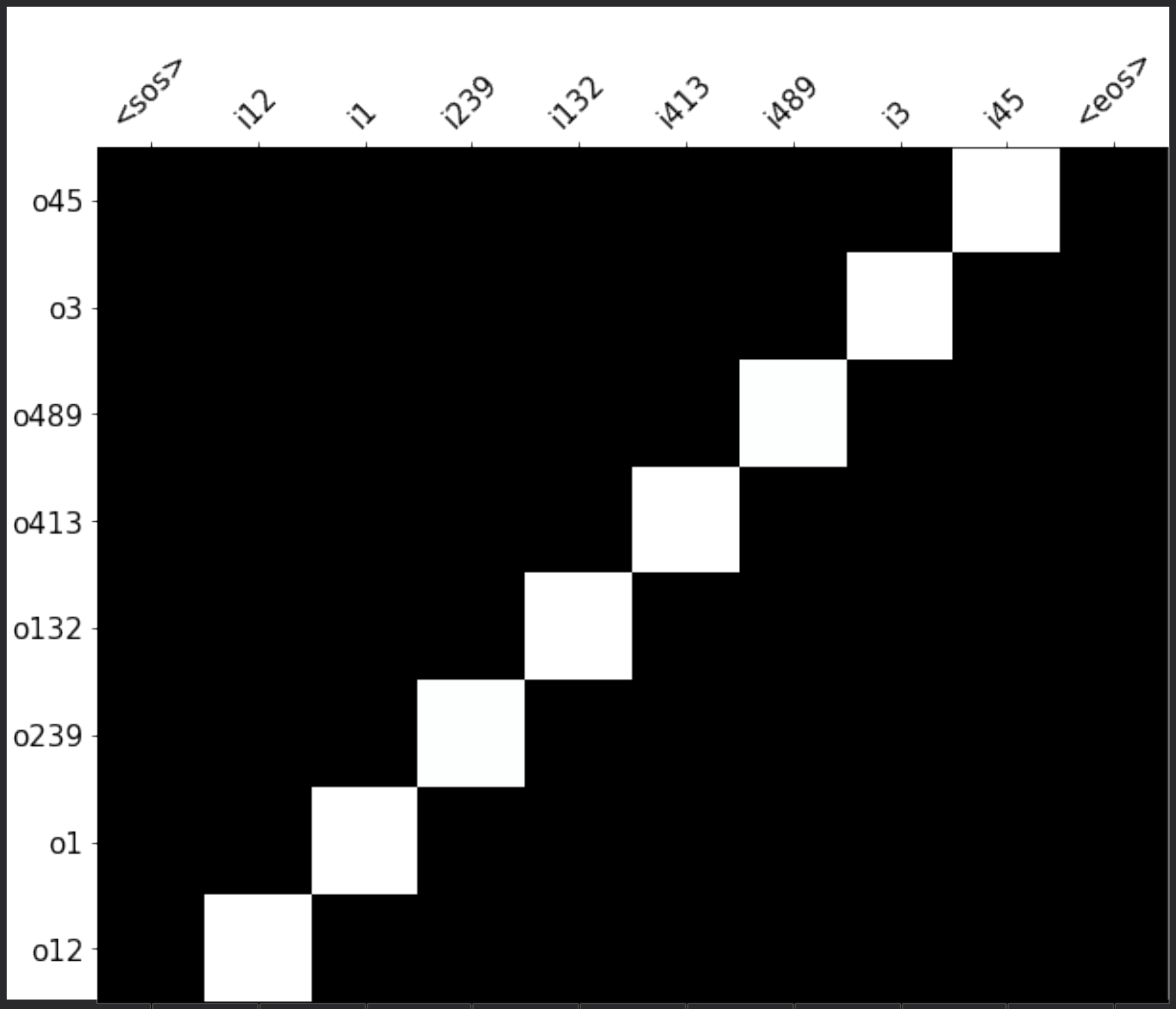
Sequence Reverse



Sequence Reverse

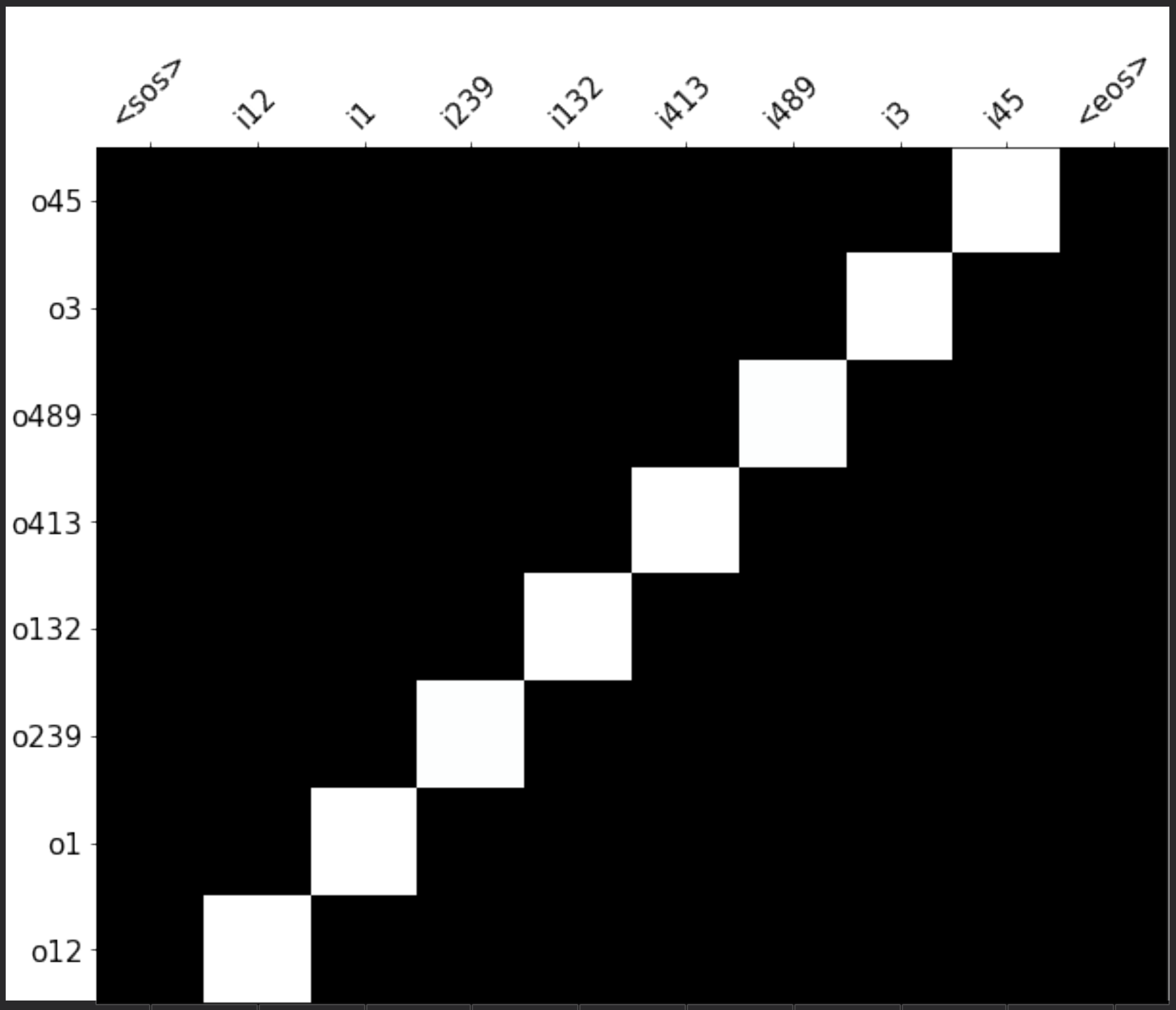


Sequence Reverse

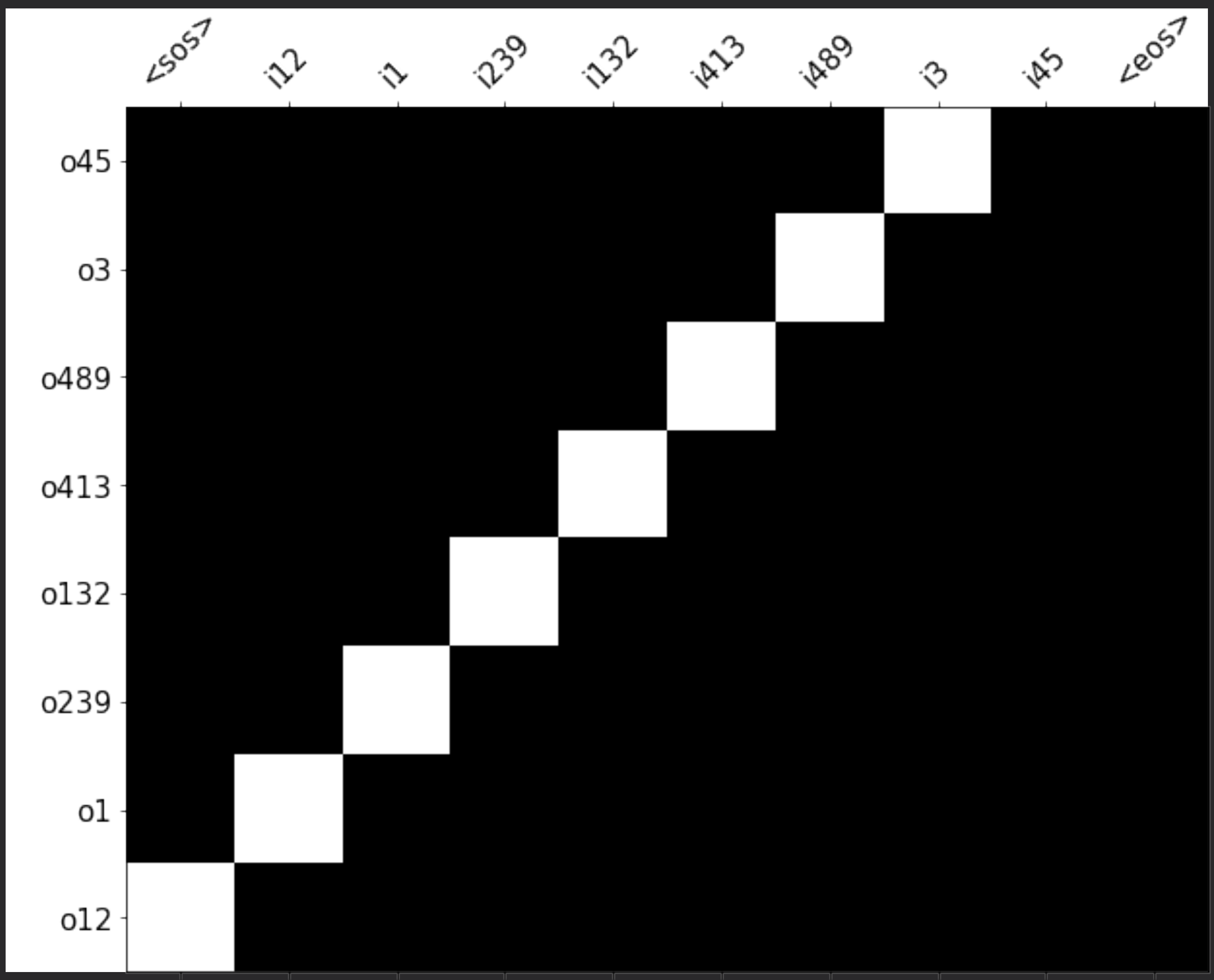


Original

Sequence Reverse

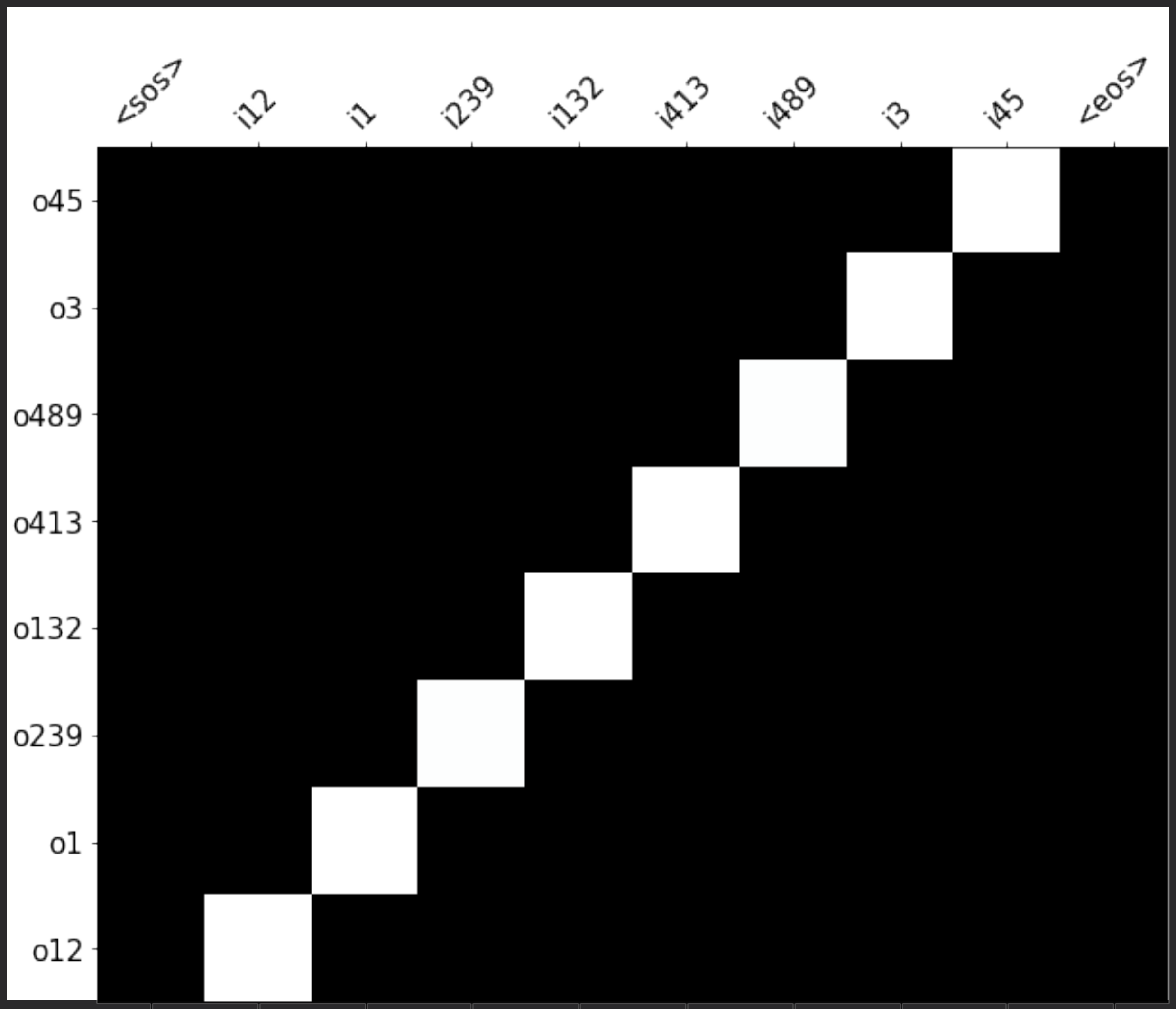


Original



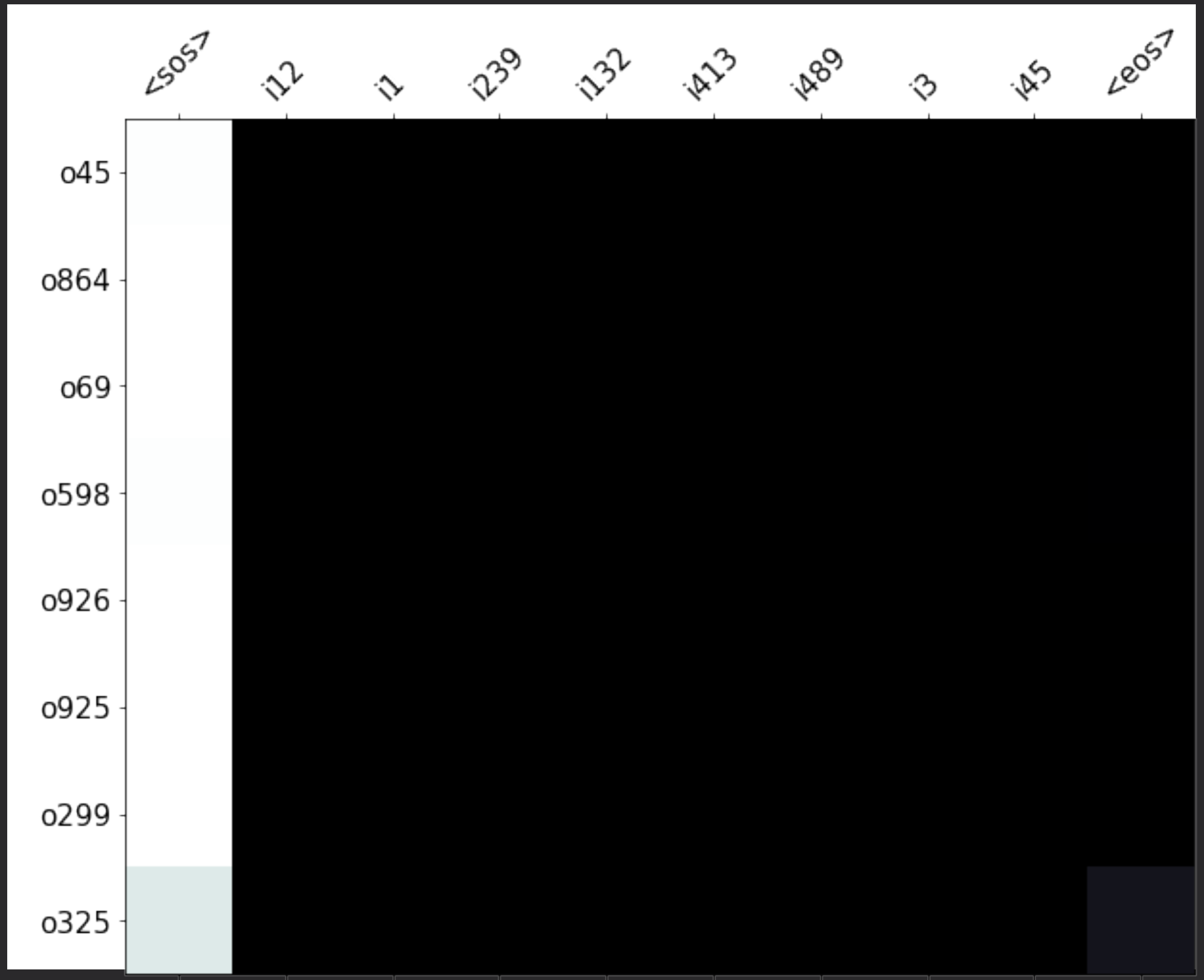
Manipulated

Sequence Reverse



Original

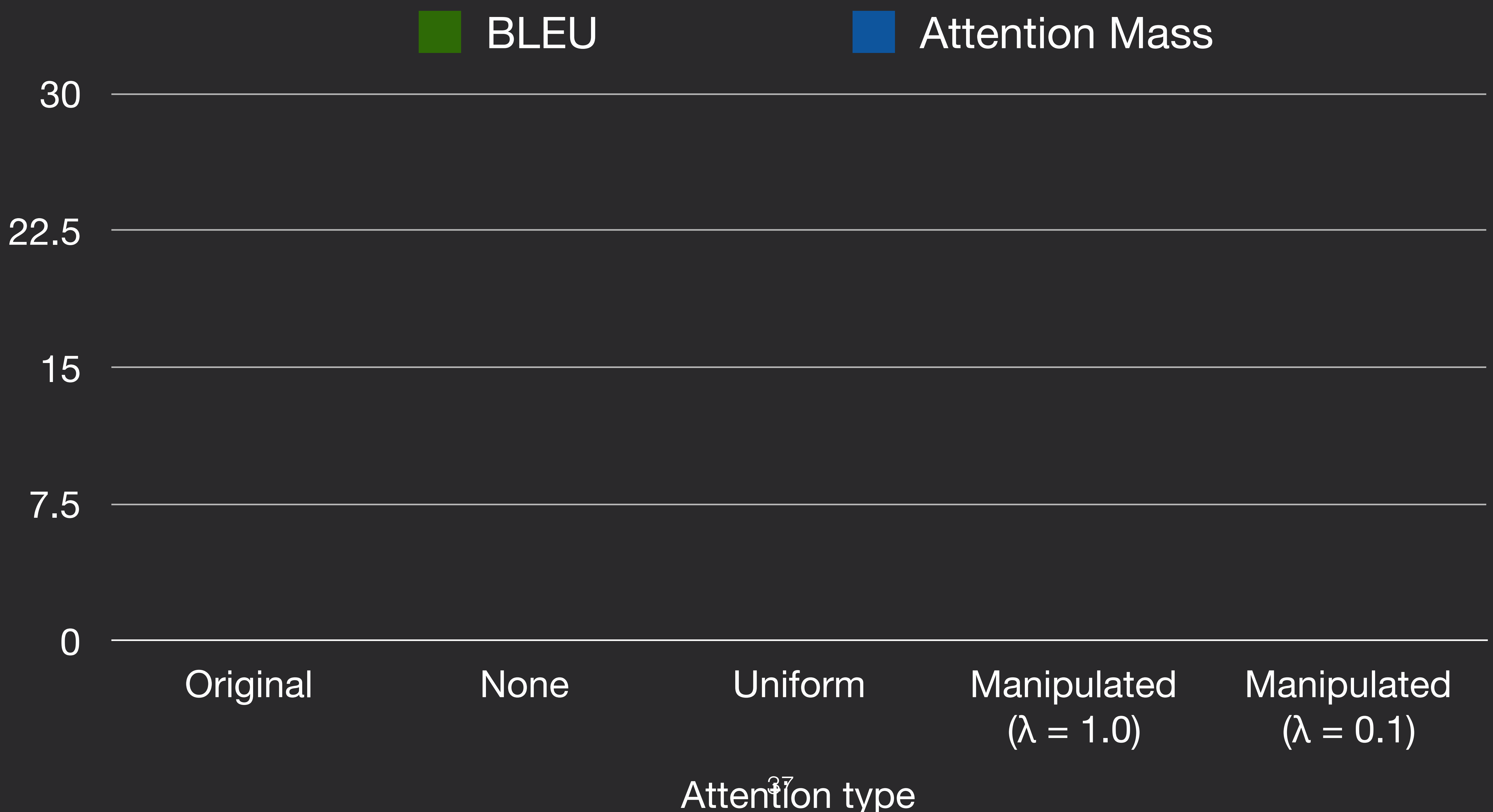
A different seed



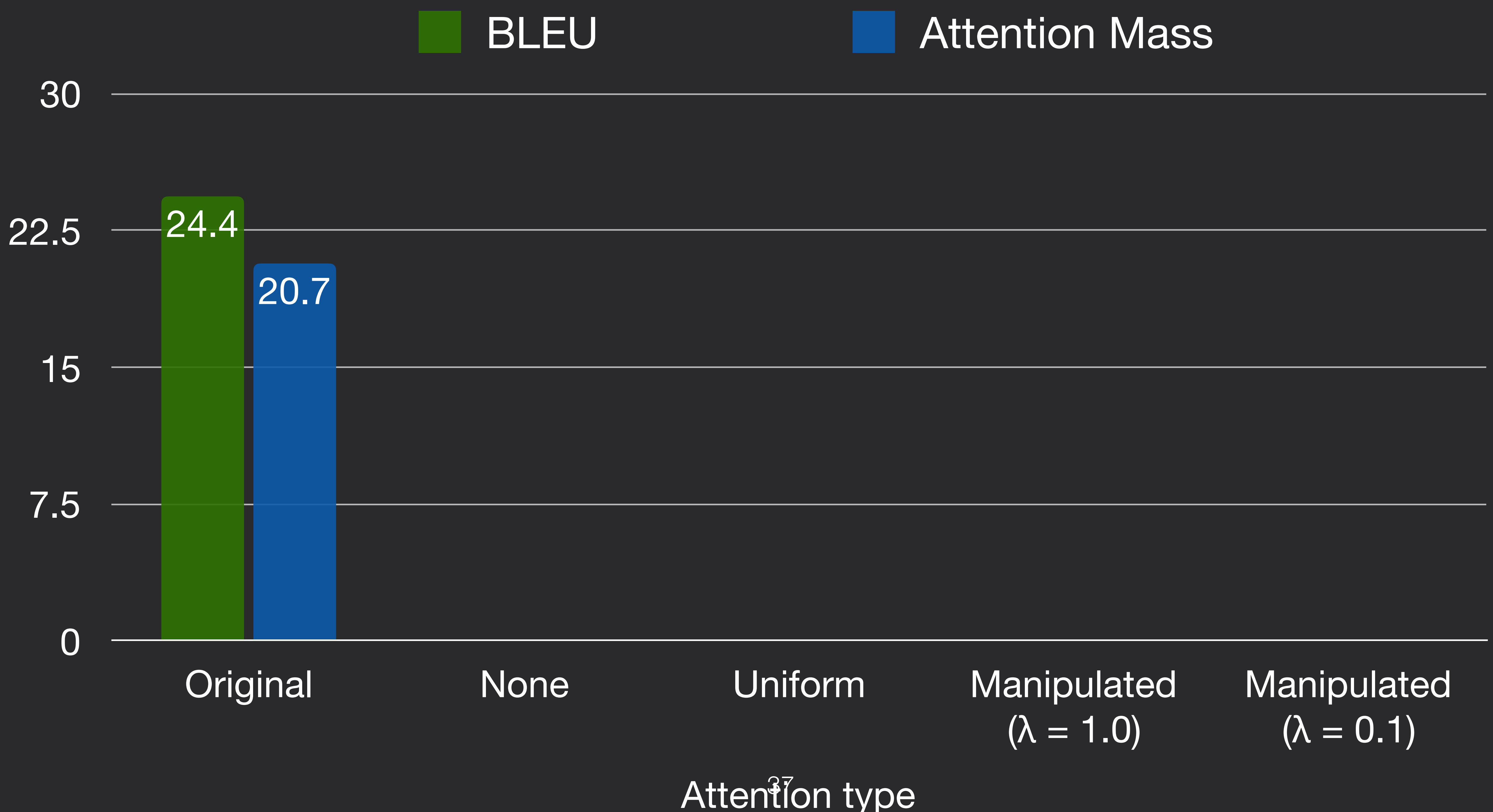
Manipulated

English German MT

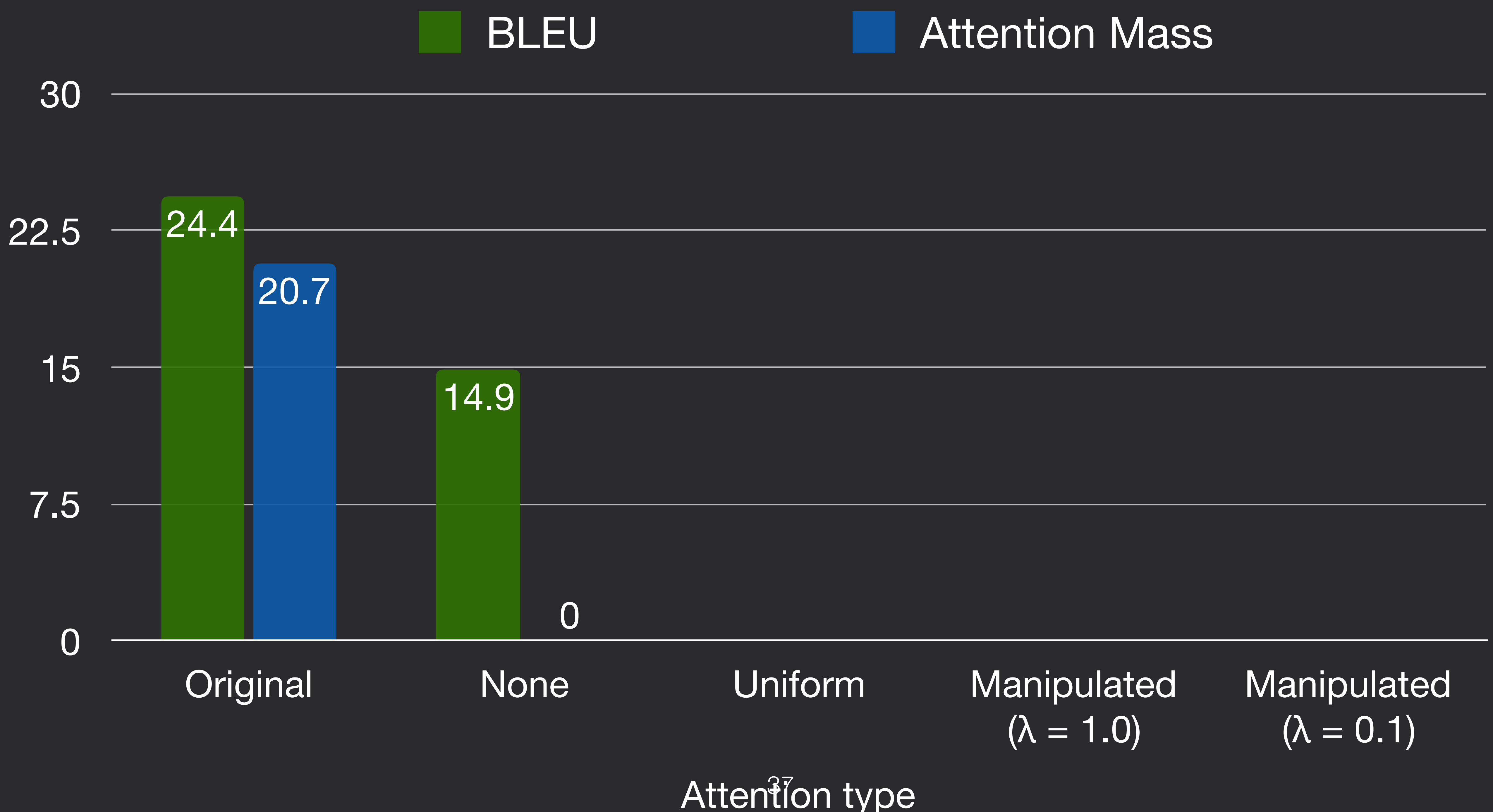
English German MT



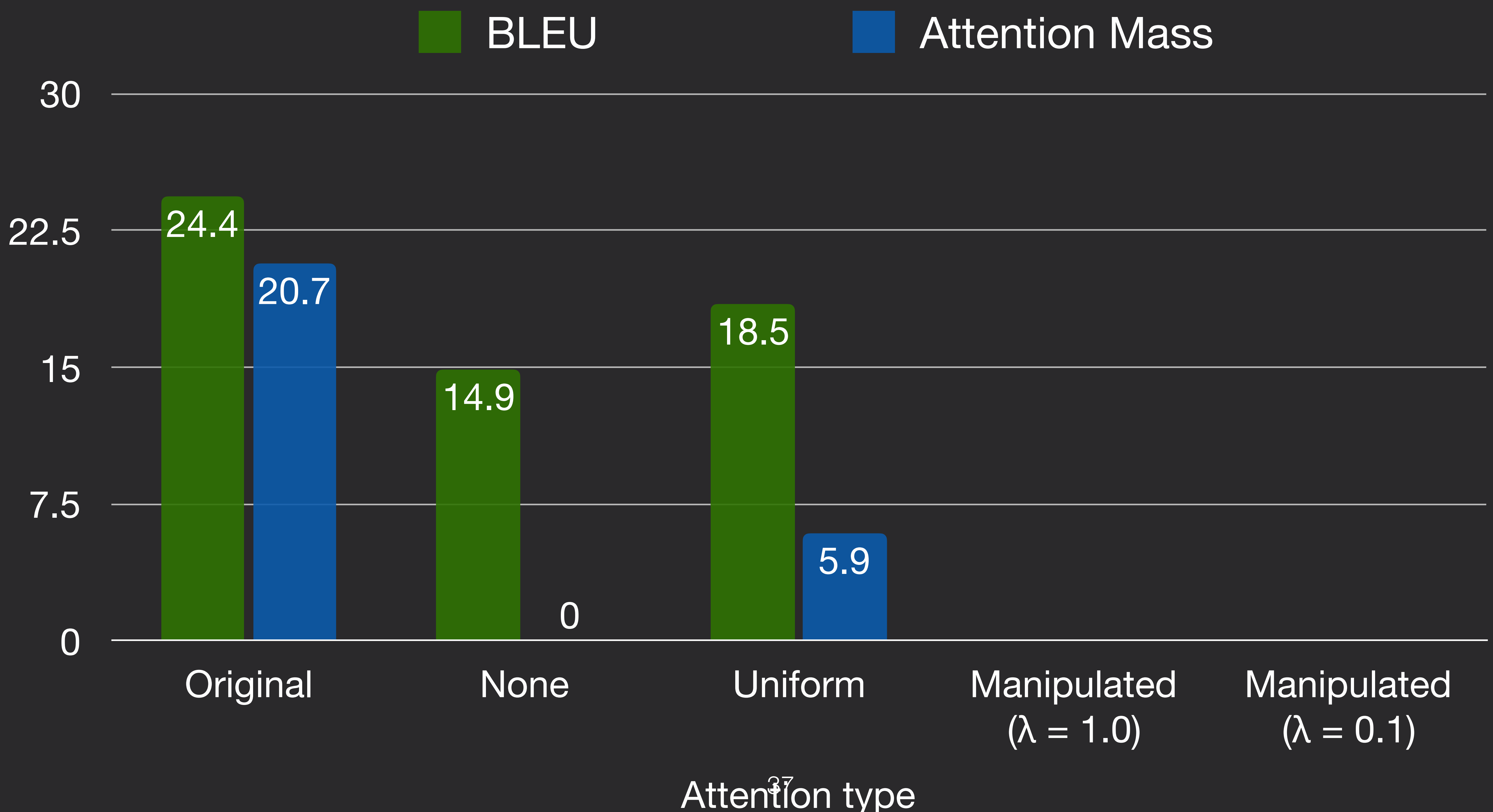
English German MT



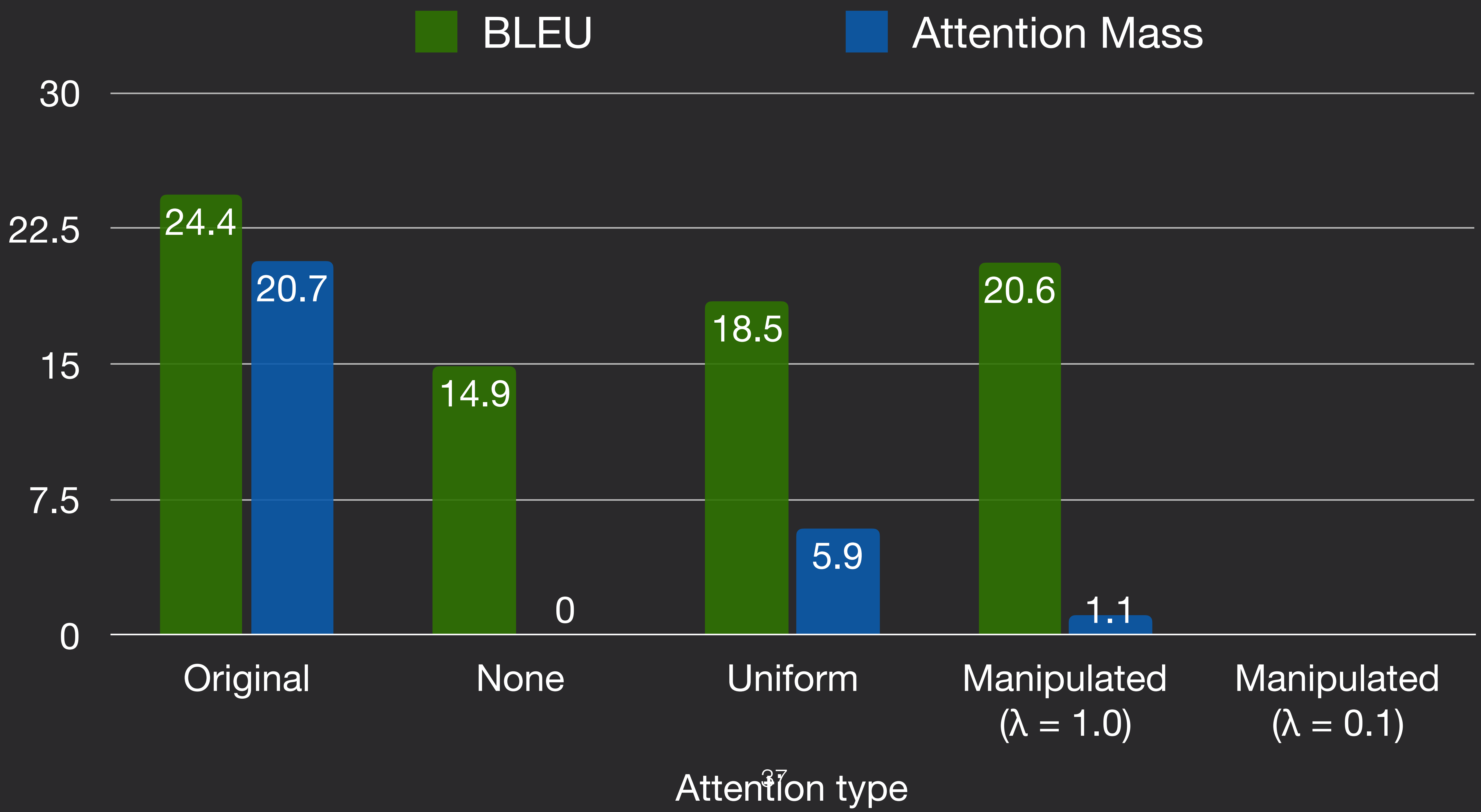
English German MT



