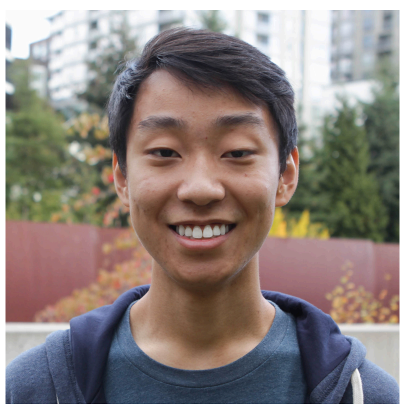
Linguistic Knowledge and Transferability of Contextual Representations





Nelson F. Liu | Matt Gardner | Yonatan Belinkov | Matthew E. Peters | Noah A. Smith

NAACL 2019—June 3, 2019

[McCann et al., 2017; Peters et al., 2018a; Devlin et al., 2019, *inter alia*]

Contextual Word Representations Are Extraordinarily Effective

- Contextual word representations (from ***contextualizers*** like ELMo or BERT) work well on many NLP tasks.
- But **why** do they work so well?
- Better understanding enables principled enhancement.
- **This work:** studies a few questions about their generalizability and transferability.

(1) Probing Contextual Representations

Question: Is the information necessary for a variety of core NLP tasks linearly recoverable from contextual word representations?

Answer: Yes, to a great extent! Tasks with lower performance may require fine-grained linguistic knowledge.

(2) How Does Transferability Vary?

Question: How does transferability vary across contextualizer layers?

Answer: First layer in LSTMs is the most transferable. Middle layers for transformers.

(3) Why Does Transferability Vary?

Question: **Why** does transferability vary across contextualizer layers?

Answer: It depends on pretraining task-specificity!

(4) Alternative Pretraining Objectives

Question: How does language model pretraining compare to alternatives?

Answer: Even with 1 million tokens, language model pretraining yields the most transferable representations.

But, transferring between related tasks does help.

[Shi et al., 2016; Adi et al., 2017]

Probing Models

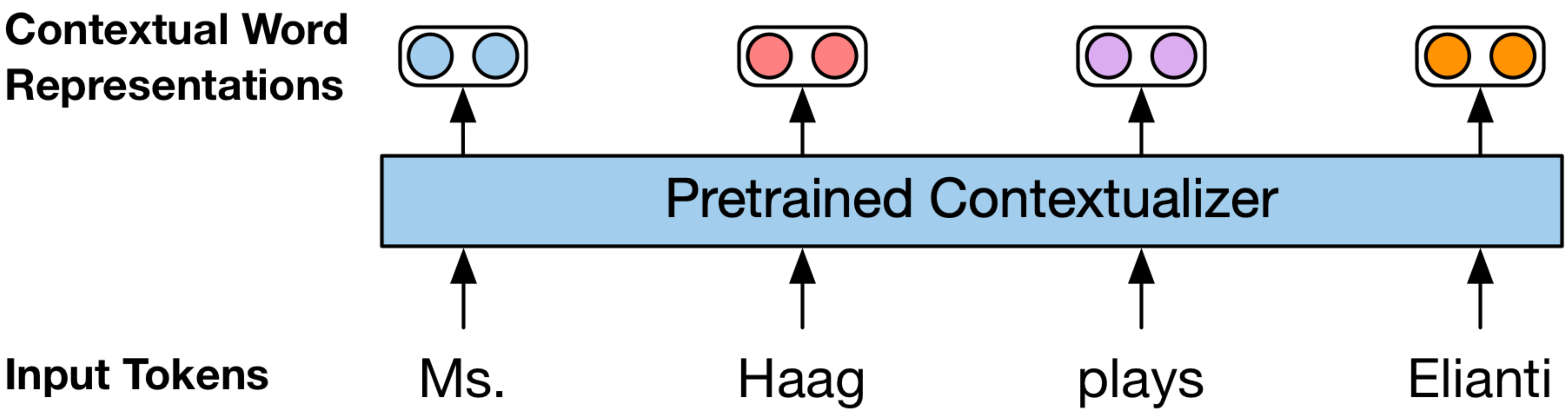
[Shi et al., 2016; Adi et al., 2017]

Probing Models

Input Tokens Ms. Haag plays Elianti

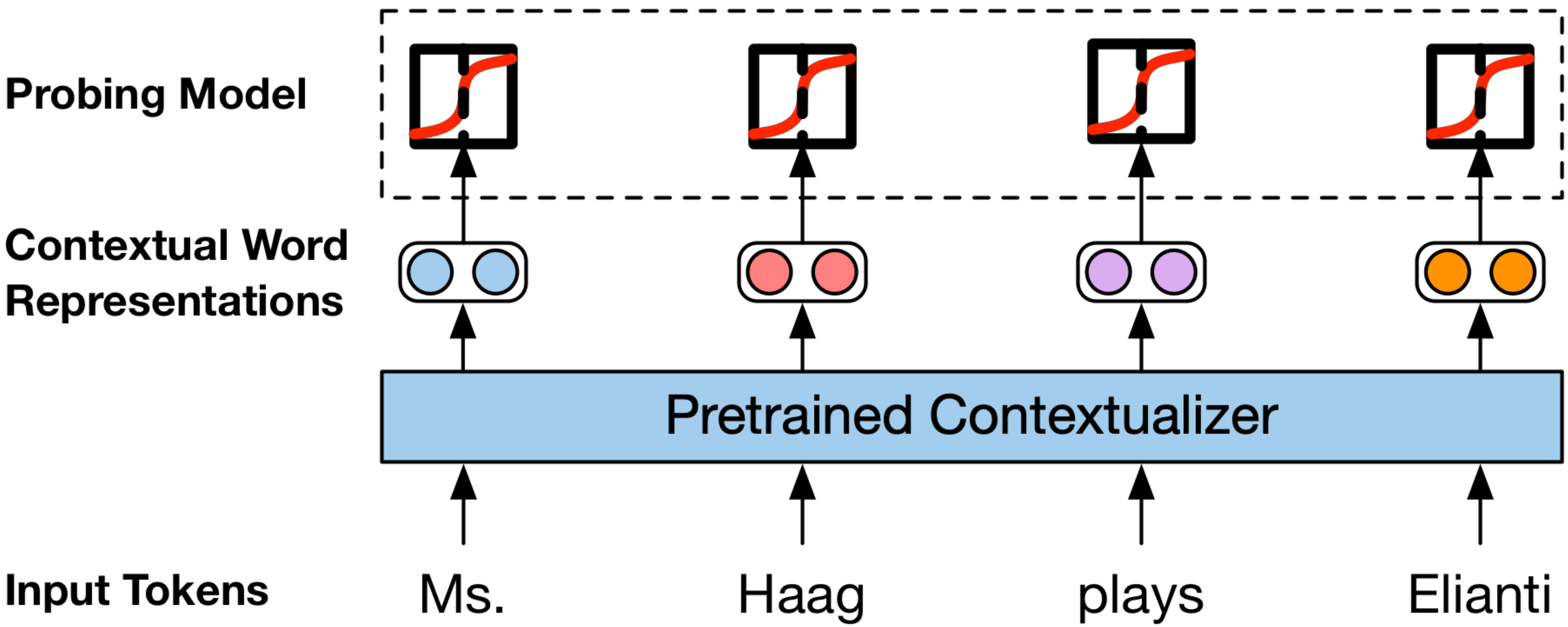
[Shi et al., 2016; Adi et al., 2017]

Probing Models

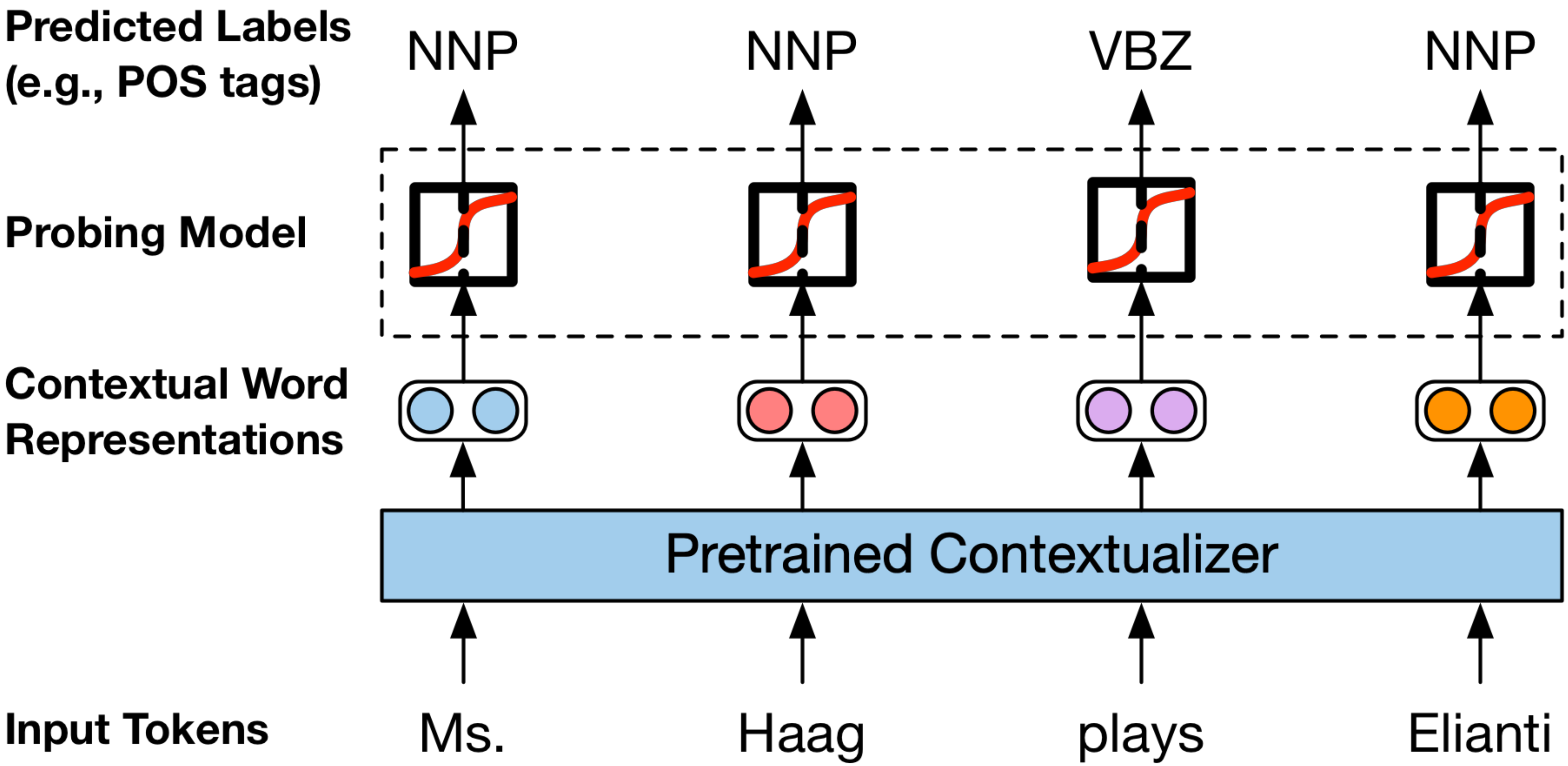


[Shi et al., 2016; Adi et al., 2017]

Probing Models



Probing Models



[Belinkov, 2018; Blevins et al., 2018; Tenney et al., 2019]

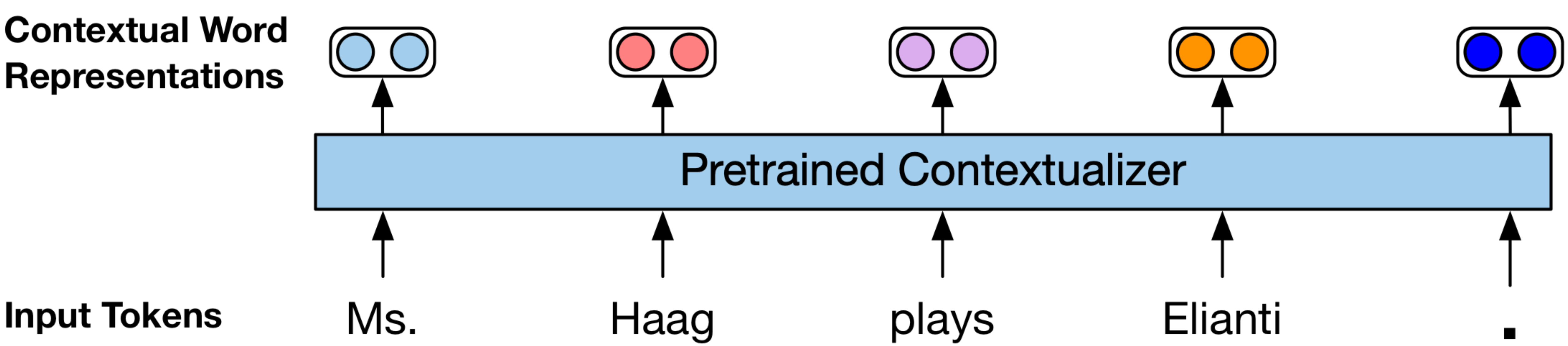
Pairwise Probing

[Belinkov, 2018; Blevins et al., 2018; Tenney et al., 2019]

Pairwise Probing

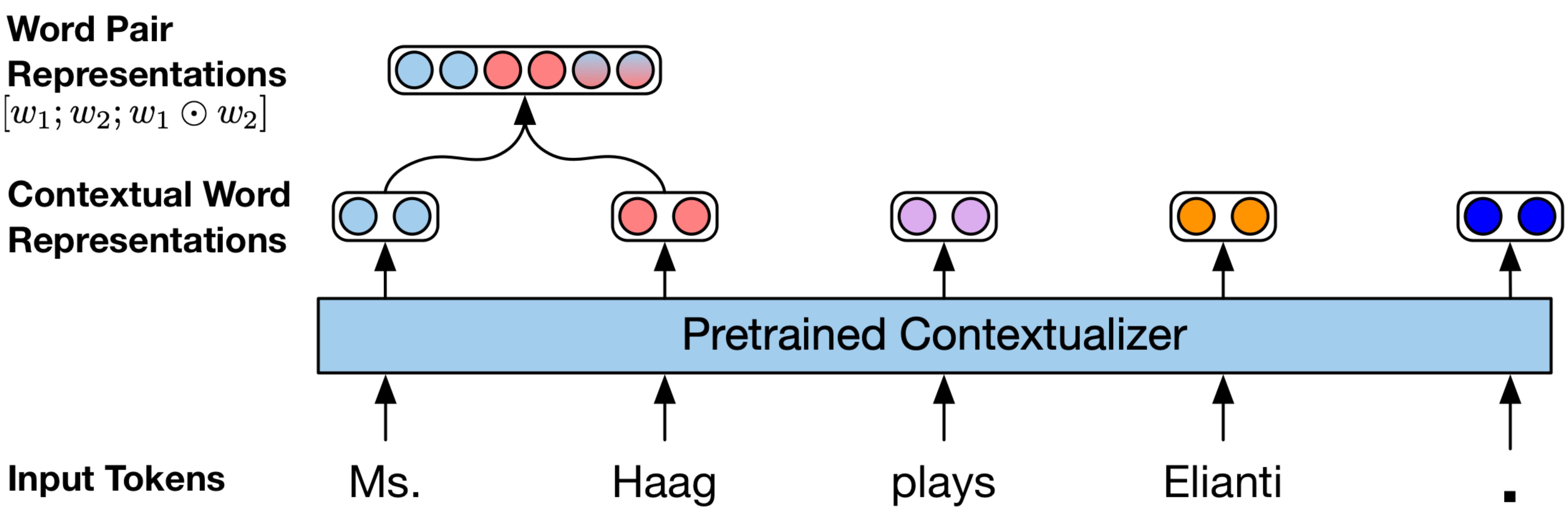
Input Tokens	Ms.	Haag	plays	Elianti	.

Pairwise Probing



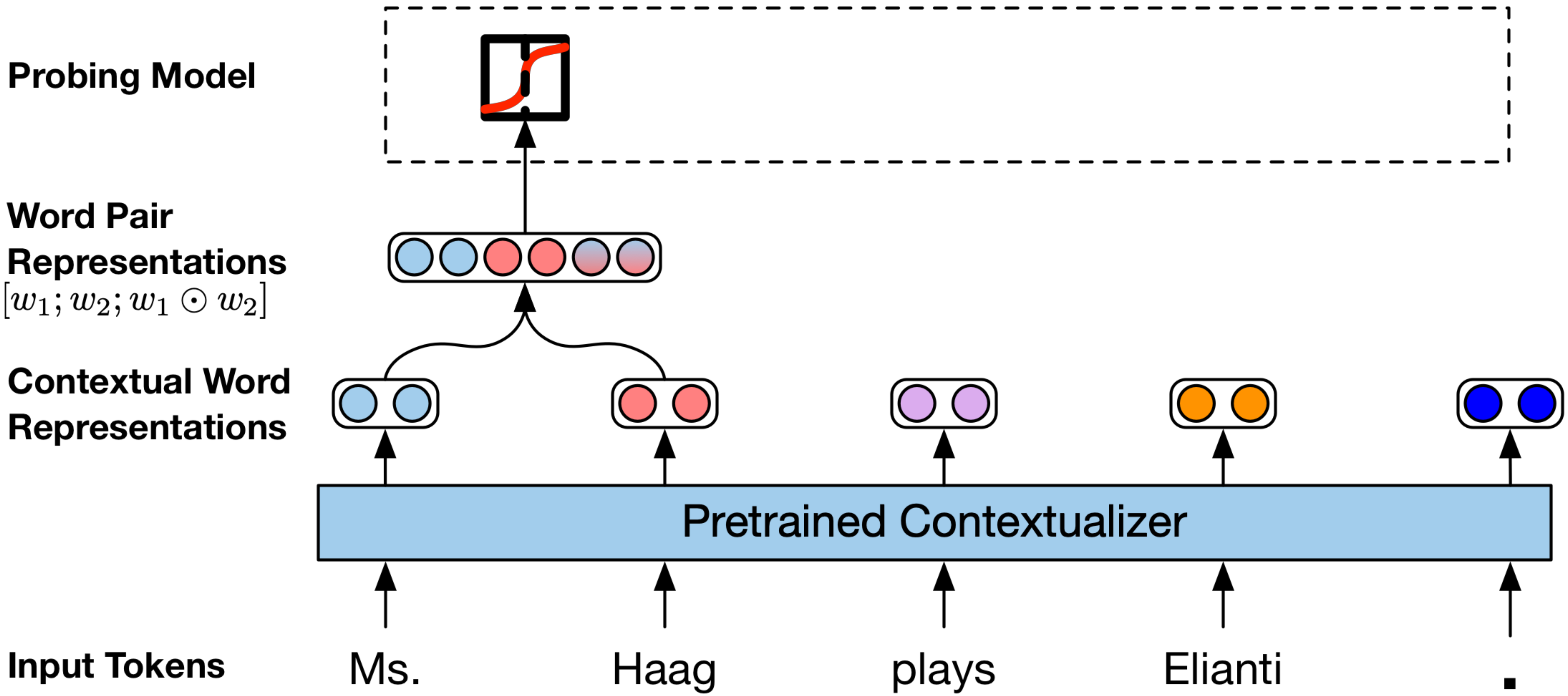
[Belinkov, 2018; Blevins et al., 2018; Tenney et al., 2019]

Pairwise Probing



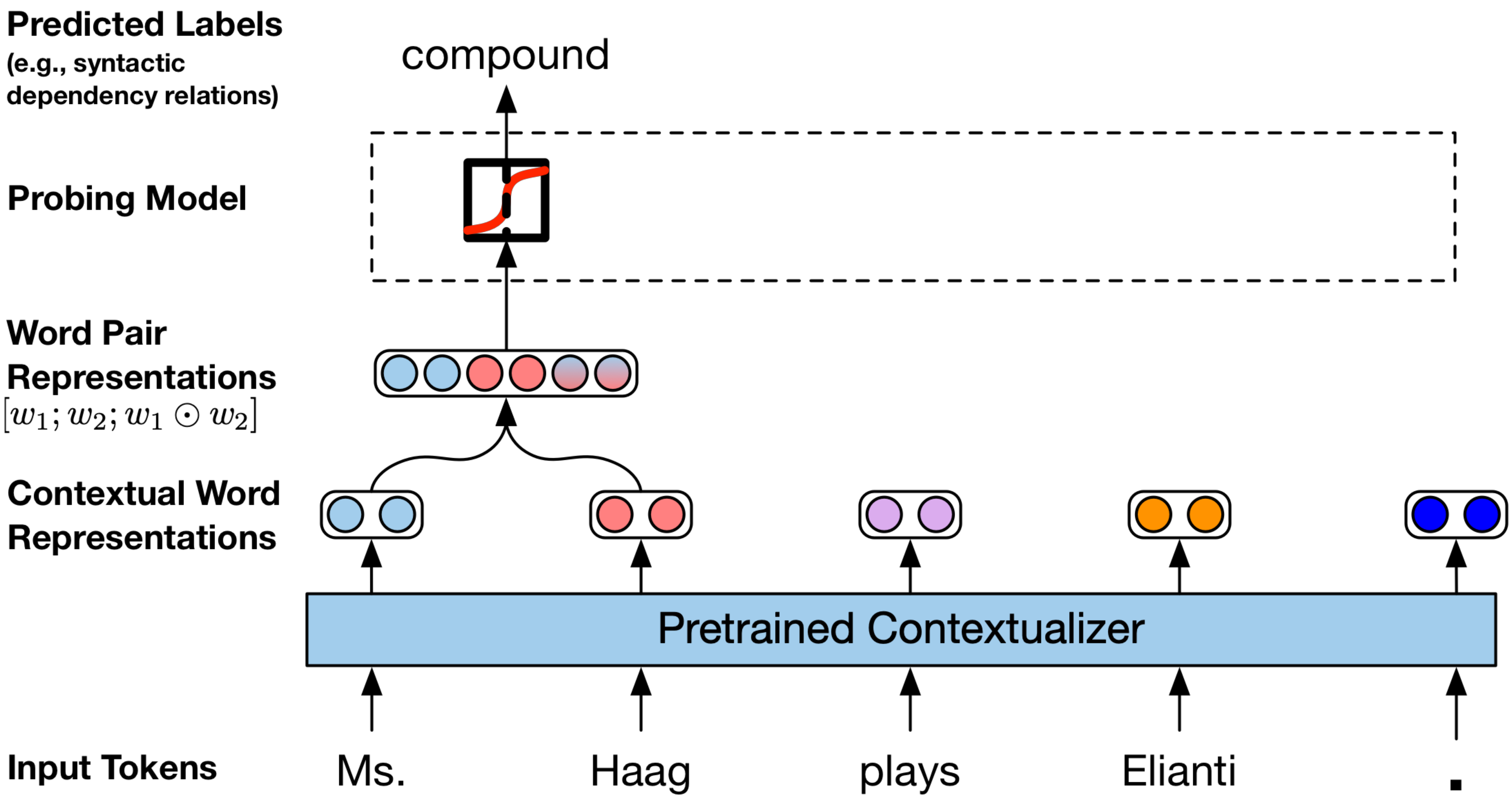
[Belinkov, 2018; Blevins et al., 2018; Tenney et al., 2019]