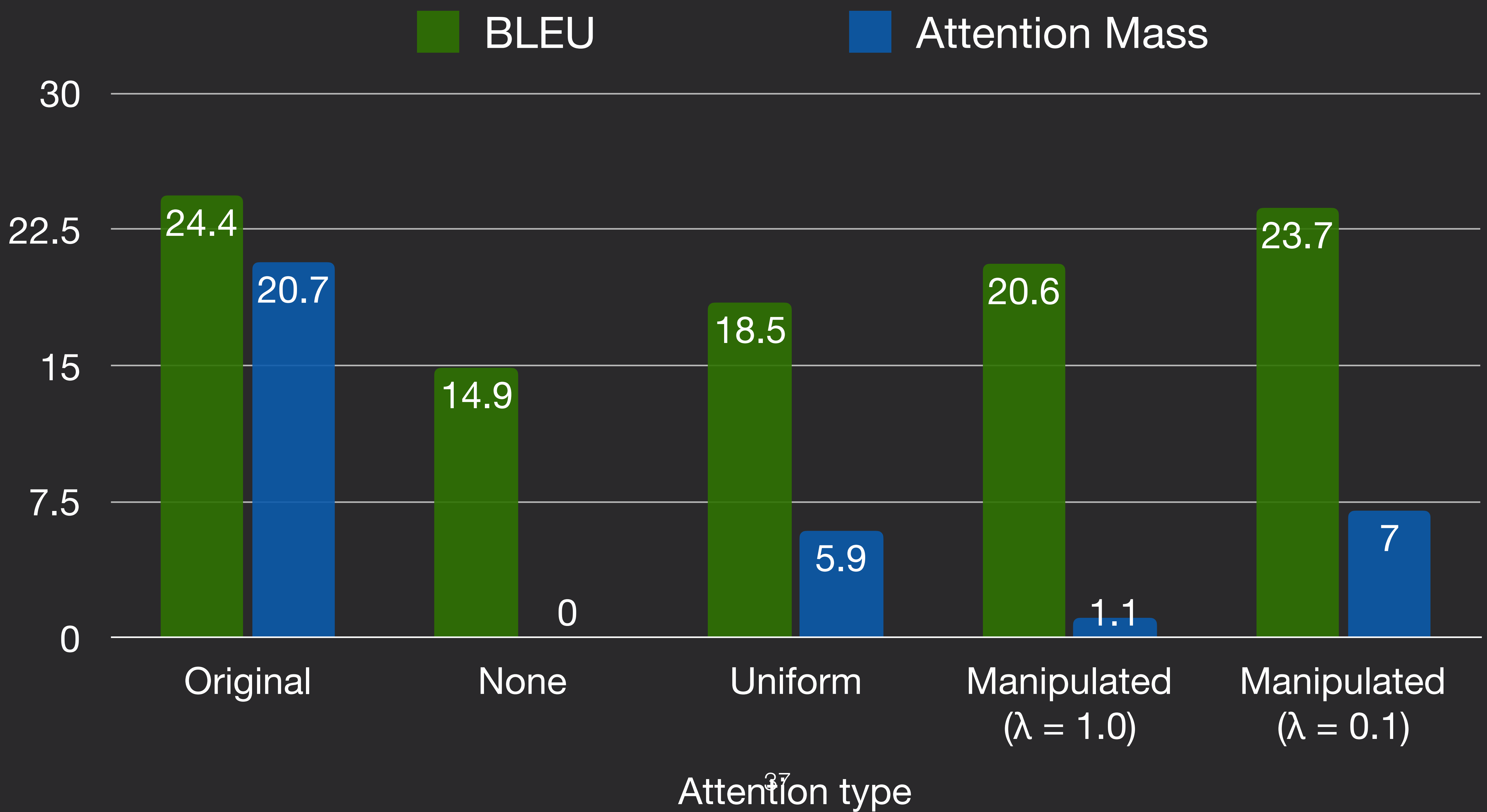
English German MT



English German MT



English German MT



Alternative workarounds

- Through recurrent connections, if they exist.
- Increase in the magnitude of representations corresponding to impermissible tokens.

Human studies

- We present inputs for the task *Occupation Prediction* and the predicted outputs (*Physician or Surgeon*) by one of the models
- We notify the annotators that the input tokens are highlighted on the basis of an “explanation method” (attention weights)
- We ask the annotators two rating questions

Human studies

- Q1: Do you think that this prediction was influenced by the gender of the individual?