Pairwise Probing



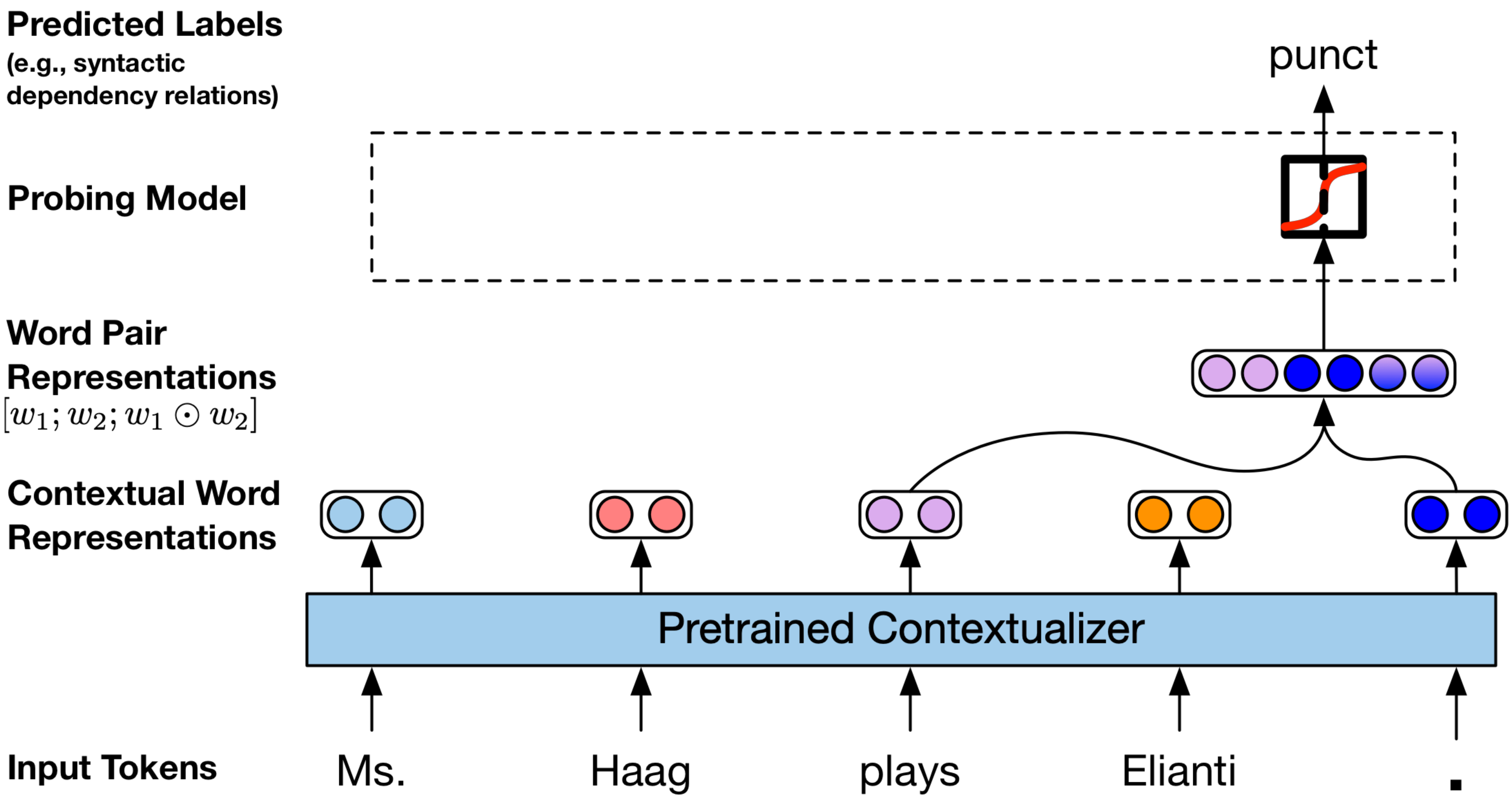
[Belinkov, 2018; Blevins et al., 2018; Tenney et al., 2019]

Pairwise Probing



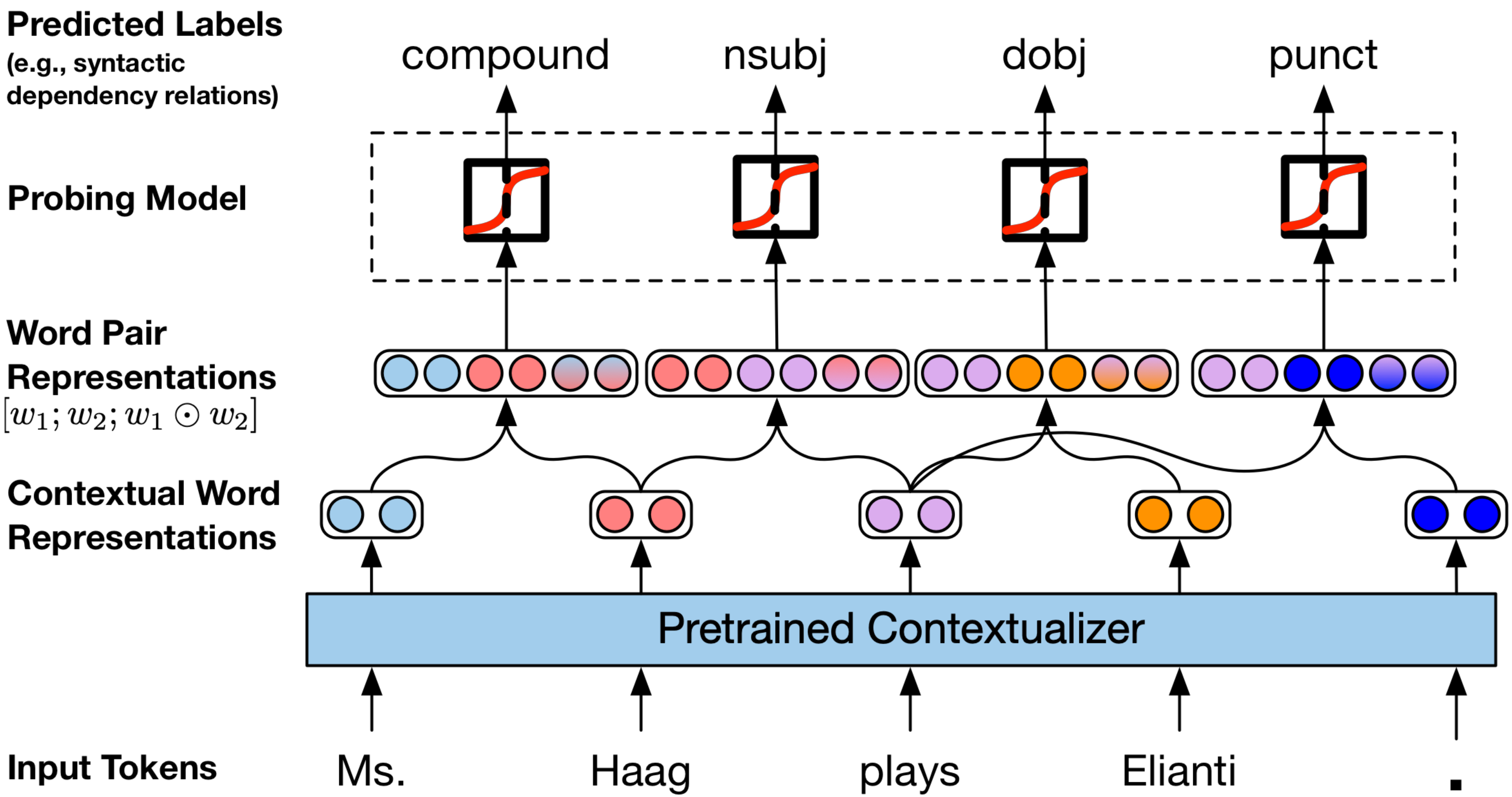
[Belinkov, 2018; Blevins et al., 2018; Tenney et al., 2019]

Pairwise Probing



[Belinkov, 2018; Blevins et al., 2018; Tenney et al., 2019]

Pairwise Probing



Probing Model Setup

- Contextualizer weights are always frozen.
- Results are from the highest-performing contextualizer layer.
- We use a linear probing model.

Contextualizers Analyzed

[Peters et al., 2018a,b]

Contextualizers Analyzed

ELMo

Bidirectional language model (BiLM) pretraining on 1B Word Benchmark

2-layer LSTM (ELMo original)

4-layer LSTM (ELMo 4-layer)

6-layer Transformer (ELMo transformer)

[Peters et al., 2018a,b; Radford et al., 2018]

Contextualizers Analyzed

ELMo

Bidirectional language model (BiLM) pretraining on 1B Word Benchmark

- 2-layer LSTM (ELMo original)
- 4-layer LSTM (ELMo 4-layer)
- 6-layer Transformer (ELMo transformer)

OpenAI Transformer

Left-to-right language model pretraining on uncased BookCorpus

- 12-layer transformer

[Peters et al., 2018a,b; Radford et al., 2018; Devlin et al., 2019]

Contextualizers Analyzed

ELMo
Bidirectional language model (BiLM) pretraining on 1B Word Benchmark

2-layer LSTM (ELMo original)

4-layer LSTM (ELMo 4-layer)

6-layer Transformer (ELMo transformer)

OpenAI Transformer
Left-to-right language model pretraining on uncased BookCorpus

12-layer transformer

BERT (cased)
Masked language model pretraining on BookCorpus + Wikipedia

12-layer transformer (BERT base)

24-layer transformer (BERT large)

(1) Probing Contextual Representations

Question: Is the information necessary for a variety of core NLP tasks linearly recoverable from contextual word representations?

Answer: Yes, to a great extent! Tasks with lower performance may require fine-grained linguistic knowledge.

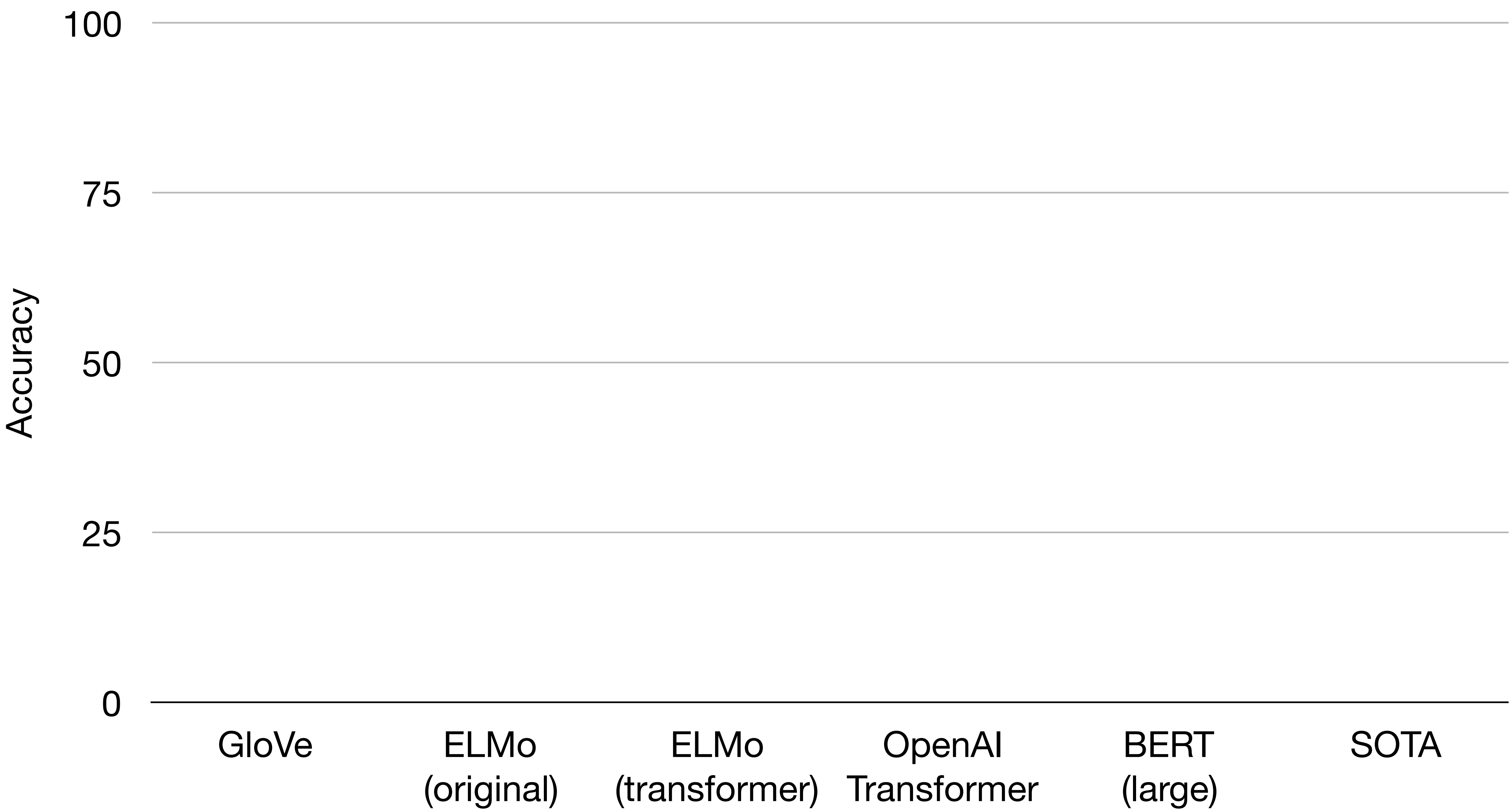
Examined 17 Diverse Probing Tasks

- Part-of-Speech Tagging
- CCG Supertagging
- Semantic Tagging
- Preposition supersense disambiguation
- Event Factuality
- Syntactic Constituency Ancestor Tagging
- Syntactic Chunking
- Named entity recognition
- Grammatical error detection
- Conjunct identification
- Syntactic Dependency Arc Prediction
- Syntactic Dependency Arc Classification
- Semantic Dependency Arc Prediction
- Semantic Dependency Arc Classification
- Coreference Arc Prediction

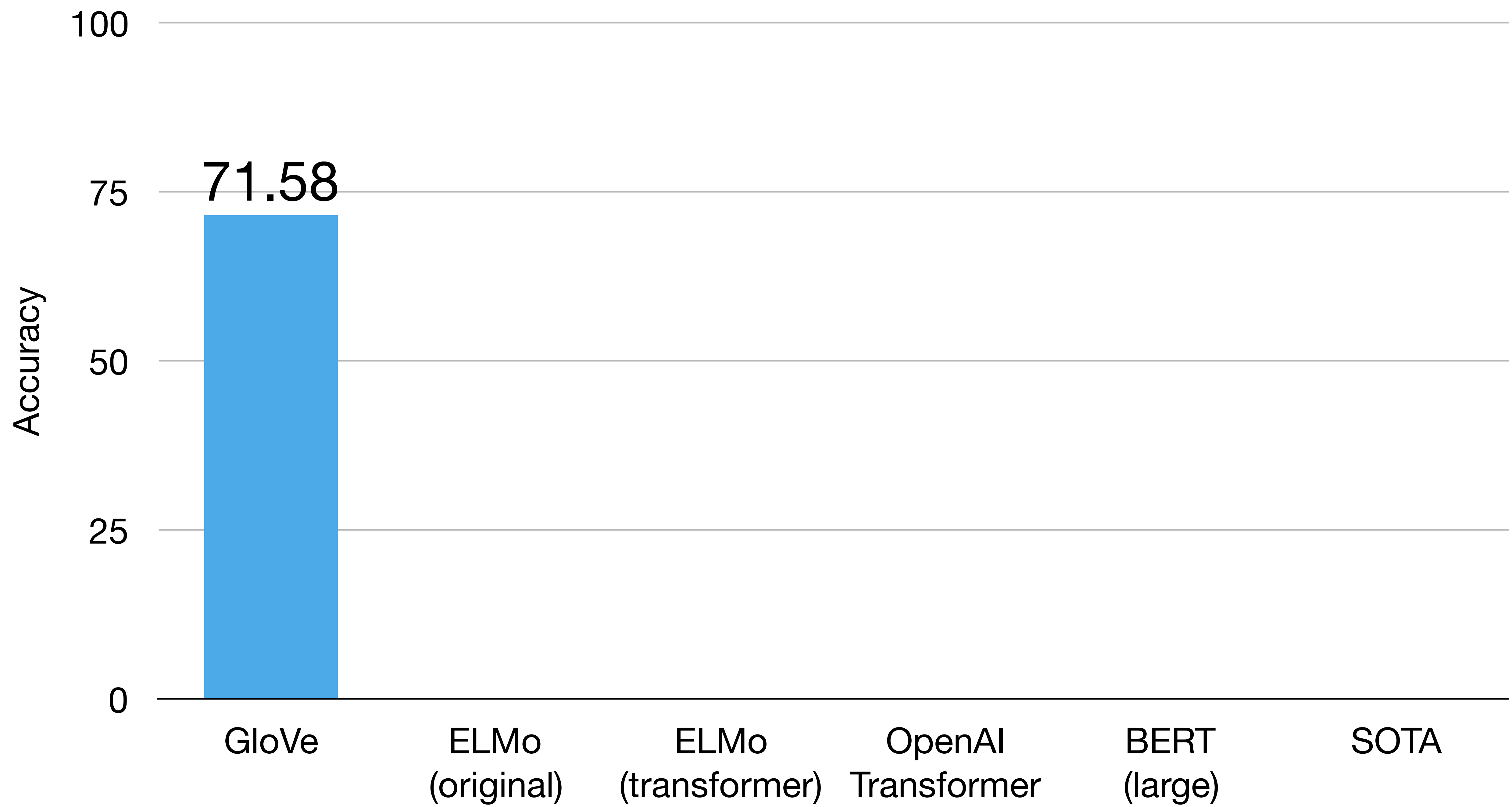
Linear Probing Models Rival Task-Specific Architectures

- Part-of-Speech Tagging
- CCG Supertagging
- Semantic Tagging
- Preposition supersense disambiguation
- Event Factuality
- Syntactic Constituency Ancestor Tagging
- Syntactic Chunking
- Named entity recognition
- Grammatical error detection
- Conjunct identification
- Syntactic Dependency Arc Prediction
- Syntactic Dependency Arc Classification
- Semantic Dependency Arc Prediction
- Semantic Dependency Arc Classification
- Coreference Arc Prediction

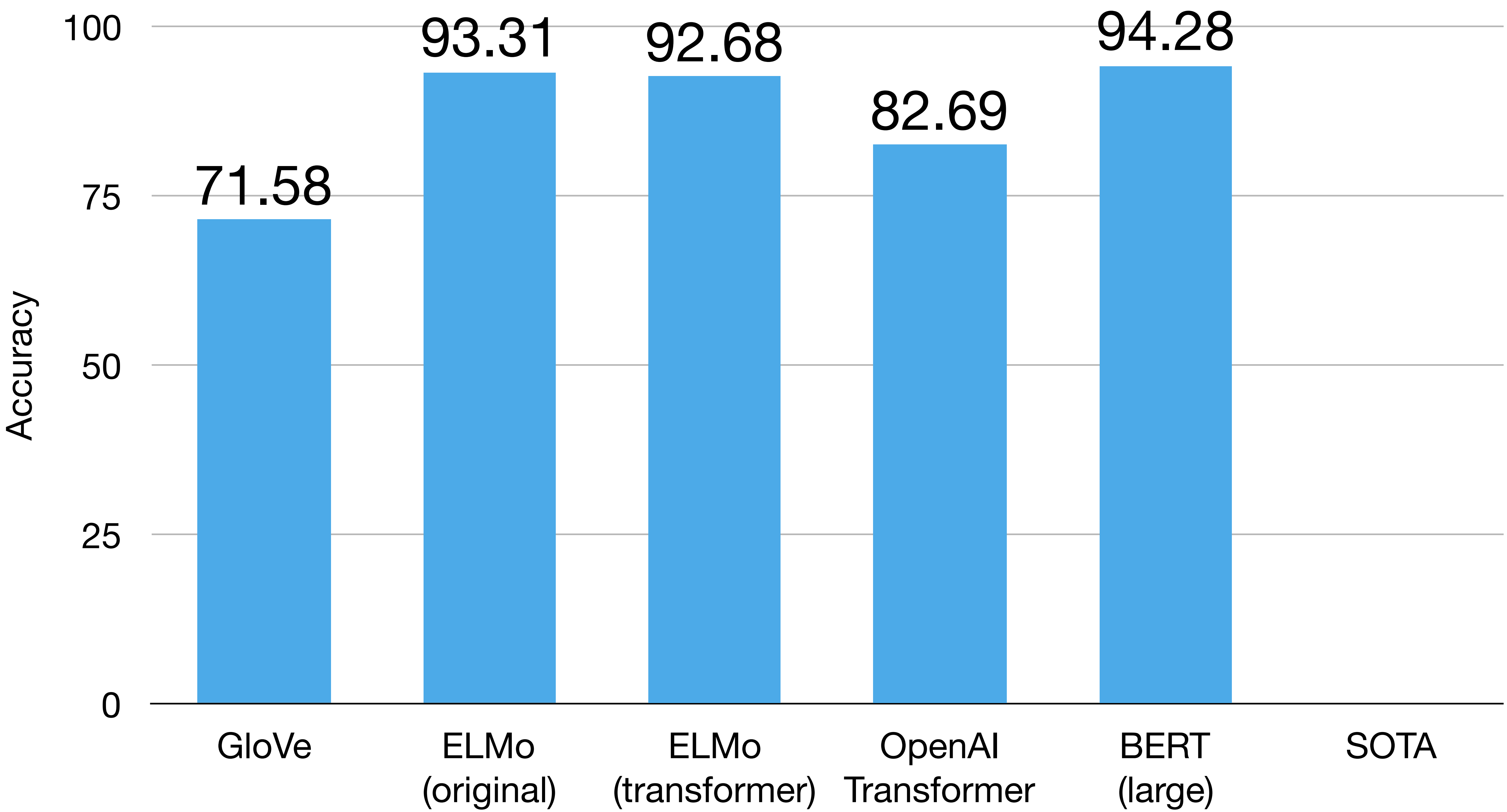
CCG Supertagging



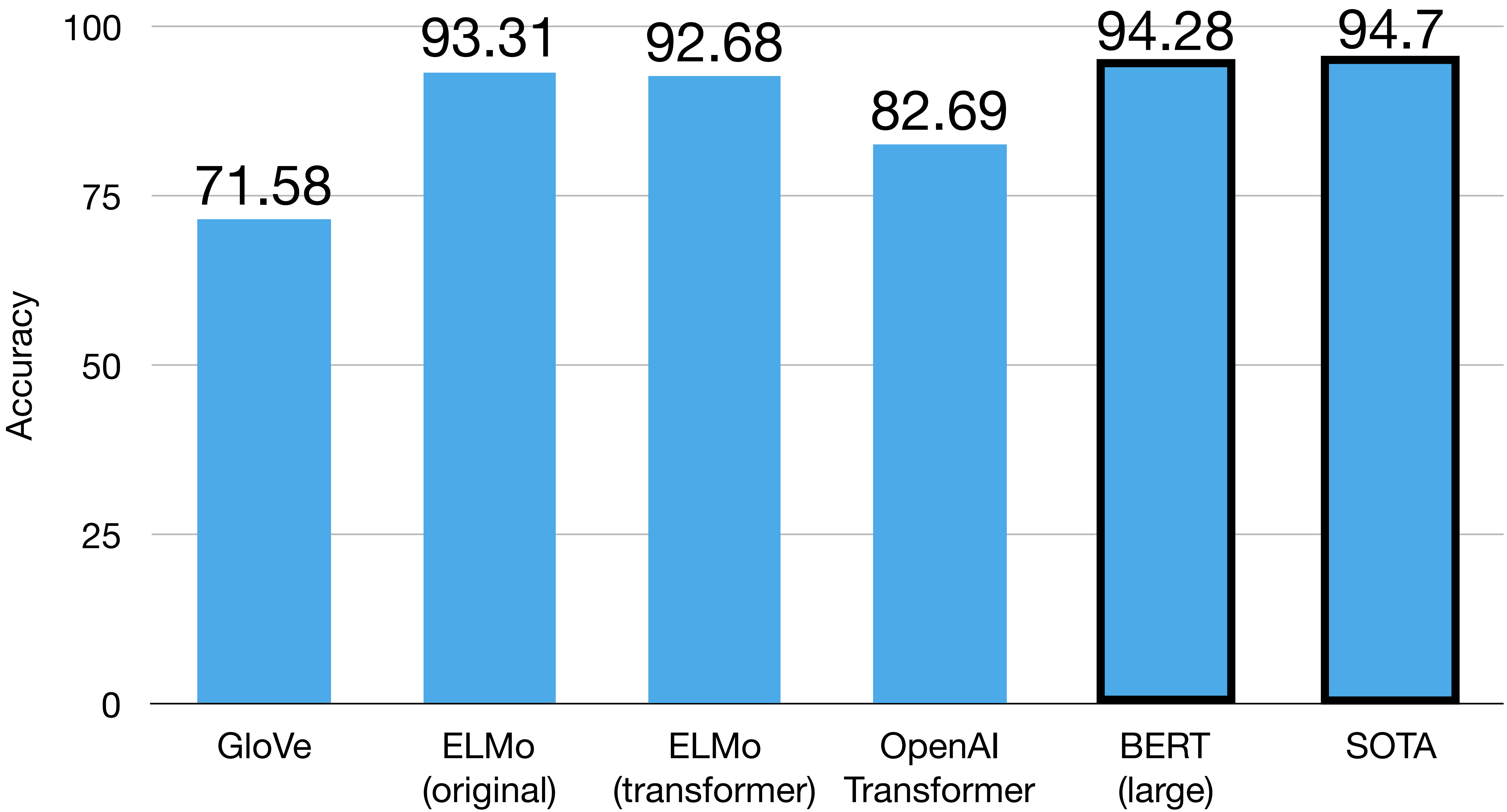
CCG Supertagging



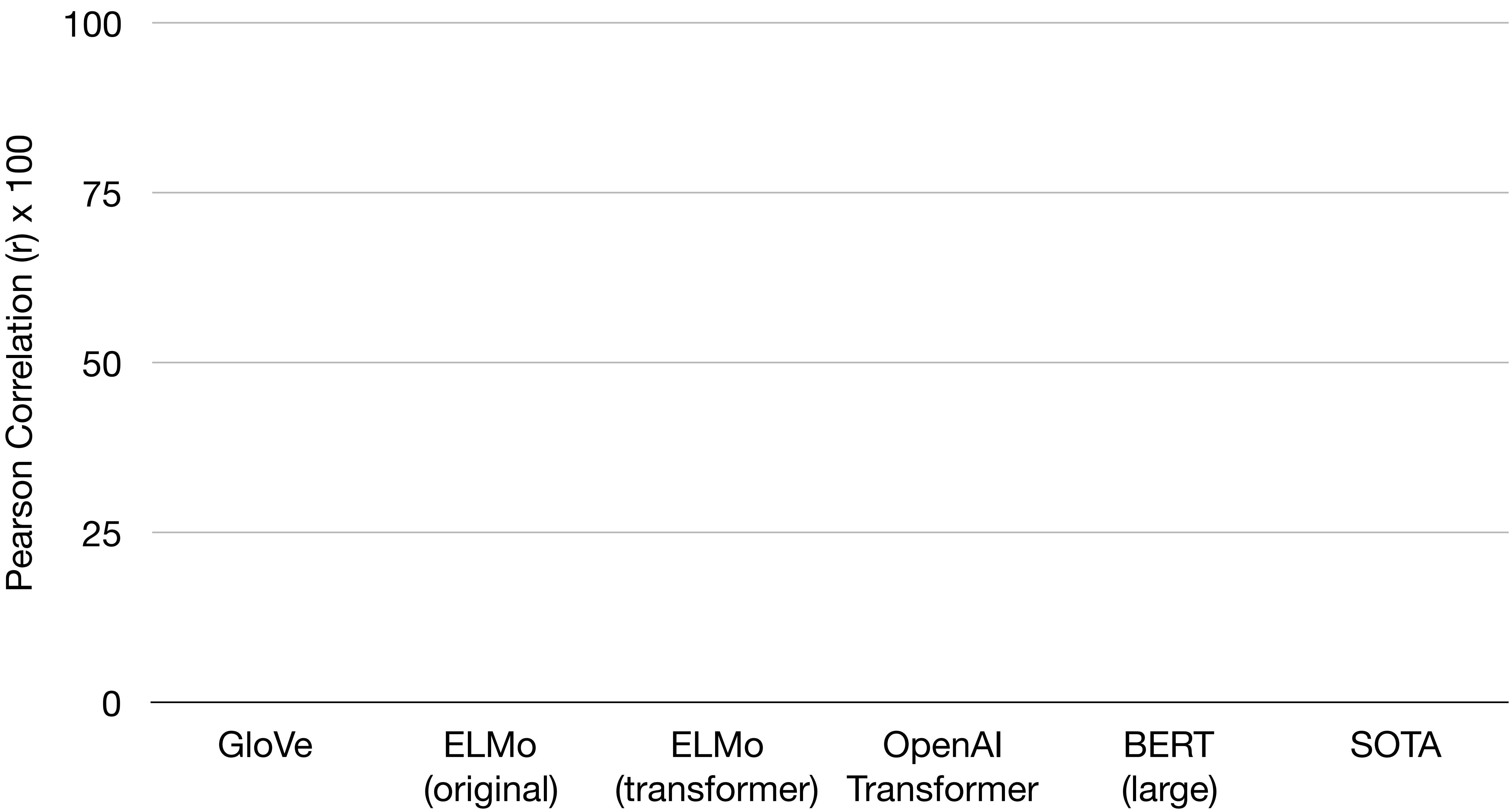
CCG Supertagging



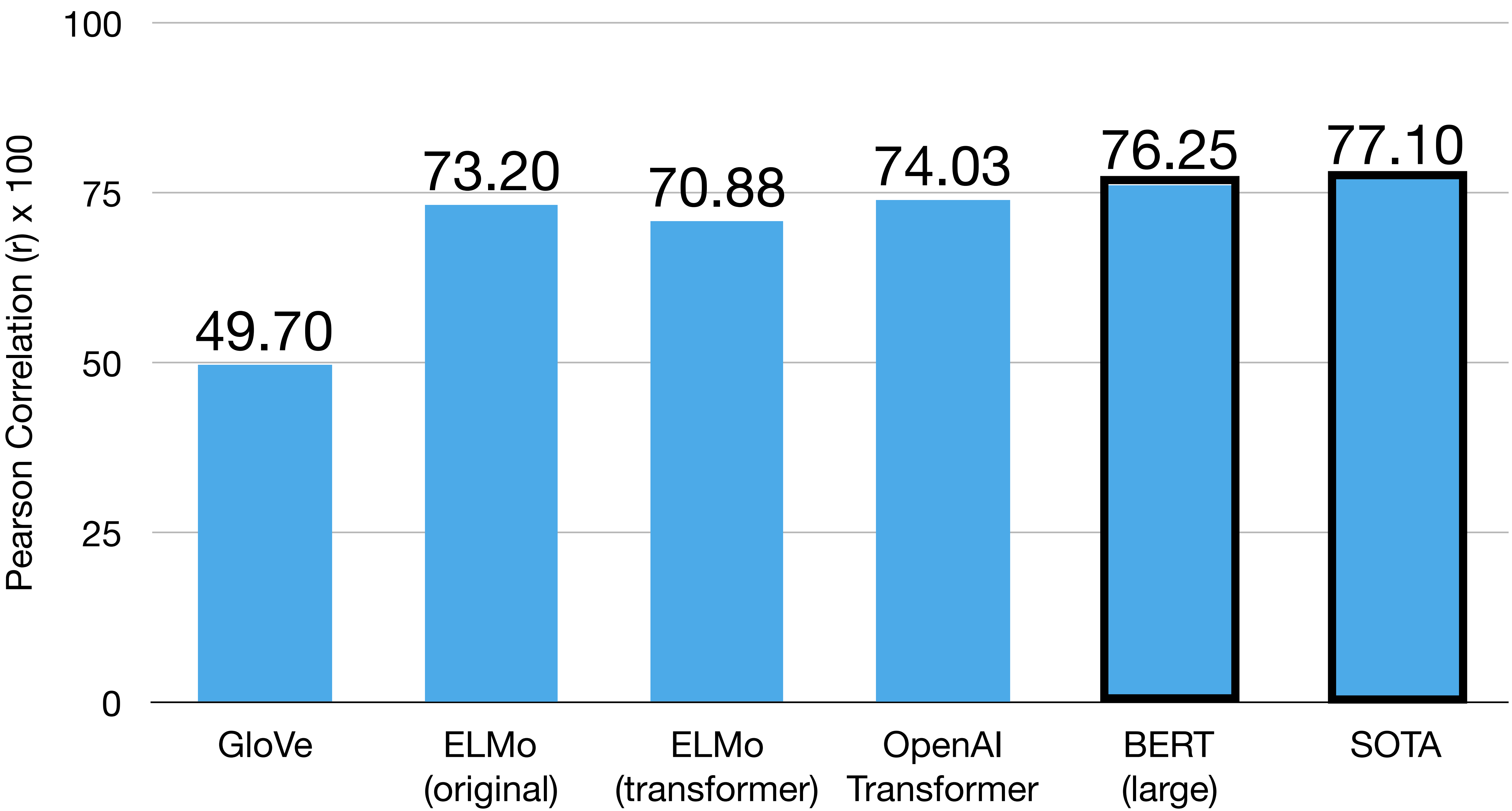
CCG Supertagging



Event Factuality



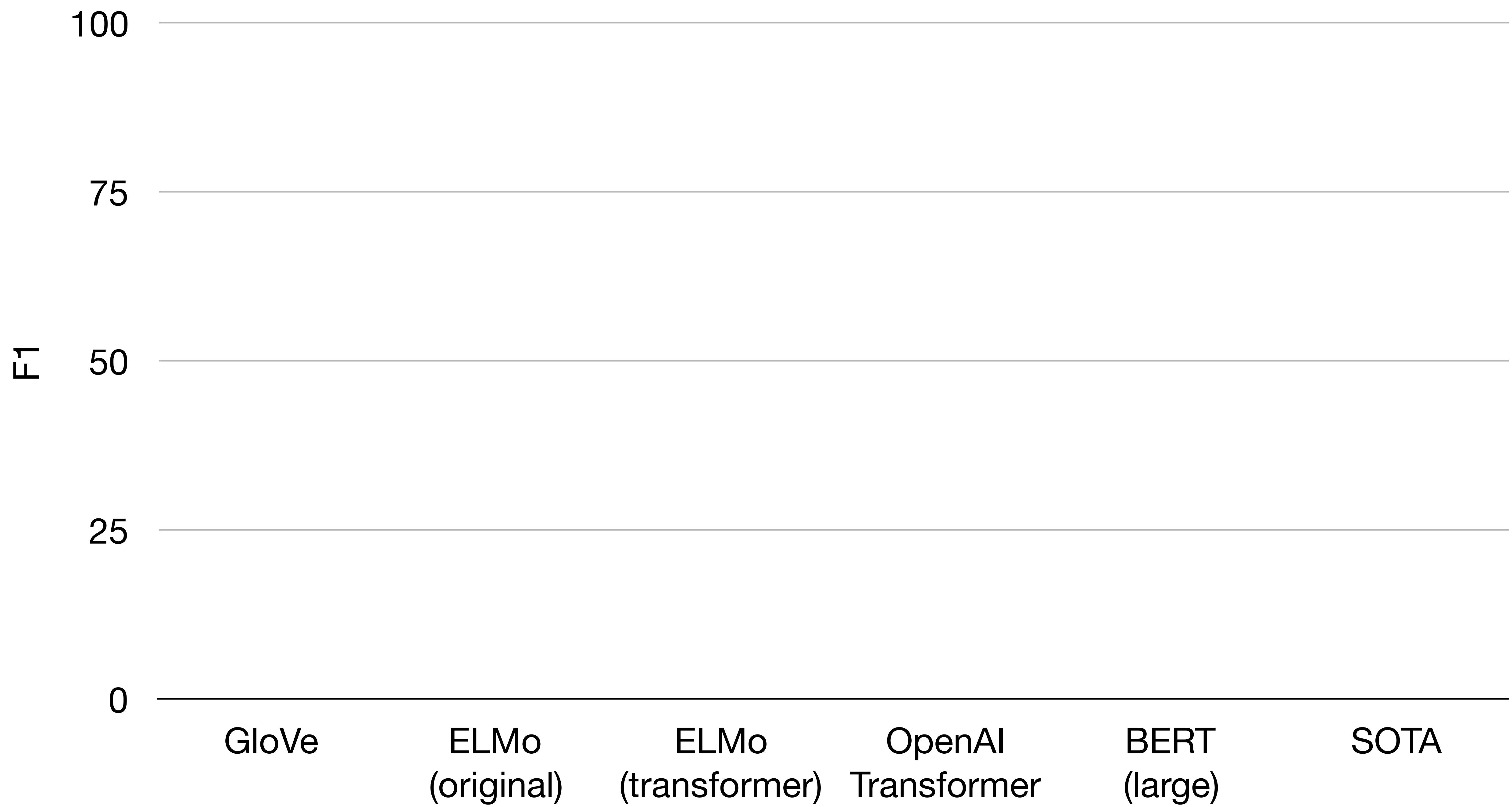
Event Factuality



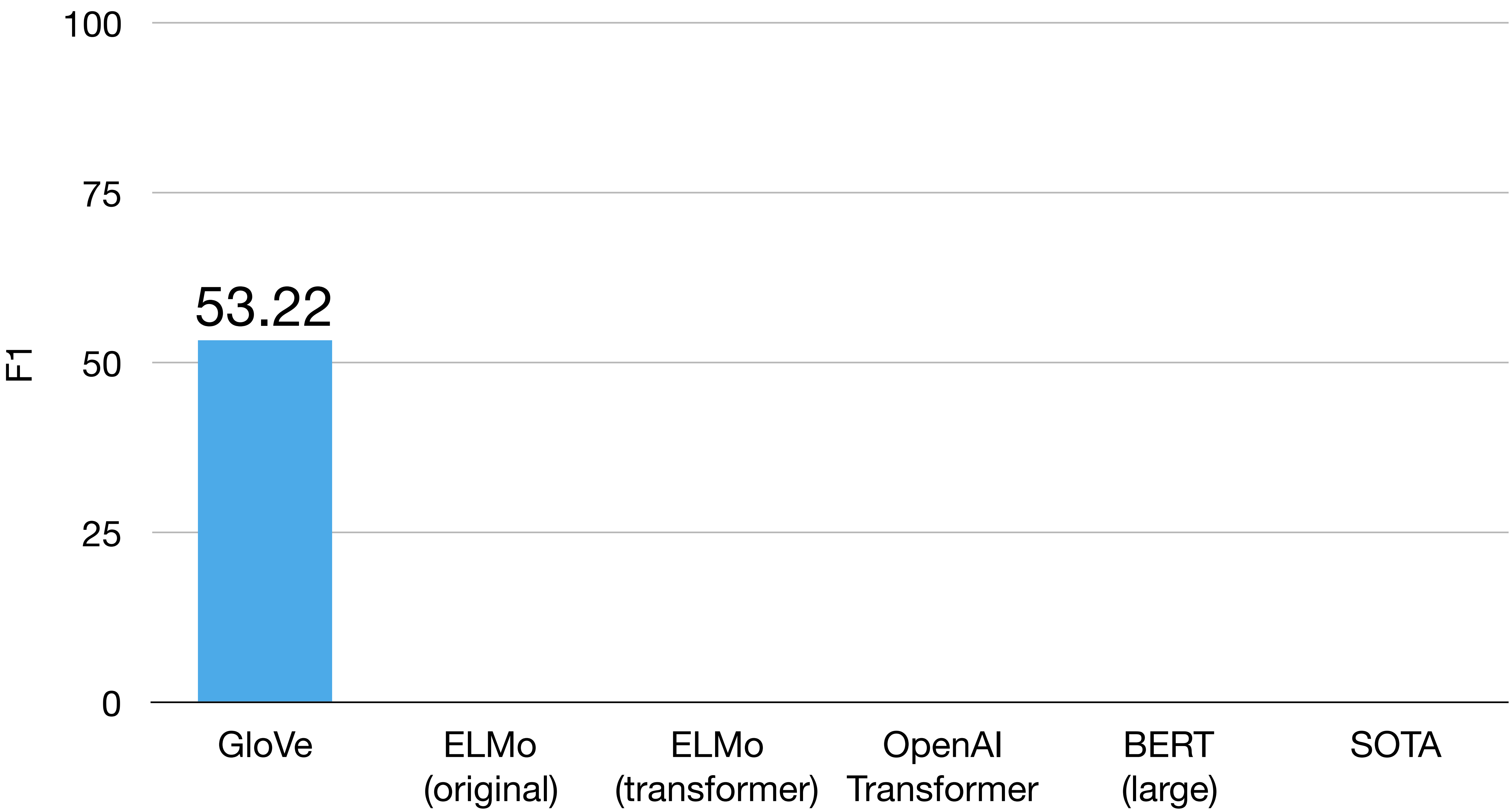
But Linear Probing Models Underperform on Some Tasks

- Tasks that linear model + contextual word representation performs poorly may require more fine-grained linguistic knowledge.
- In these cases, task-specific contextualization leads to especially large gains. See the paper for more details.