Human studies

- Q1: Do you think that this prediction was influenced by the gender of the individual?

Manipulation type	Input example Predicted label - Physician	Percentage of sentences

Human studies

- Q1: Do you think that this prediction was influenced by the gender of the individual?

Manipulation type	Input example Predicted label - Physician	Percentage of sentences
No manipulation	ms. UNK practices medicine in UNK and specializes in urological surgery. ms. UNK is affiliated with menorah medical center	66%

Human studies

- Q1: Do you think that this prediction was influenced by the gender of the individual?

Manipulation type	Input example Predicted label - Physician	Percentage of sentences
No manipulation	ms. UNK practices medicine in UNK and specializes in urological surgery. ms. UNK is affiliated with menorah medical center	66%
Ours	ms. UNK practices medicine in UNK and specializes in urological surgery. ms. UNK is affiliated with menorah medical center	0%

Human studies

- Q1: Do you think that this prediction was influenced by the gender of the individual?

Manipulation type	Input example Predicted label - Physician	Percentage of sentences
No manipulation	ms. UNK practices medicine in UNK and specializes in urological surgery. ms. UNK is affiliated with menorah medical center	66%
Ours	ms. UNK practices medicine in UNK and specializes in urological surgery. ms. UNK is affiliated with menorah medical center	0%
Weigraff et al, 2019	ms. UNK practices medicine in UNK and specializes in urological surgery. ms. UNK is affiliated with menorah medical center	0%

Human studies

- Q2: Do you believe that highlighted tokens capture the model's prediction?

Manipulation type	Input example Predicted label - Physician	Rating (1 to 4)
No manipulation	ms. UNK practices medicine in UNK and specializes in urological surgery. ms. UNK is affiliated with menorah medical center	3.0 / 4
Ours	ms. UNK practices medicine in UNK and specializes in urological surgery. ms. UNK is affiliated with menorah medical center	2.67 / 4
Weigraff et al, 2019	ms. UNK practices medicine in UNK and specializes in urological surgery. ms. UNK is affiliated with menorah medical center	1.0 / 4

Outline

1. What Is attention mechanism?
2. Attention-as-explanations
3. Manipulating attention weights
4. Results and discussion
5. **Conclusion**

Conclusion

Conclusion

- In organic cases, typically attention is high for the 'right' tokens. Consistent across different seeds.

Conclusion

- In organic cases, typically attention is high for the 'right' tokens. Consistent across different seeds.
- Often attention is easy to manipulate with negligible drop in accuracy.

Conclusion

- In organic cases, typically attention is high for the 'right' tokens. Consistent across different seeds.
- Often attention is easy to manipulate with negligible drop in accuracy.
- Models with manipulated attention often perform better compared against models with no or uniform attention.

Conclusion

- In organic cases, typically attention is high for the 'right' tokens. Consistent across different seeds.
- Often attention is easy to manipulate with negligible drop in accuracy.
- Models with manipulated attention often perform better compared against models with no or uniform attention.
- Multiple possible ways to find alternate mechanisms that are not consistent with one another.

THANK YOU FOR YOUR ATTENTION

Questions?

Discussion points

- "maybe we can come up with techniques and metrics to compute the reliability of attention for an explanation, for a general model"
- "While the paper points out a major problem in the way attention is conceived, it does not make any effort to offer a solution."

Discussion points

- "I would have loved to see some more work on showing that if [accuracy] scores were retained even after changing the attention weights, then what exactly is the model focussing on for its predictions"