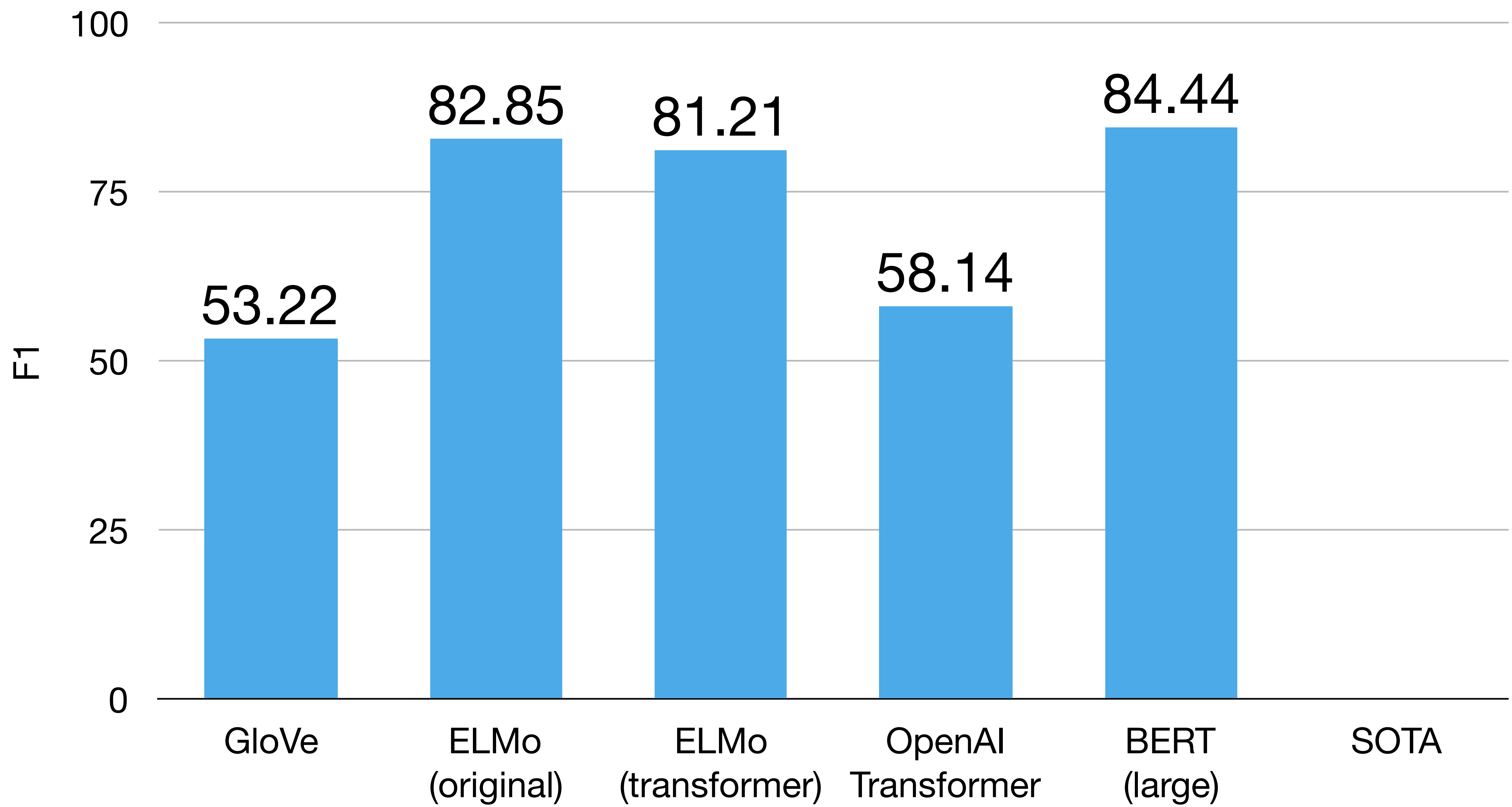
Named Entity Recognition



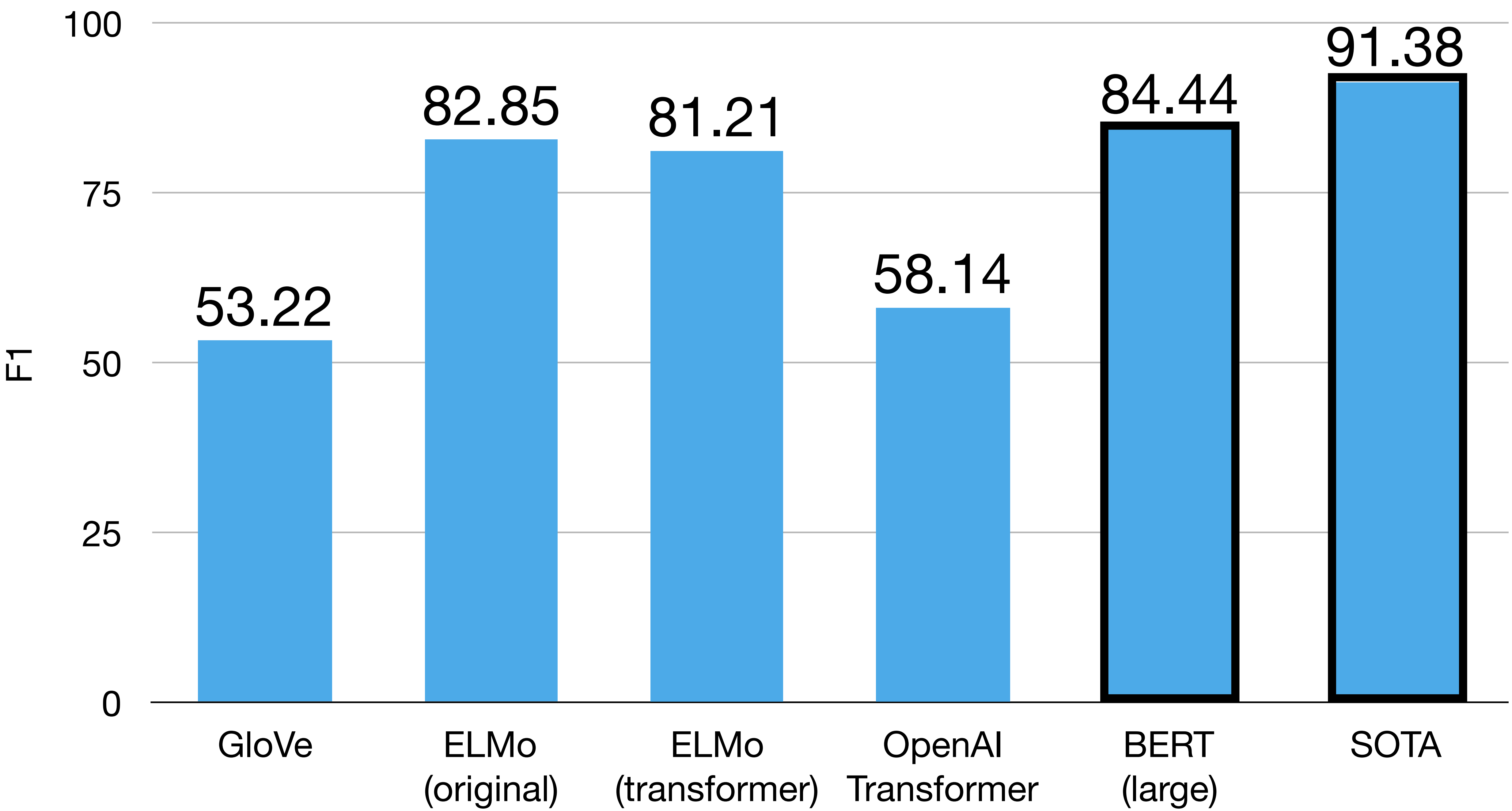
Named Entity Recognition



Named Entity Recognition



Named Entity Recognition



(2) How Does Transferability Vary?

Question: How does transferability vary across contextualizer layers?

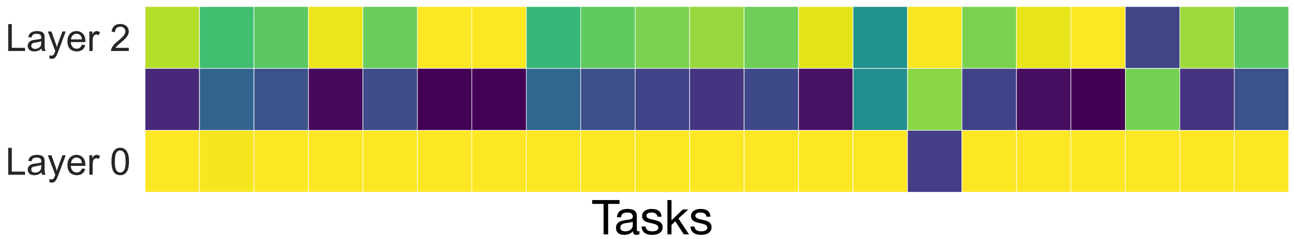
Answer: First layer in LSTMs is the most transferable. Middle layers for transformers.

Layerwise Patterns in Transferability

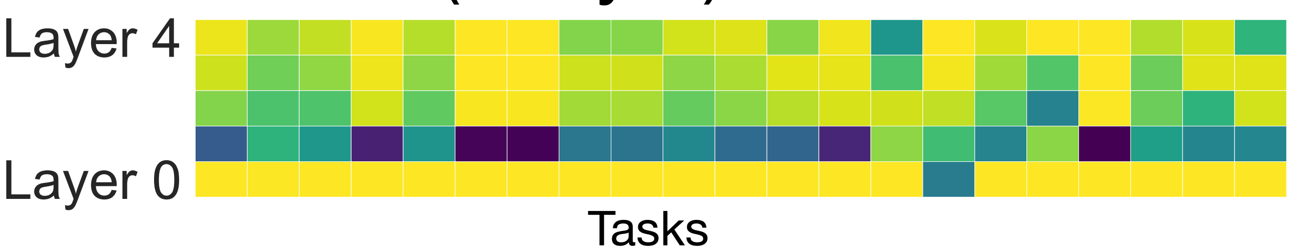
Layerwise Patterns in Transferability

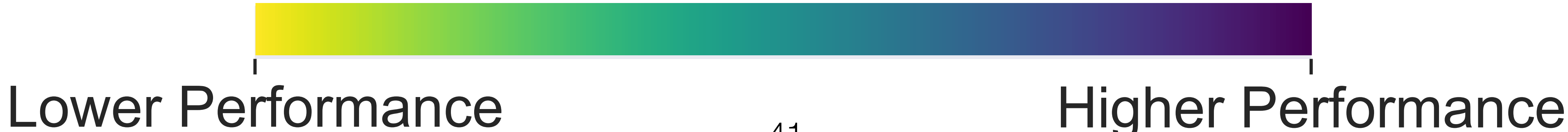
LSTM-based Contextualizers

ELMo (original)



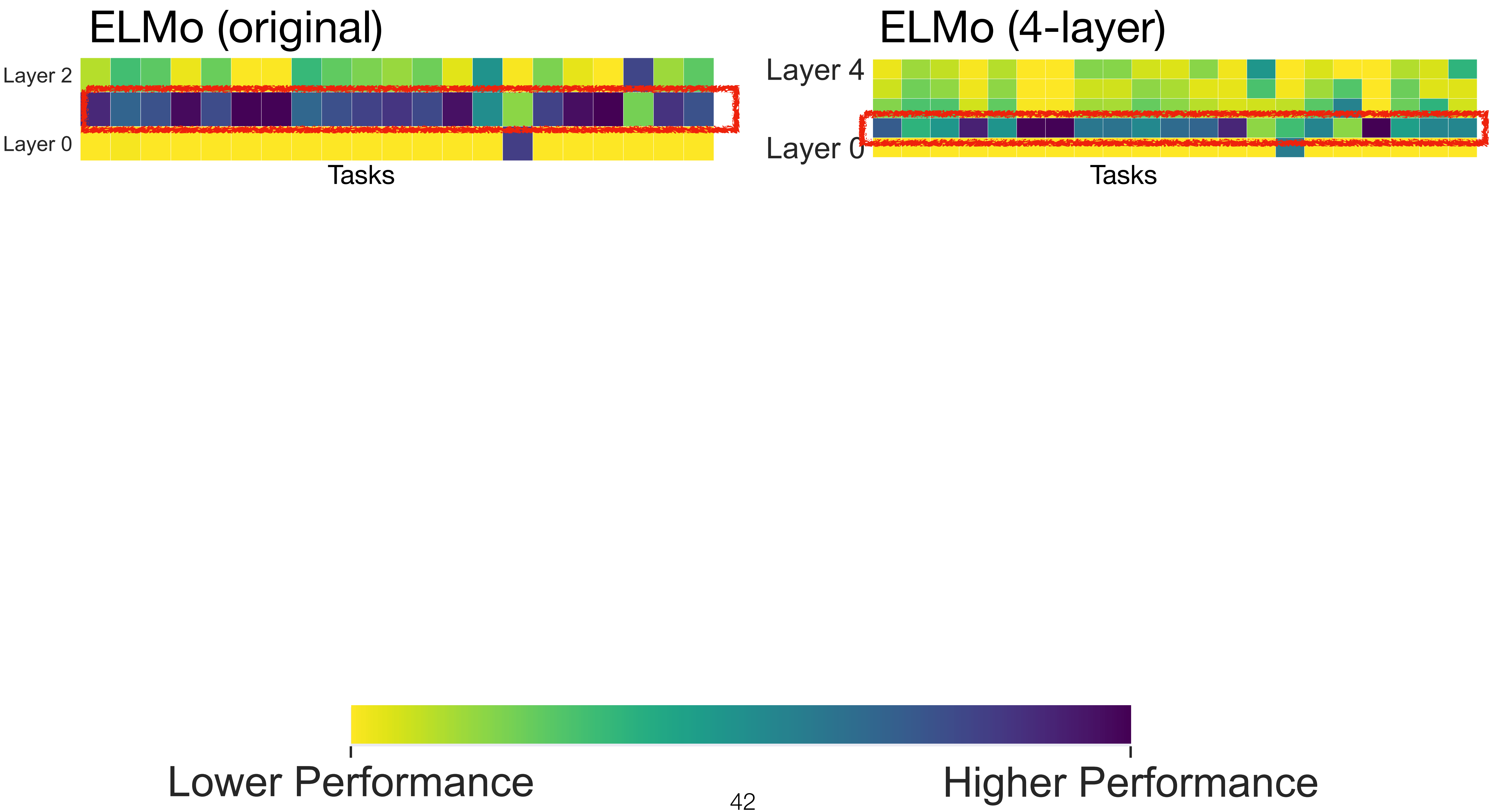
ELMo (4-layer)





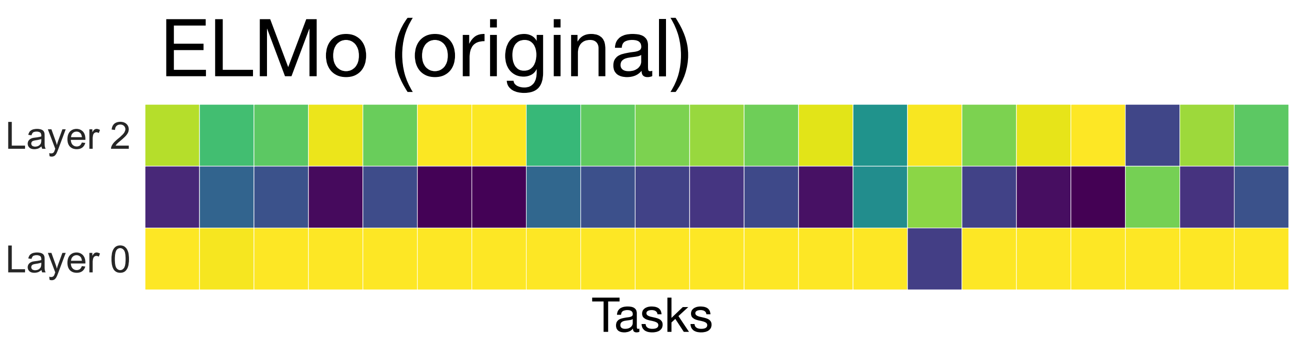
Layerwise Patterns in Transferability

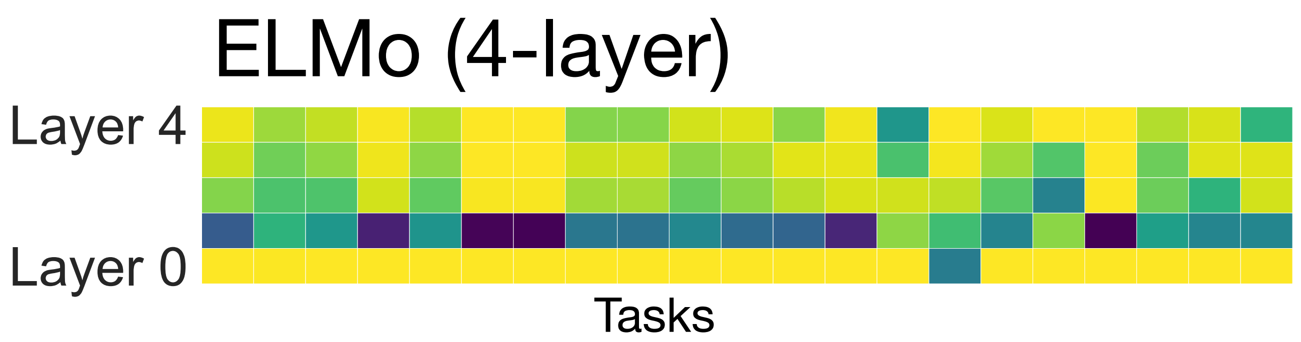
LSTM-based Contextualizers



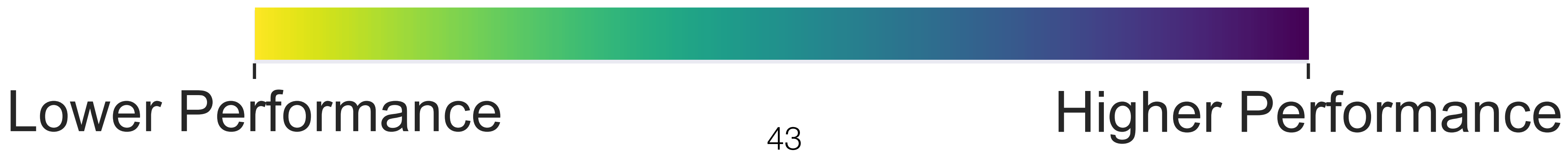
Layerwise Patterns in Transferability

LSTM-based Contextualizers





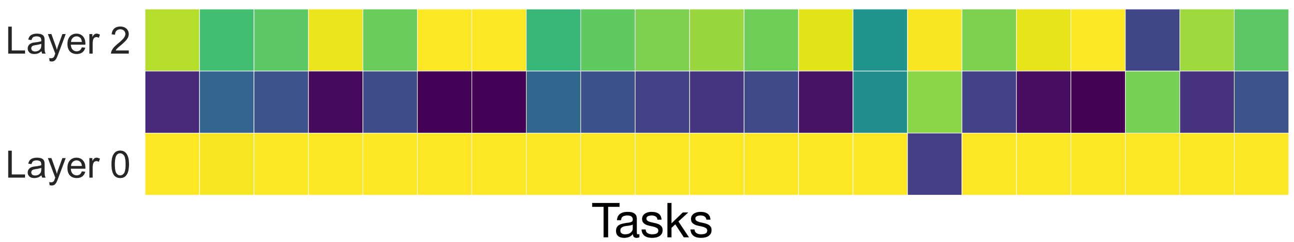
Transformer-based Contextualizers



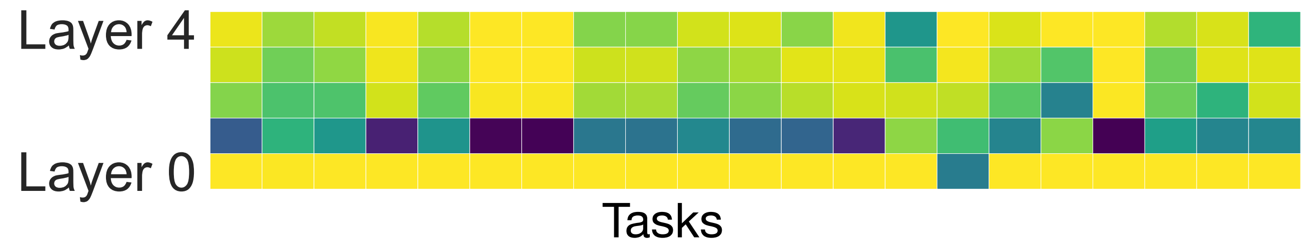
Layerwise Patterns in Transferability

LSTM-based Contextualizers

ELMo (original)

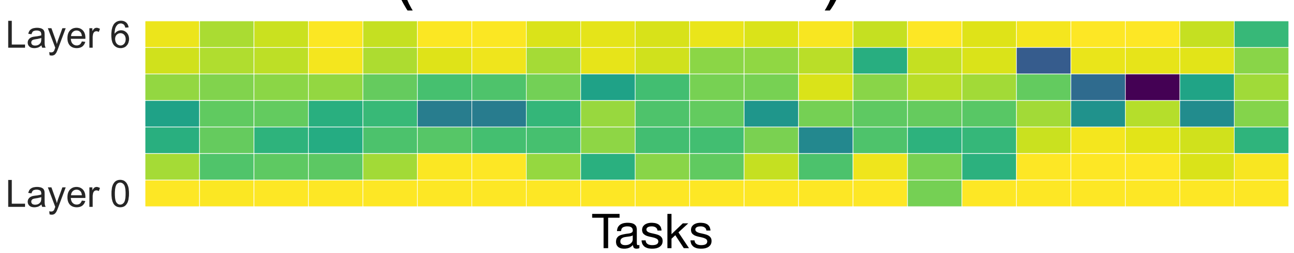


ELMo (4-layer)

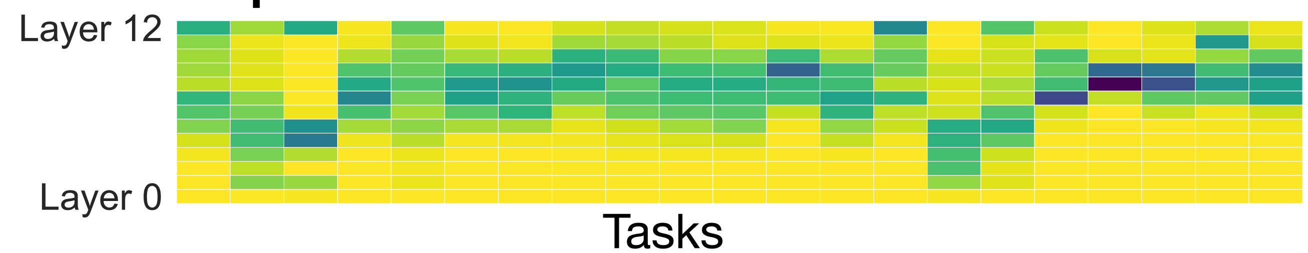


Transformer-based Contextualizers

ELMo (transformer)

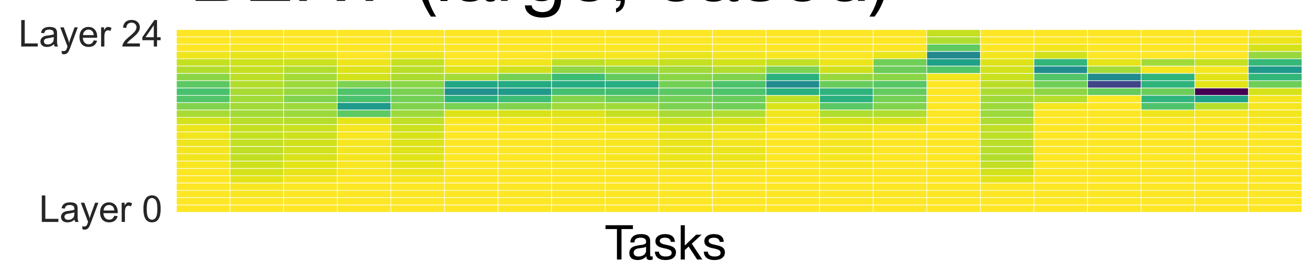


OpenAI Transformer



BERT (base, cased)

BERT (large, cased)



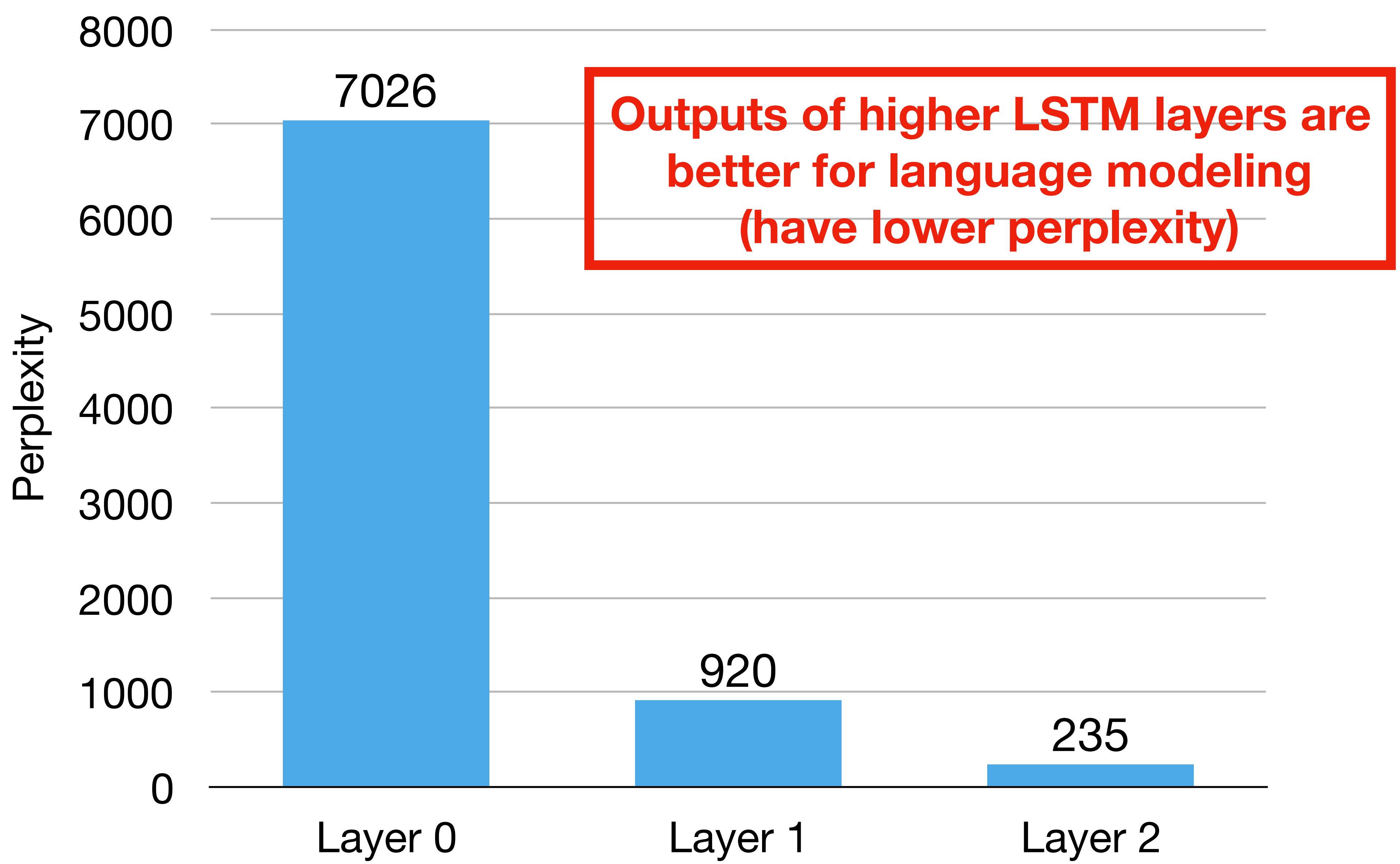
(3) Why Does Transferability Vary?

Question: **Why** does transferability vary across contextualizer layers?

Answer: It depends on pretraining task-specificity!

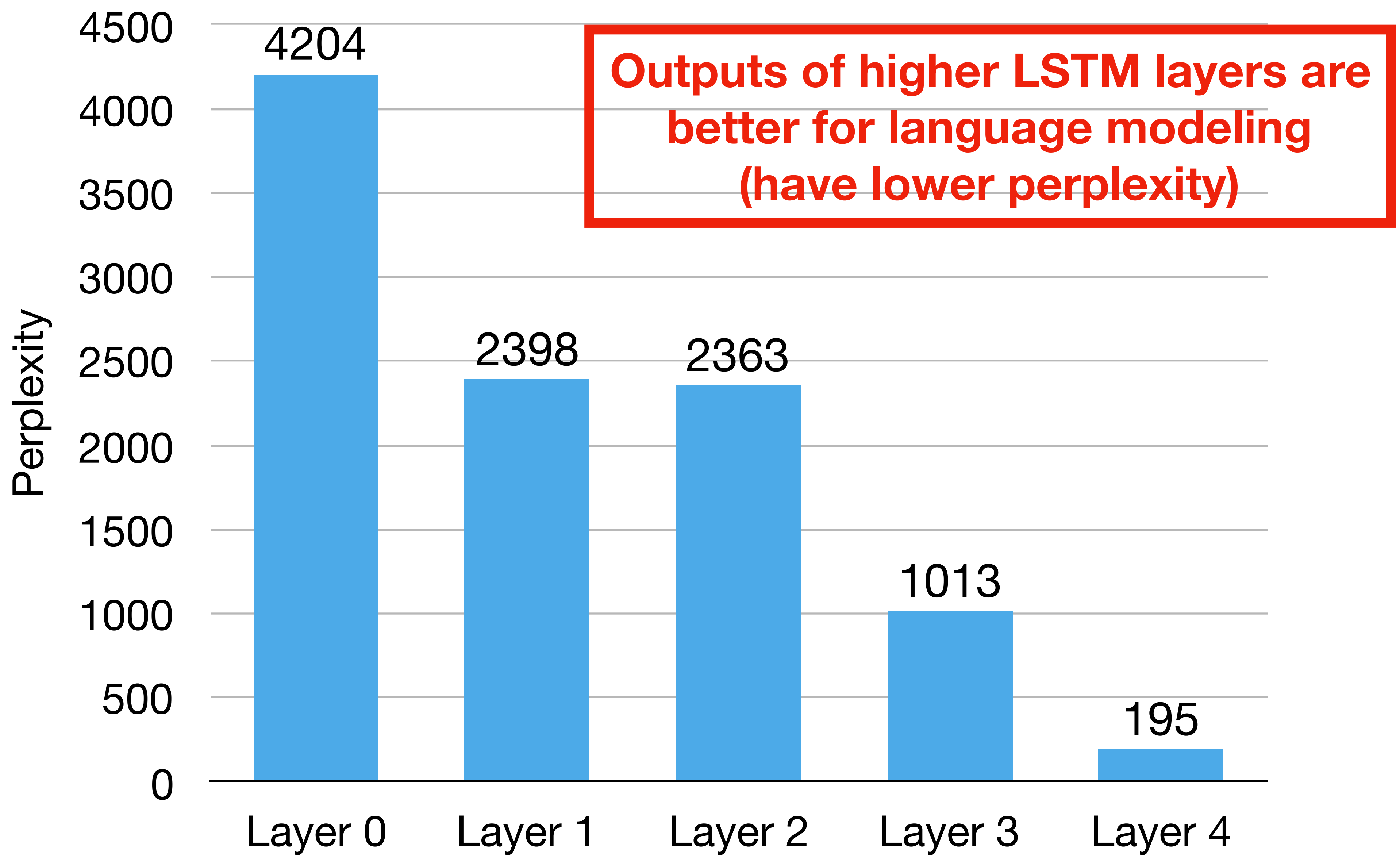
Layerwise Patterns Dictated by Perplexity

LSTM-based ELMo (original)



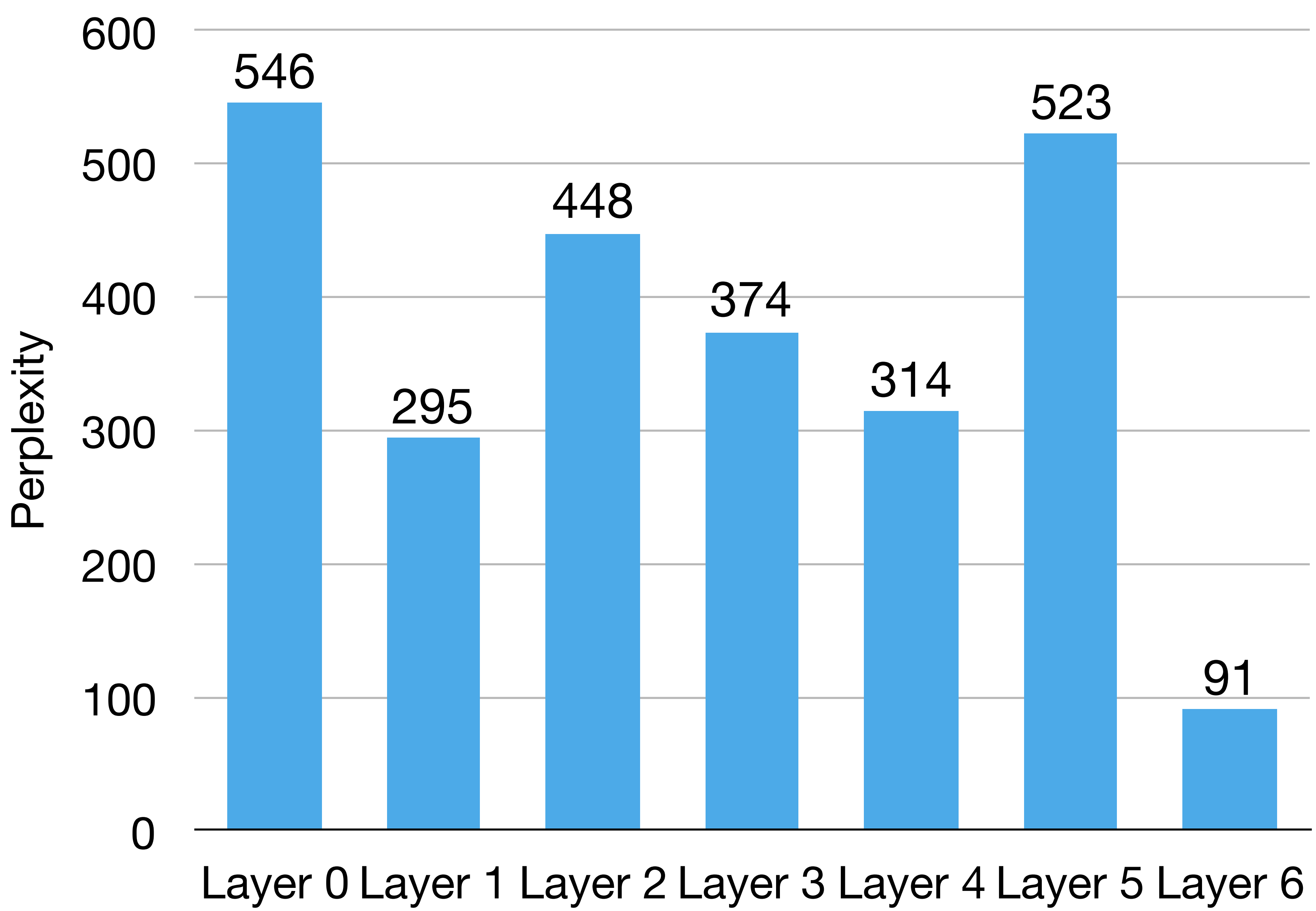
Layerwise Patterns Dictated by Perplexity

LSTM-based ELMo (4-layer)



Layerwise Patterns Dictated by Perplexity

Transformer-based ELMo (6-layer)



(4) Alternative Pretraining Objectives

Question: How does language model pretraining compare to alternatives?

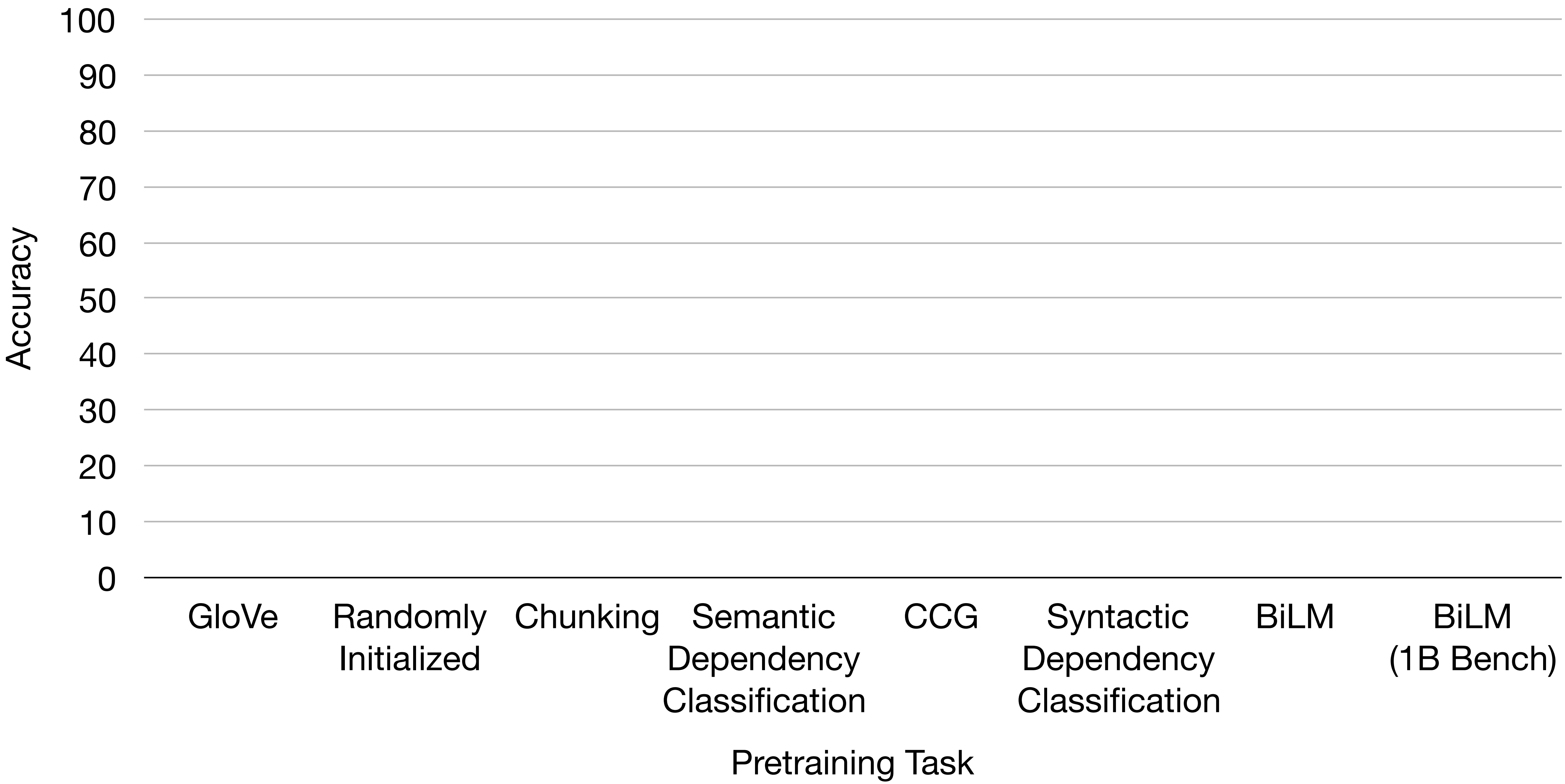
Answer: Even with 1 million tokens, language model pretraining yields the most transferable representations.

But, transferring between related tasks does help.

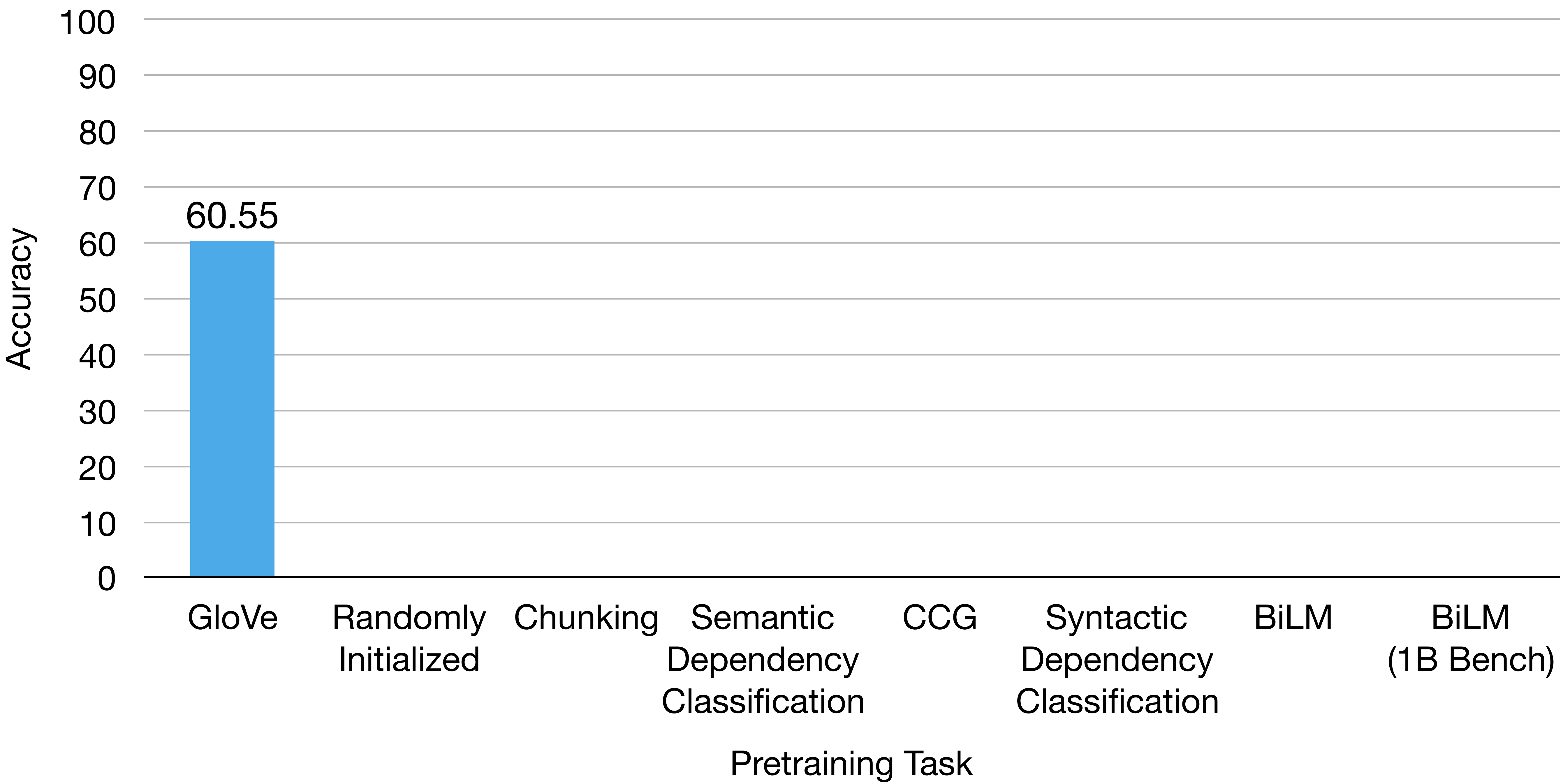
Investigating Alternatives to Language Model Pretraining

- How does the language modeling as a pretraining objective compare to explicitly supervised tasks?
- Pretrain ELMo (original)-architecture contextualizer on the Penn Treebank, with a variety of different objectives.
- Evaluate how well the resultant representations transfer to target (held-out) tasks.

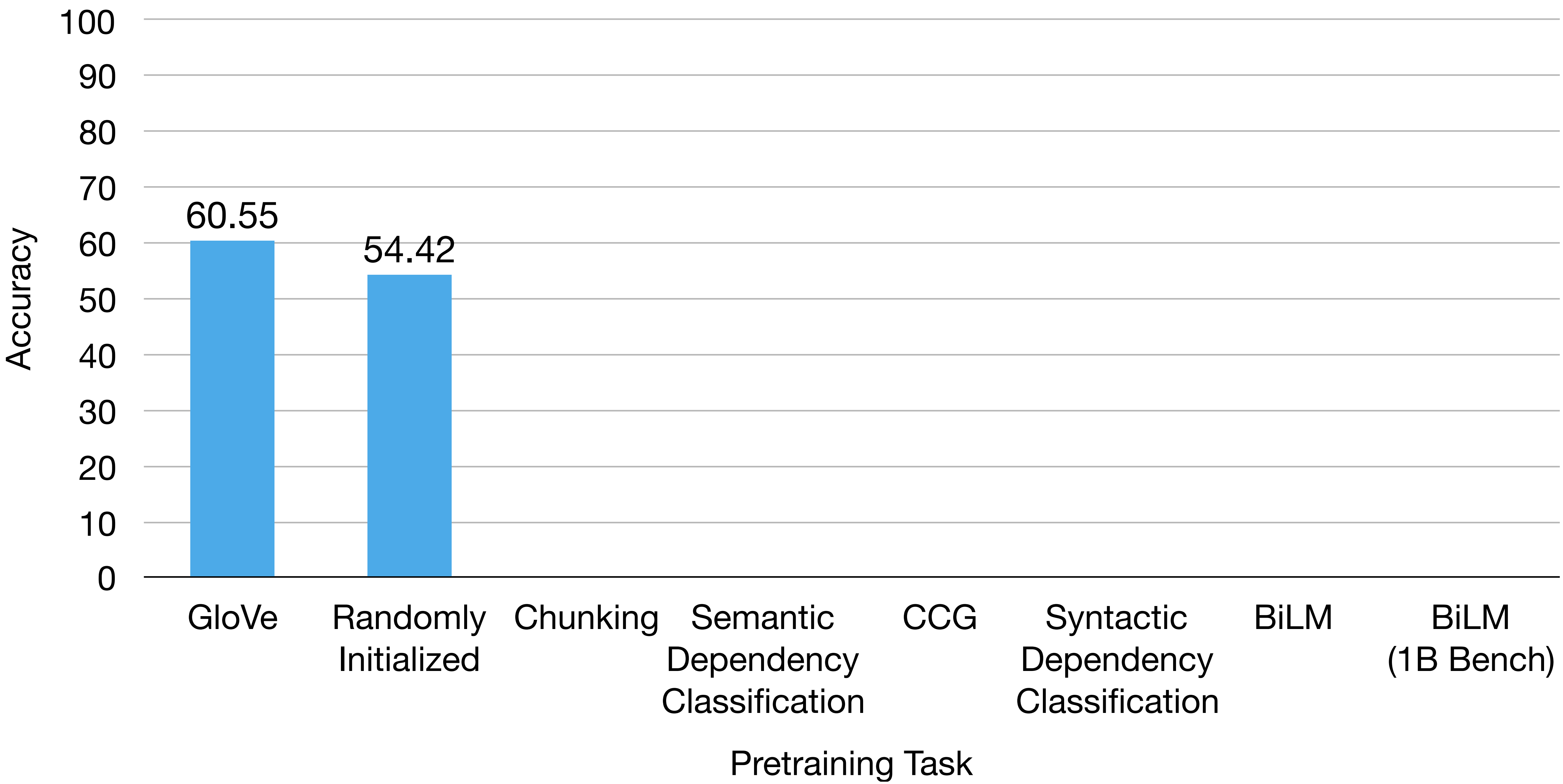
Average Across Target Tasks



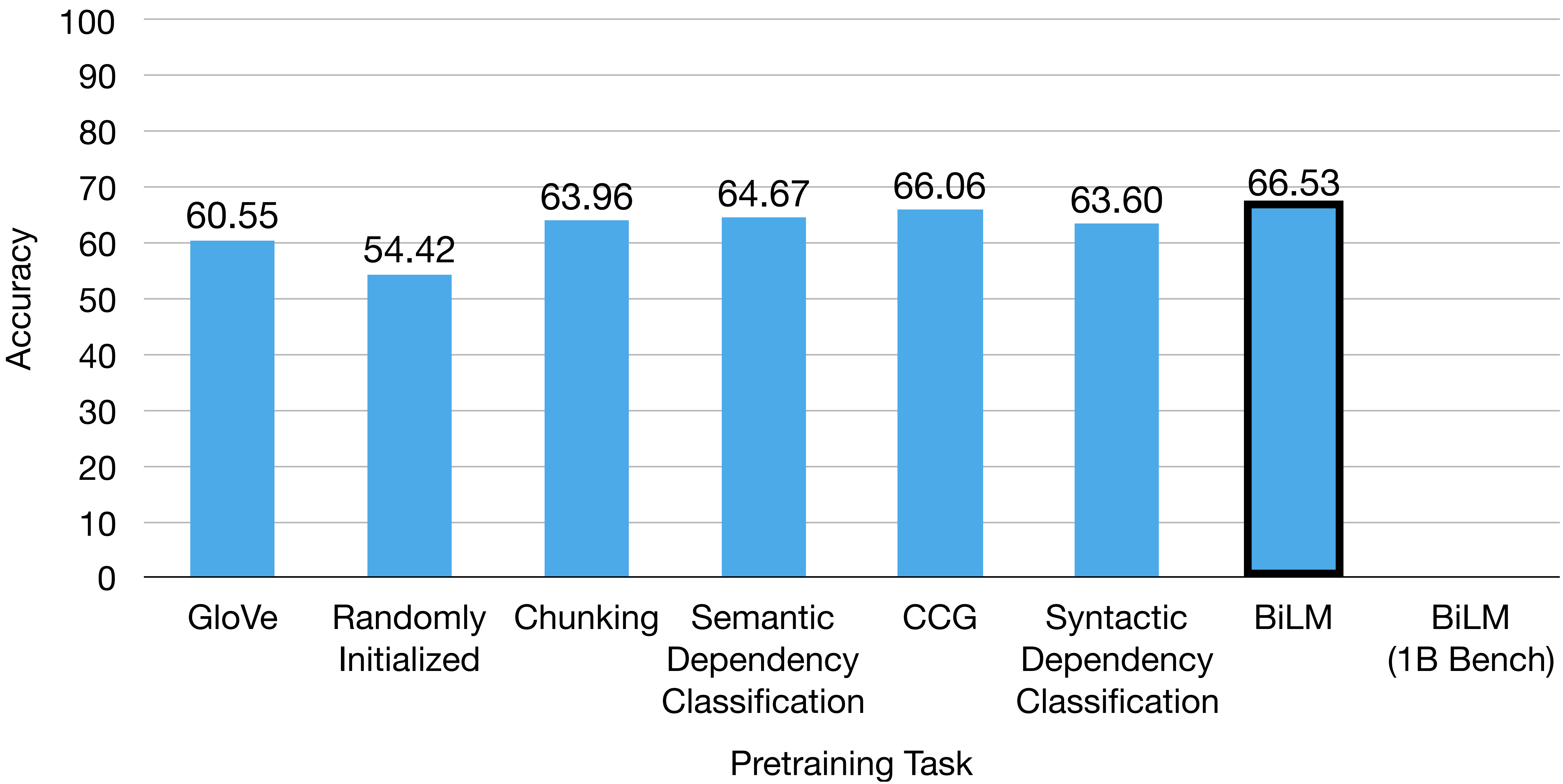
Average Across Target Tasks



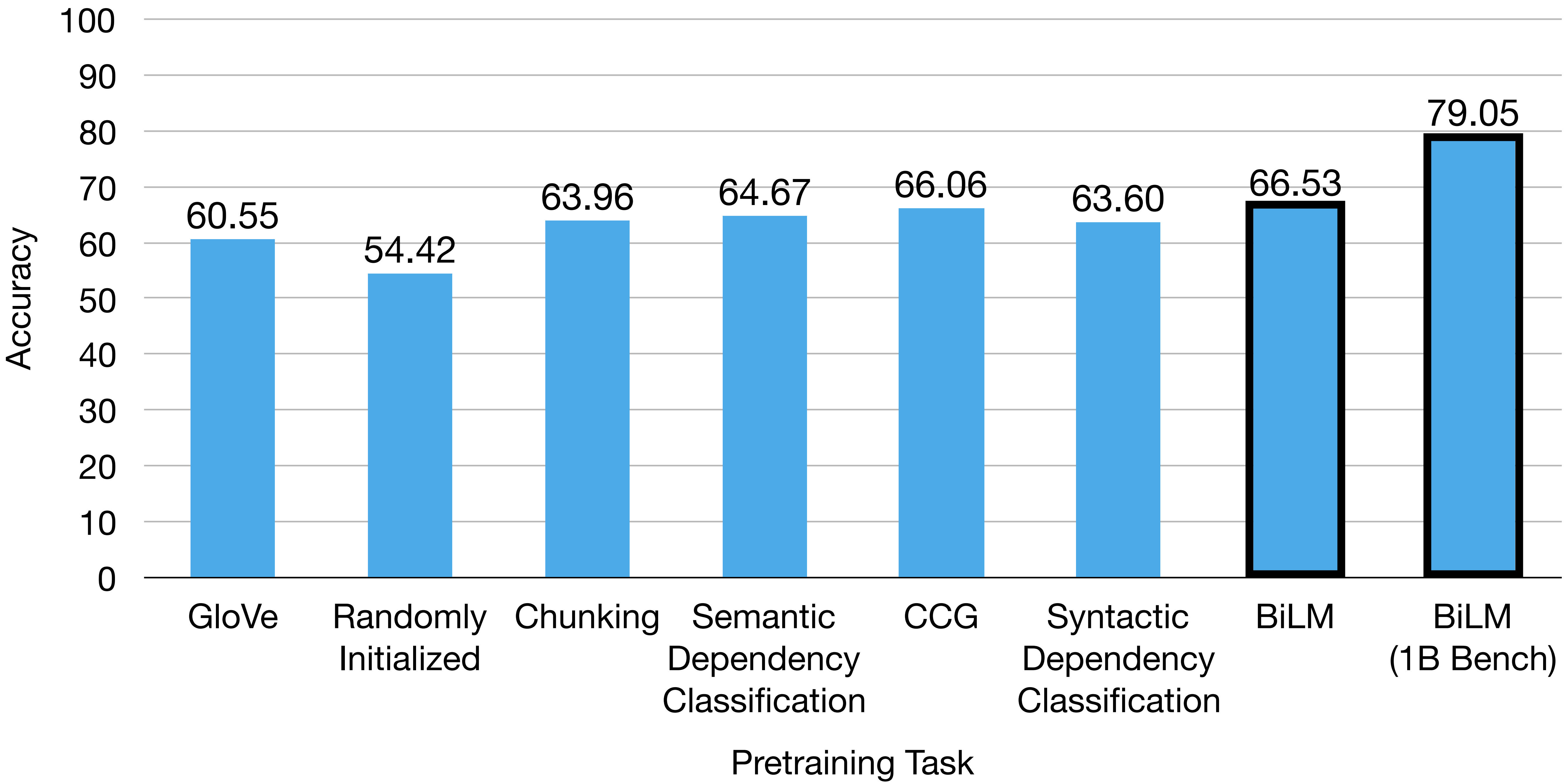
Average Across Target Tasks



Average Across Target Tasks



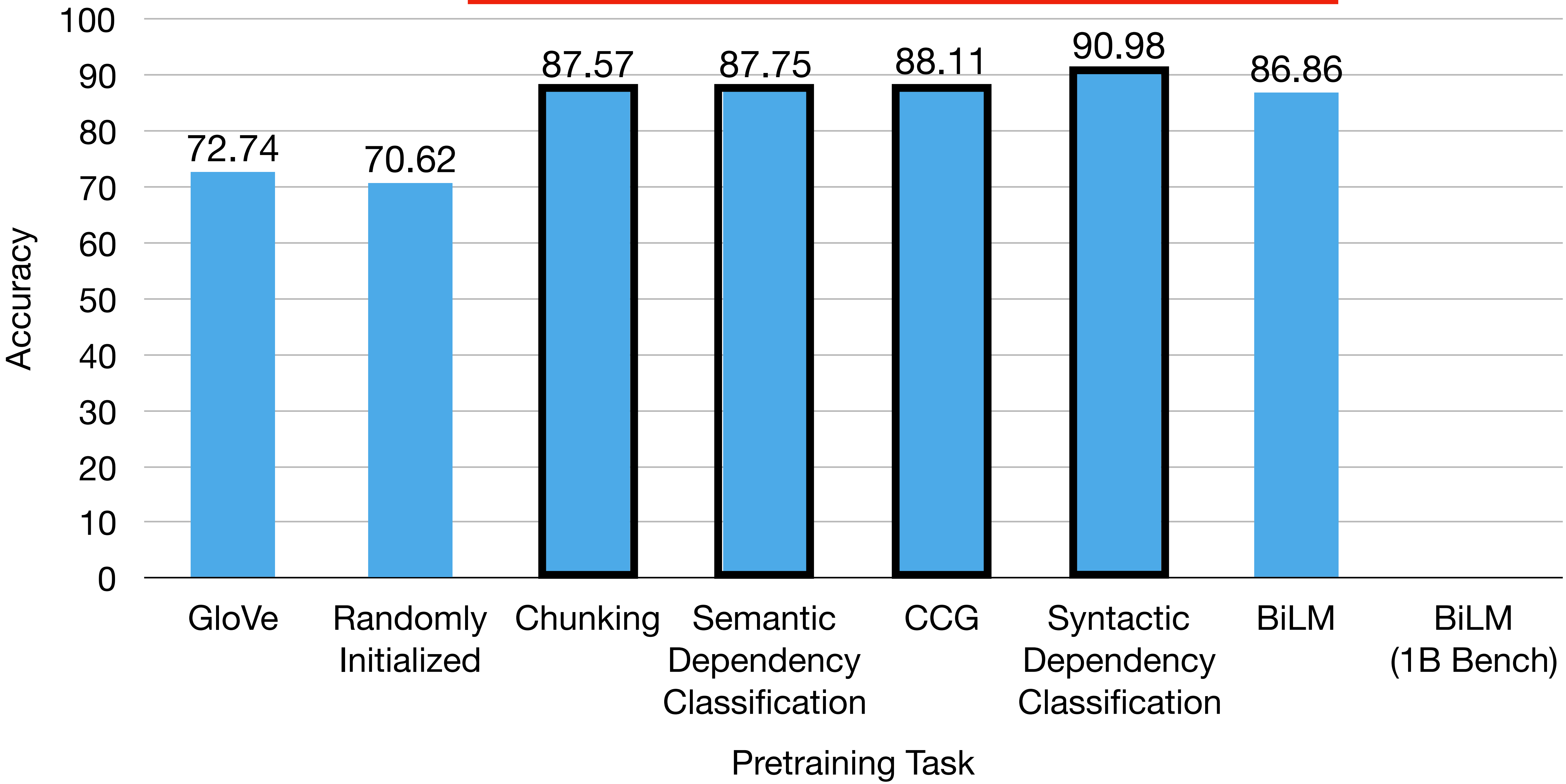
Average Across Target Tasks



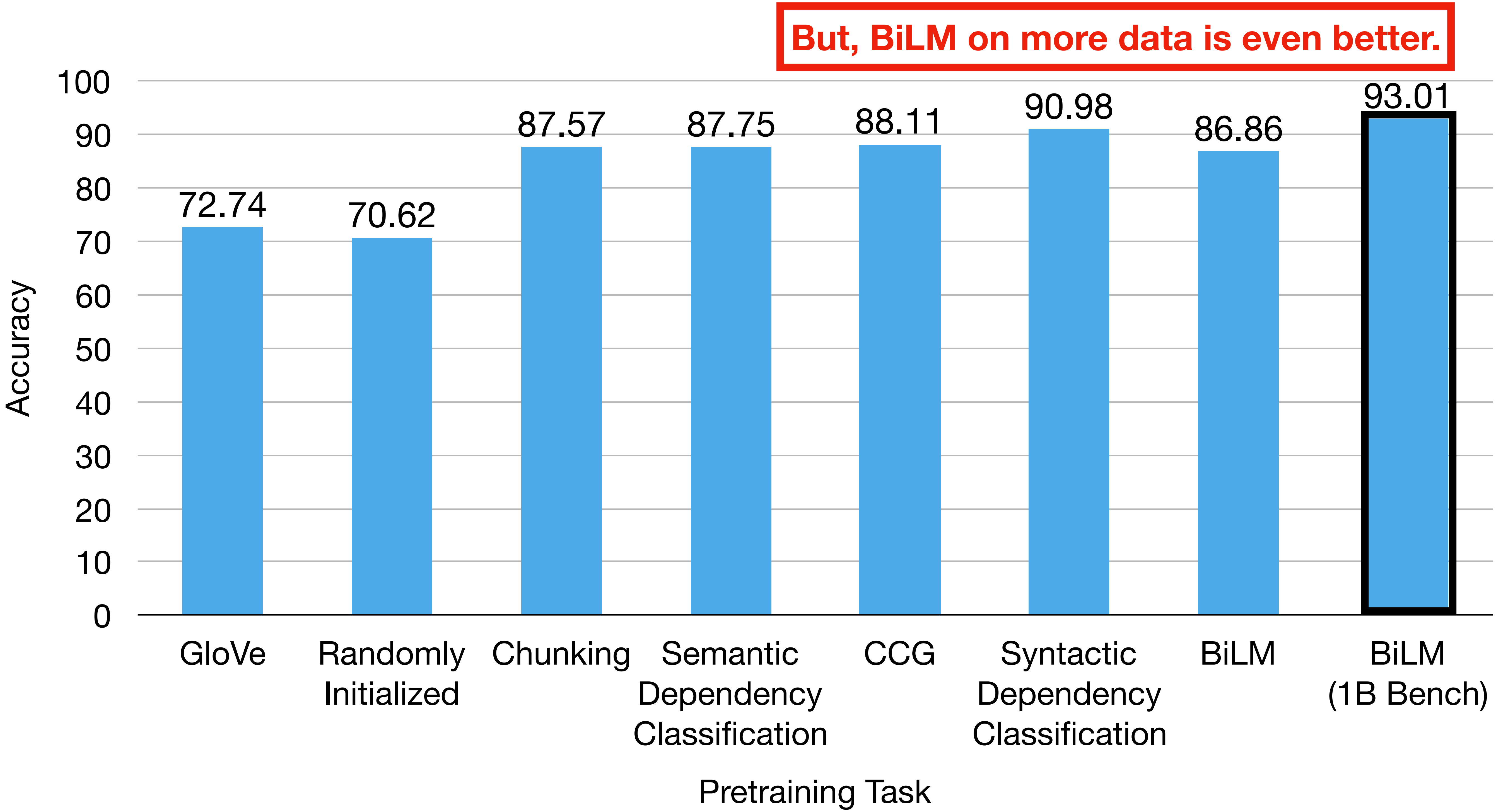
See Wang et al. (ACL 2019) "How to Get Past Sesame Street: Sentence-Level Pretraining Beyond Language Modeling" for more tasks + multitasking.

Target Task: Syntactic Dependency Classification (EWT)

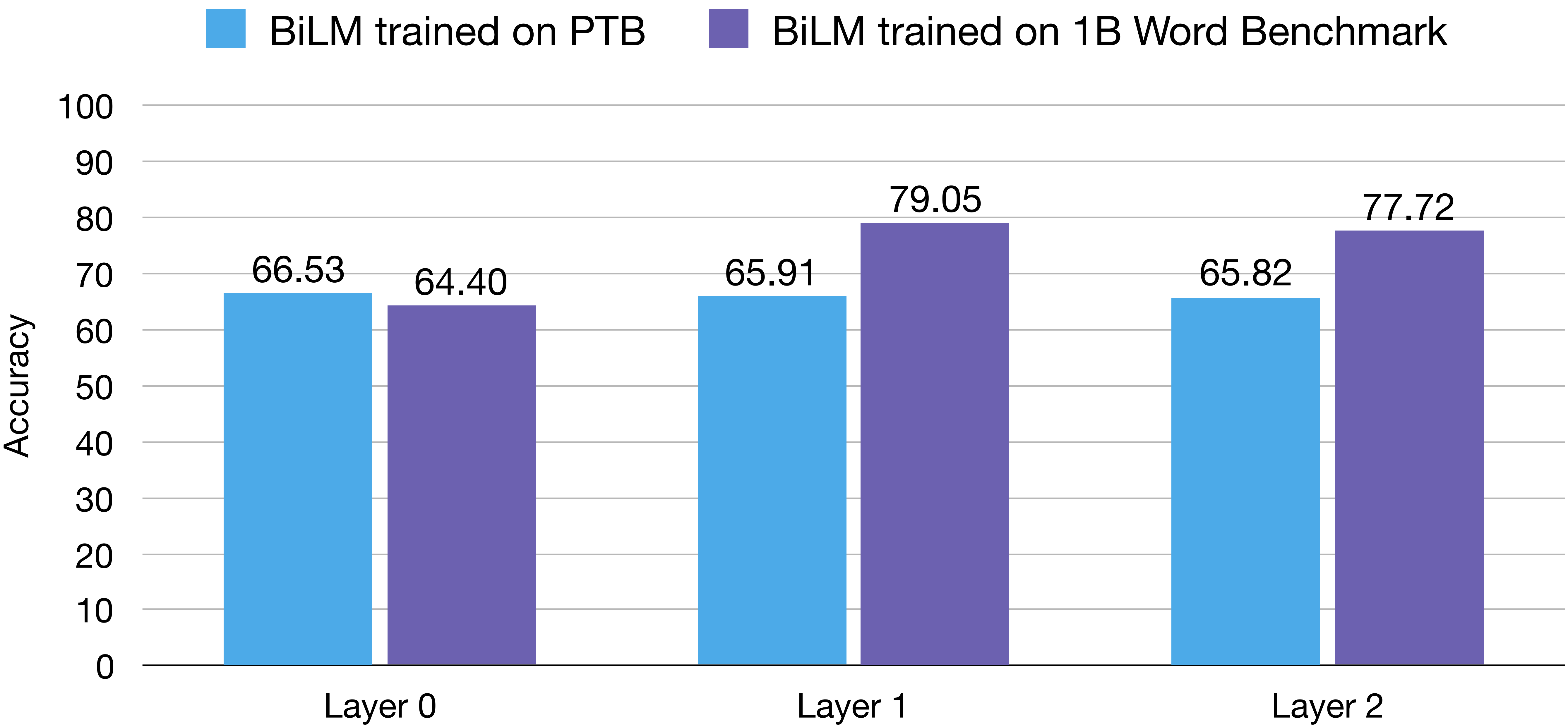
Pretraining on related tasks is better than BiLM



Target Task: Syntactic Dependency Classification (EWT)



PTB-trained BiLM vs ELMo



Also found by Saphra and Lopez (2019), check out poster 1402 on Wednesday!

Online at: bit.ly/cwr-analysis-related

Some Related Work at NAACL

Wed. June 5, 10:30 – 12:00. ML & Syntax, Hyatt Exhibit Hall:

Understanding Learning Dynamics Of Language Models with SVCCA. Naomi Saphra and Adam Lopez.

Structural Supervision Improves Learning of Non-Local Grammatical Dependencies. Ethan Wilcox et al.

Analysis Methods in Neural Language Processing: A Survey. Yonatan Belinkov and James Glass.

Wed. June 5, 16:15–16:30. Machine Learning, Nicollet B/C:

A Structural Probe for Finding Syntax in Word Representations. John Hewitt and Christopher D. Manning.

Takeaways

- Features from pretrained contextualizers are sufficient for high performance on a broad set of tasks.
- Tasks with lower performance might require fine-grained linguistic knowledge.
- Layerwise patterns in transferability exist. Dictated by how task-specific each layer is.
- Even on PTB-size data, BiLM pretraining yields the most general representations.
    - Pretraining on related tasks helps
    - *More data* helps even more!

Code: http://nelsonliu.me/papers/contextual-repr-analysis

Takeaways

Thanks!
Questions?

- Features from pretrained contextualizers are sufficient for high performance on a broad set of tasks.
- Tasks with lower performance might require fine-grained linguistic knowledge.
- Layerwise patterns in transferability exist. Dictated by how task-specific each layer is.
- Even on PTB-size data, BiLM pretraining yields the most general representations.
  - Pretraining on related tasks helps
  - *More data* helps even more!

Code: `http://nelsonliu.me/papers/contextual-repr-analysis`

Bonus Slides

Probing Task Examples

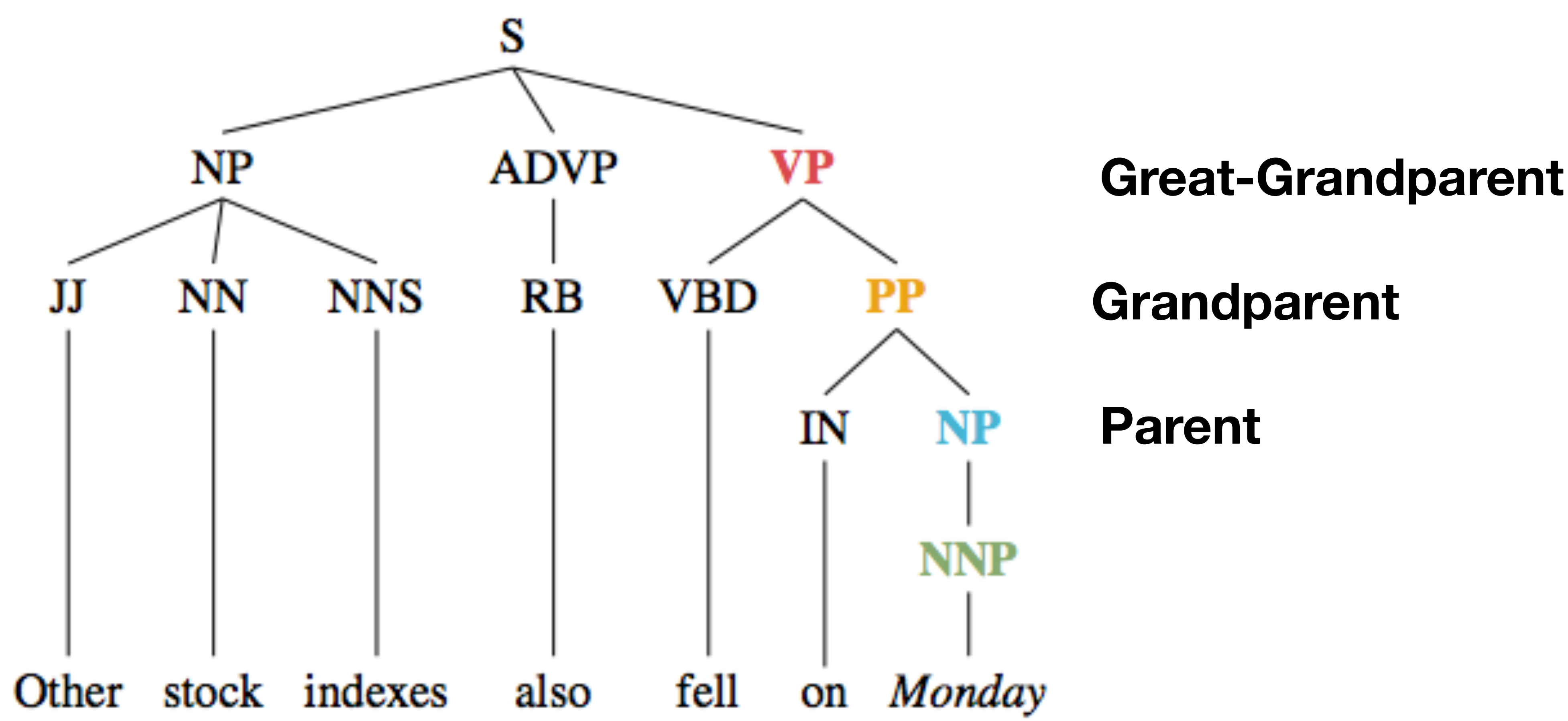
Part-of-Speech Tagging

Soon	she	was	running	the	office
RB	PRP	VBD	VBG	DT	NN

CCG Supertagging

$$
\begin{array}{cccccc}
\underline{\text{Soon}} & \underline{\text{she}} & \underline{\text{was}} & \underline{\text{running}} & \underline{\text{the}} & \underline{\text{office}} \\
S/S & NP & (S\backslash NP)/NP & NP & NP/N & N \\
\cline{3-4} \cline{5-6}
 & & \multicolumn{2}{c}{(S\backslash NP)/NP} & \multicolumn{2}{c}{NP} \\
\cline{3-6}
 & & \multicolumn{4}{c}{S\backslash NP} \\
\cline{2-6}
 & \multicolumn{5}{c}{S} \\
\cline{1-6}
\multicolumn{6}{c}{S}
\end{array}
$$

Syntactic Constituency Ancestor Tagging



Semantic Tagging

- Semantic tags abstract over redundant POS distinctions and disambiguate useful cases within POS tags.
- (1) Sarah bought **herself** a book
- (2) Sarah **herself** bought a book
- Same POS tag (Personal Pronoun), but different semantic function. (1) reflexive function, (2) emphasizing function

Preposition Supersense Disambiguation

- Classify a preposition's lexical semantic contribution (function), or the semantic role / relation it mediates (role).

- Specialized kind of word sense disambiguation.

Preposition Supersense Disambiguation

(1) I was booked **for**/DURATION 2 nights **at**/LOCUS this hotel **in**/TIME Oct 2007 .

(2) I went **to**/GOAL ohm **after**/EXPLANATION⇝TIME reading some **of**/QUANTITY⇝WHOLE the reviews .

(3) It was very upsetting to see this kind **of**/SPECIES behavior especially **in_front_of**/LOCUS **my**/SOCIALREL⇝GESTALT four year_old .

Event Factuality

- Label predicates with the factuality of events they describe.

Event "leave" did not happen.

(3) a. Jo didn't remember to **leave**.

b. Jo didn't remember **leaving**.

Event "leaving" happened.

Syntactic Chunking

[NP He] [VP reckons] [NP the current account deficit] [VP will narrow] [PP to] [NP only # 1.8 billion] [PP in] [NP September] .

Named Entity Recognition

[ORG U.N.] official [PER Ekeus] heads for [LOC Baghdad] .

Grammatical Error Detection

```
+ +    +  -          +   +        +   +    +    -       +
I like to playing the guitar and sing very louder .
```

Conjunct Identification

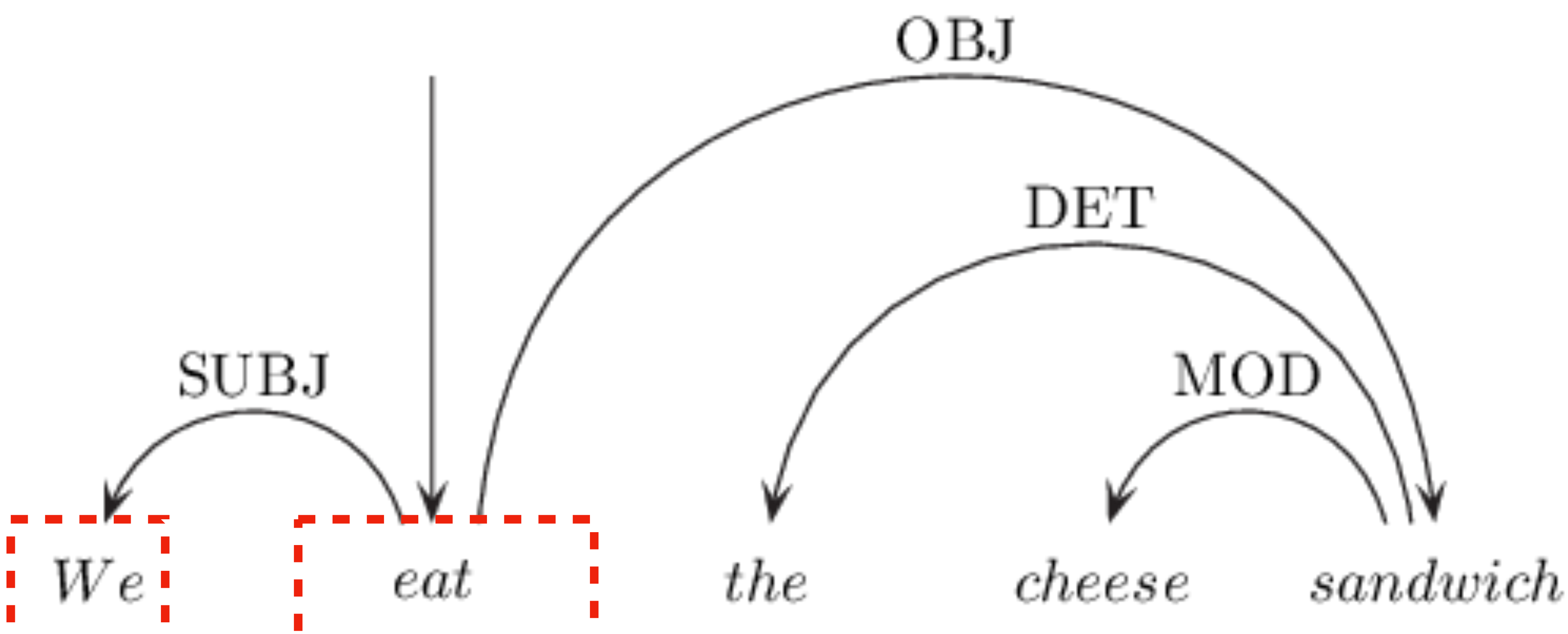
- And the city decided to treat its guests more like **[royalty]** or **[rock stars]** than factory owners.

Two Types of Pairwise Relations

- **Arc prediction tasks:** Given two random tokens, identify **whether** a relation exists between them.

- **Arc classification tasks:** Given two tokens that are known to be related, **identify what** the relation is.

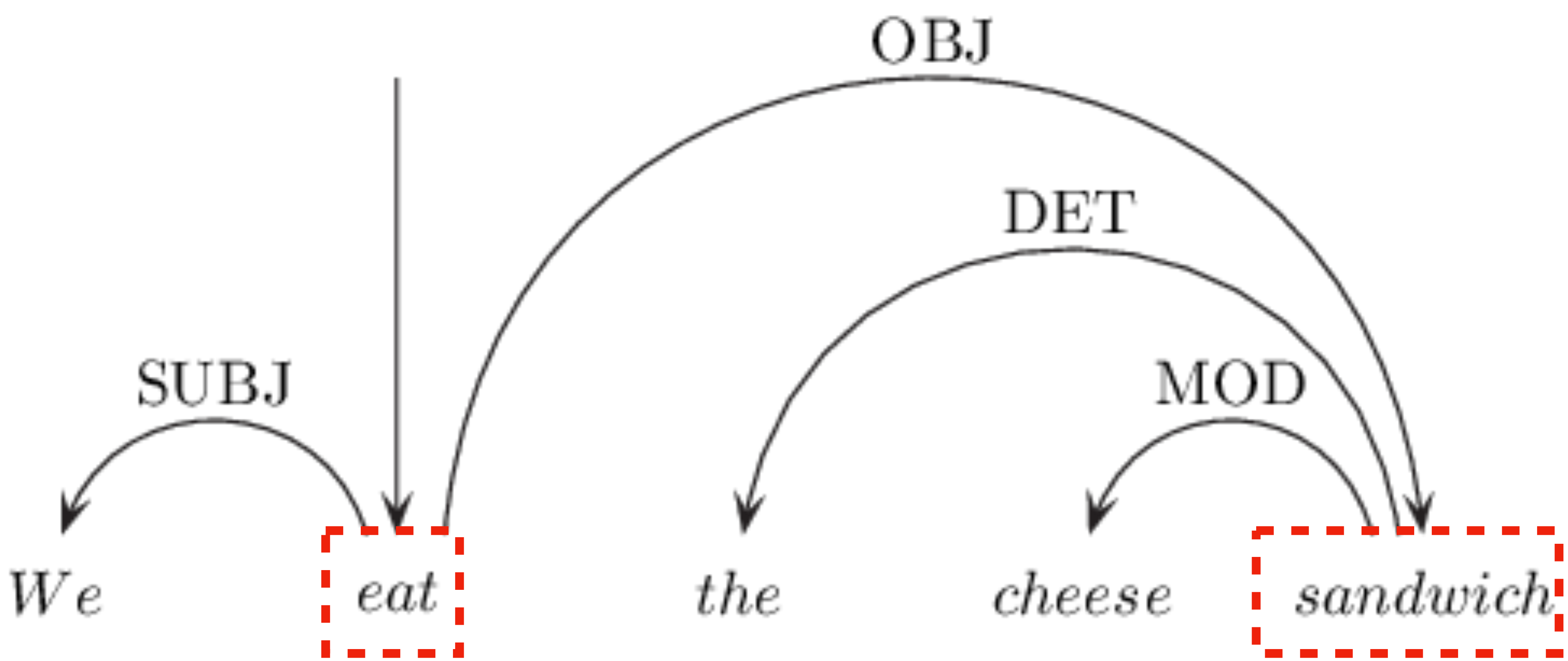
Syntactic Dependency Arc Prediction



Input Tokens

Label: True, there exists a relation

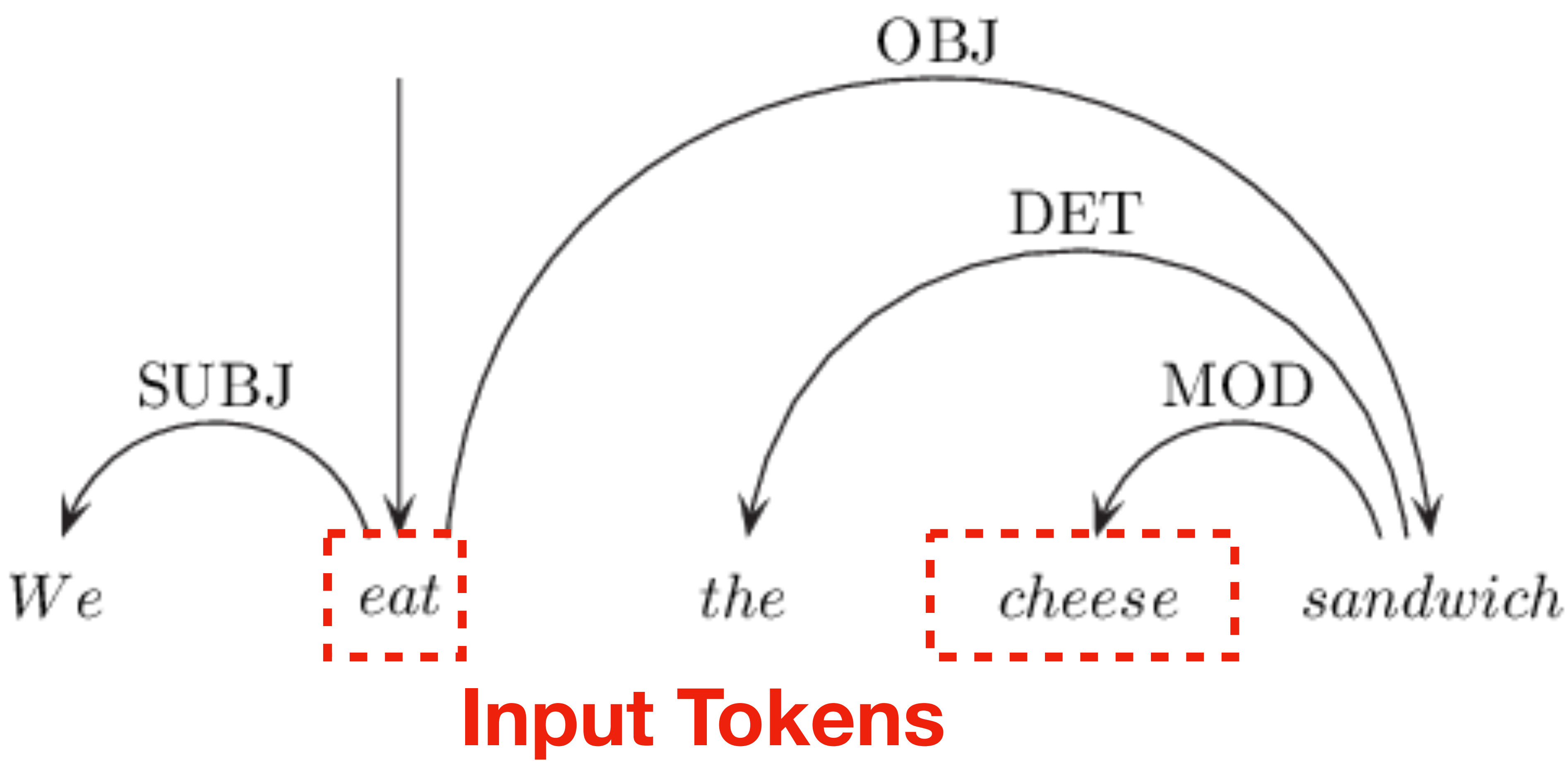
Syntactic Dependency Arc Prediction



Input Tokens

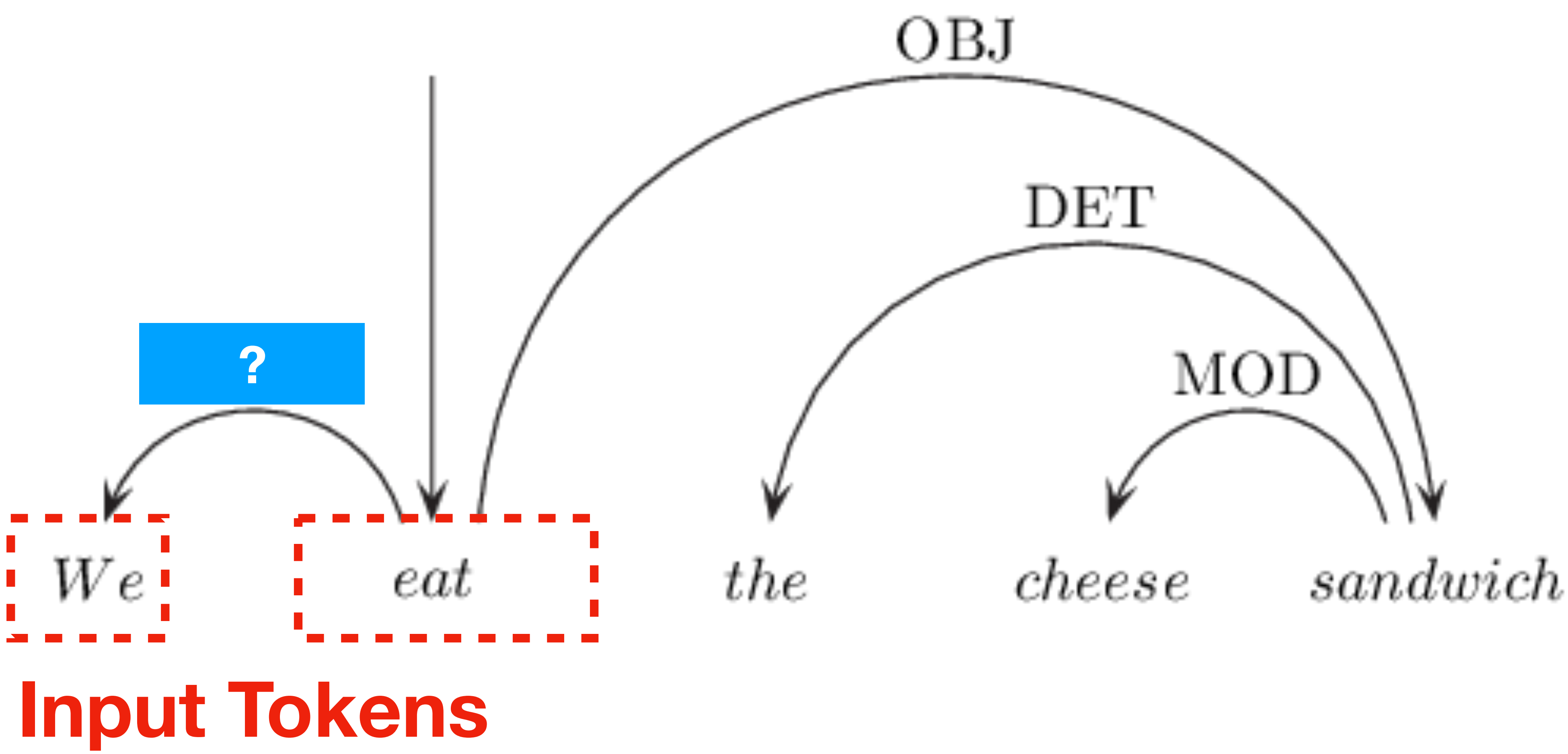
Label: True, there exists a relation

Syntactic Dependency Arc Prediction

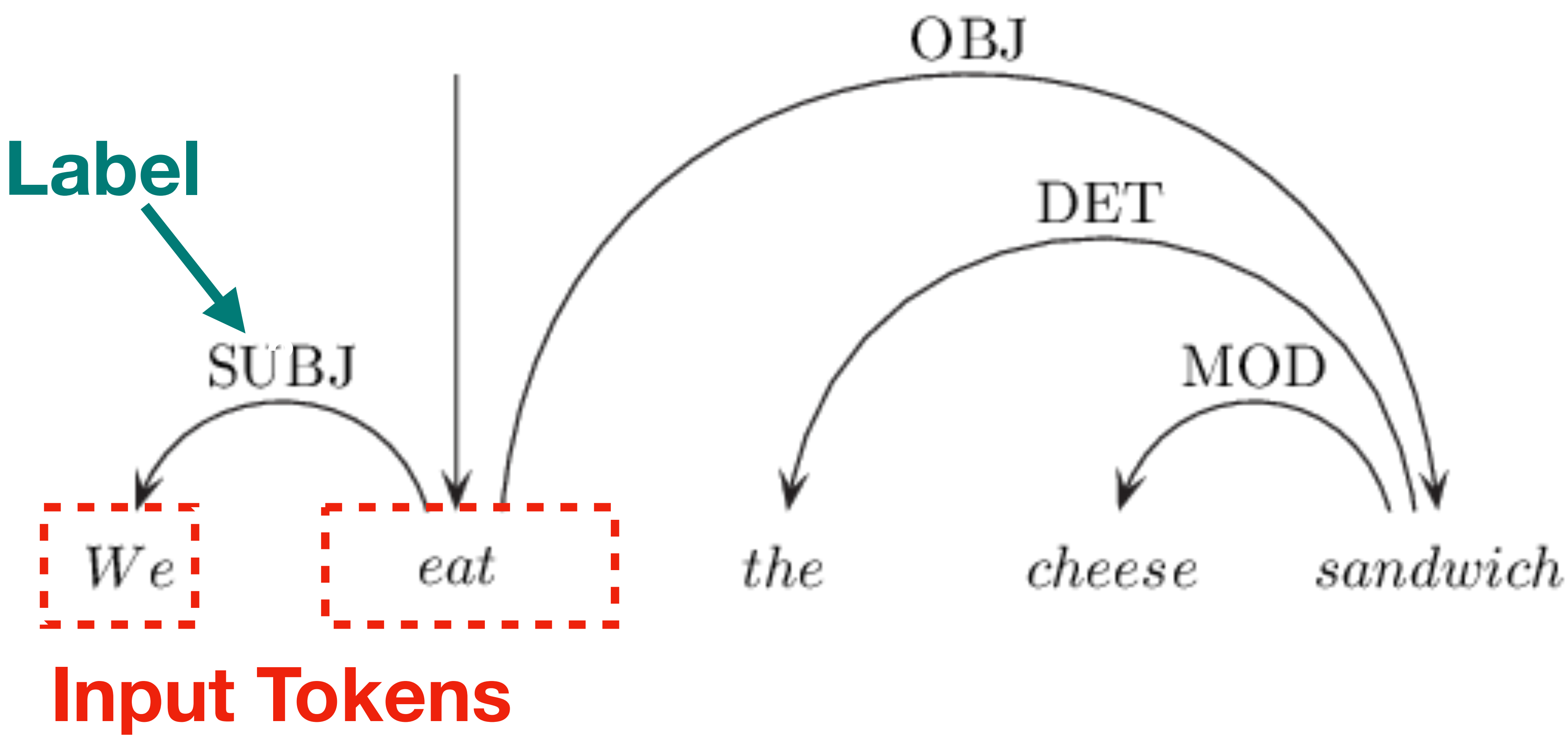


Label: False, there does not exist a relation

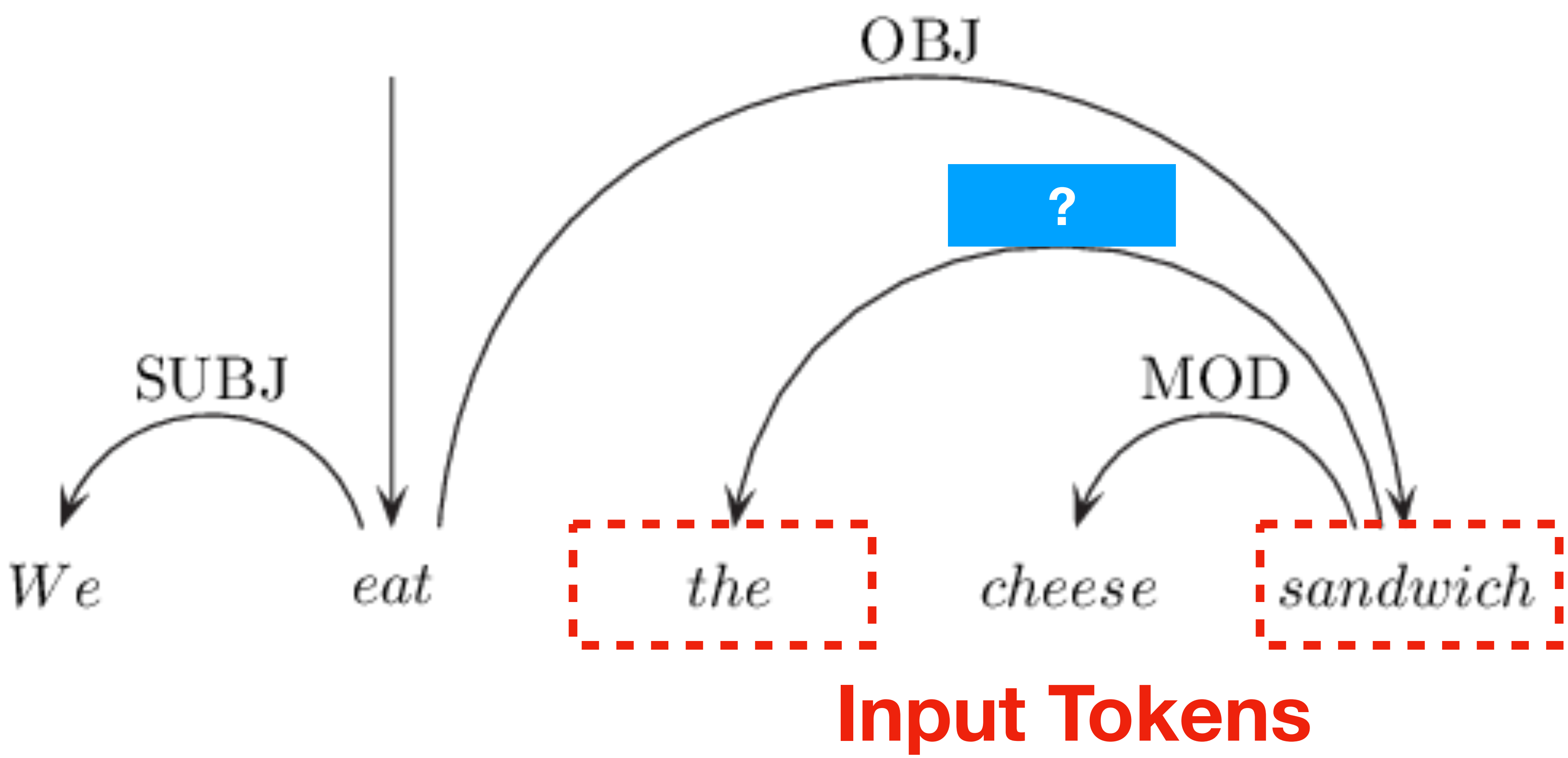
Syntactic Dependency Arc Classification



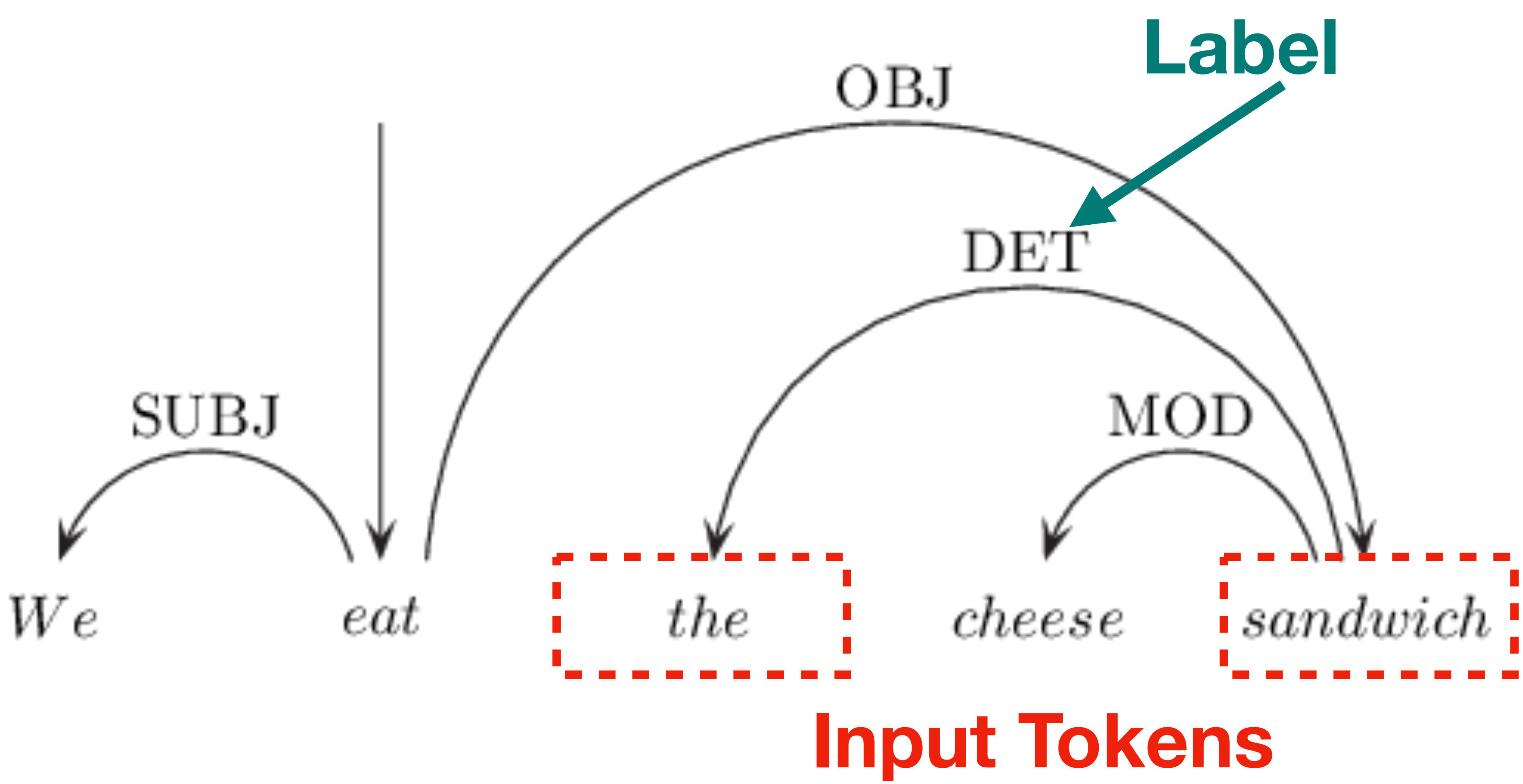
Syntactic Dependency Arc Classification



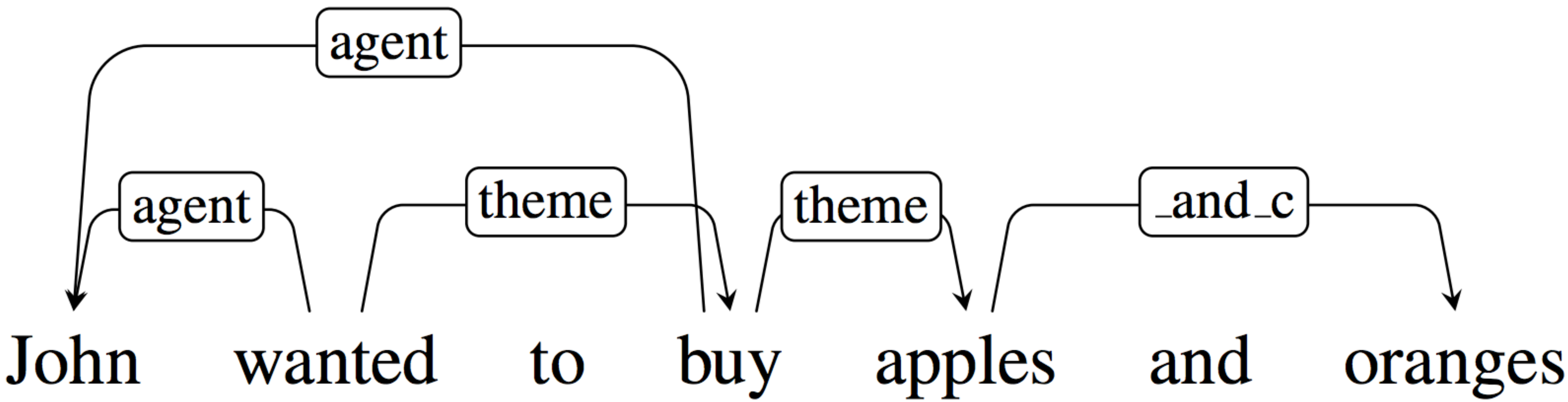
Syntactic Dependency Arc Classification



Syntactic Dependency Arc Classification



Semantic Dependencies

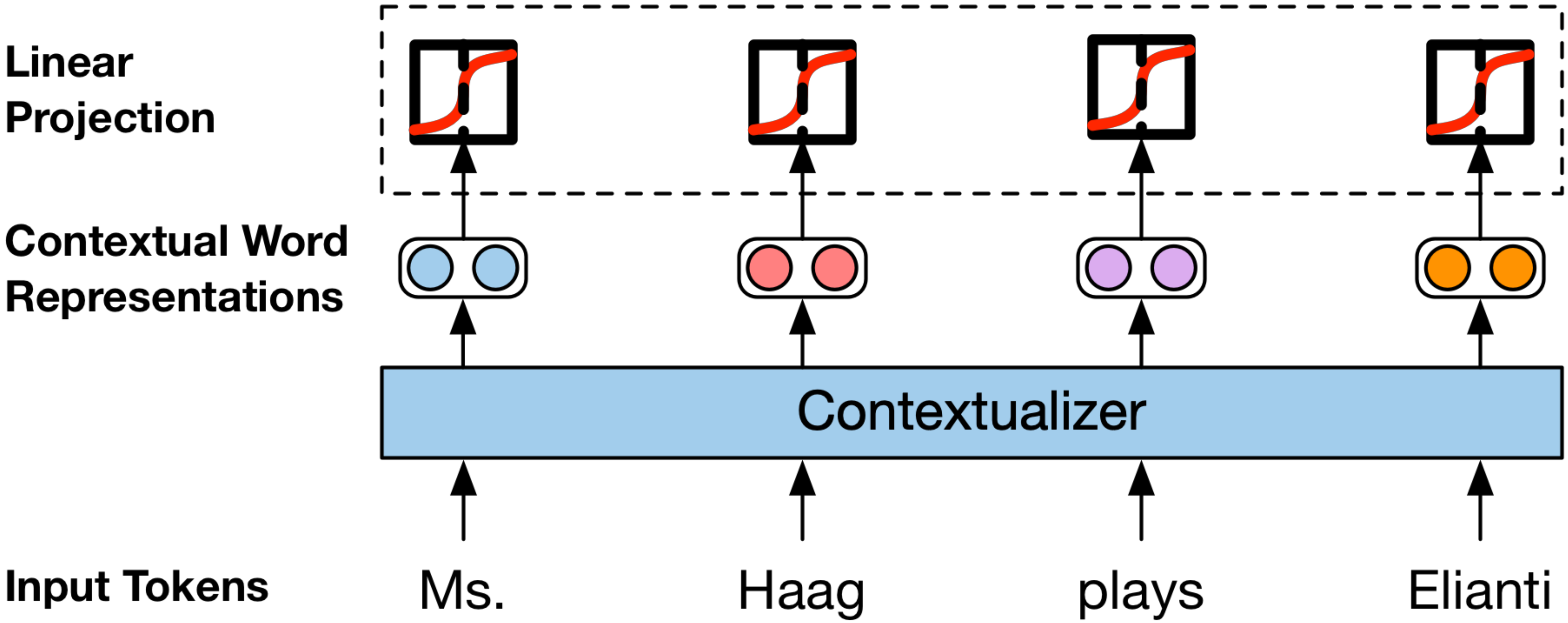


Coreference Relations

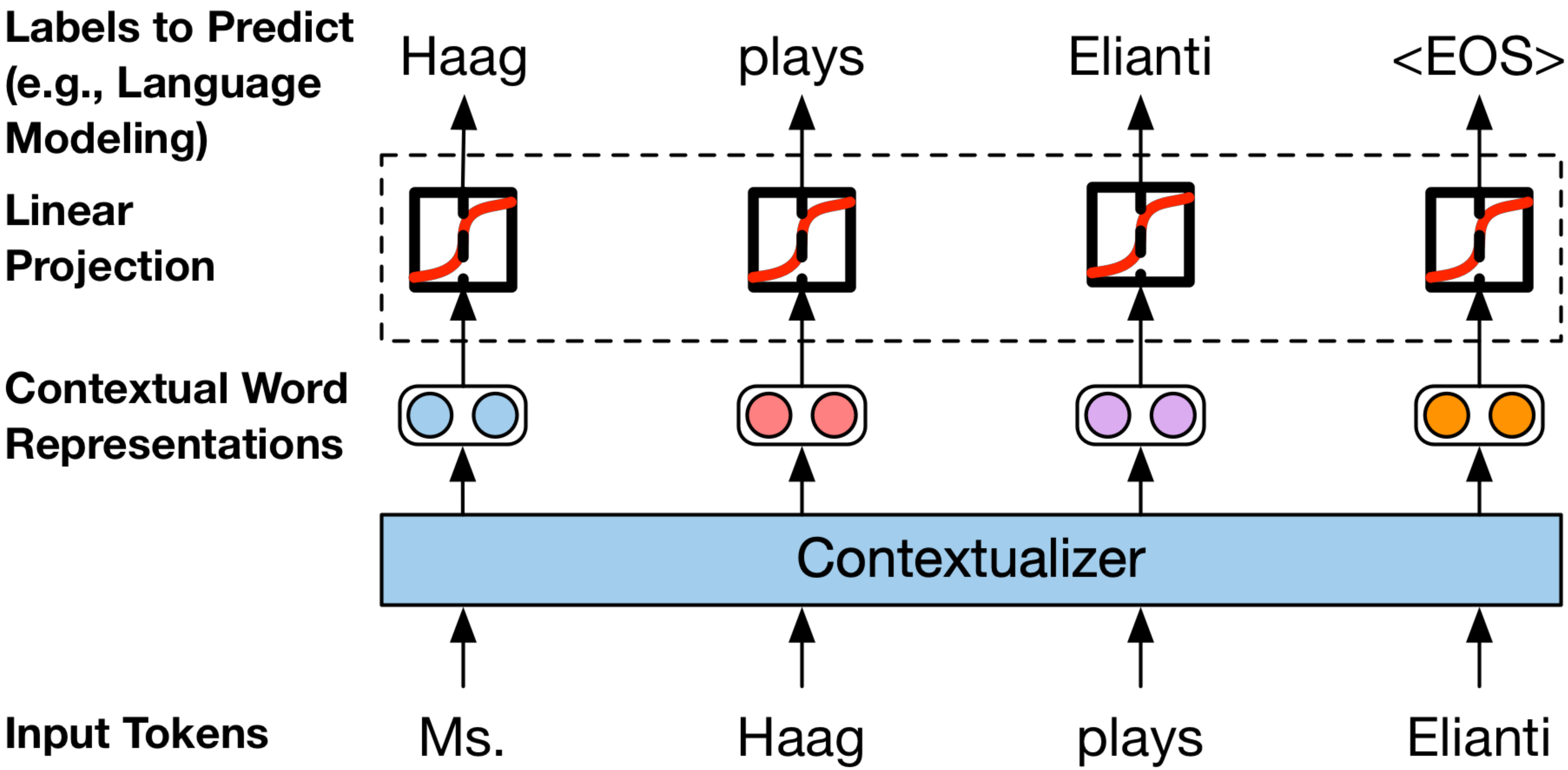
"I voted for Nader because he was most aligned with my values," she said.

Setting Up Alternative Pretraining Objectives

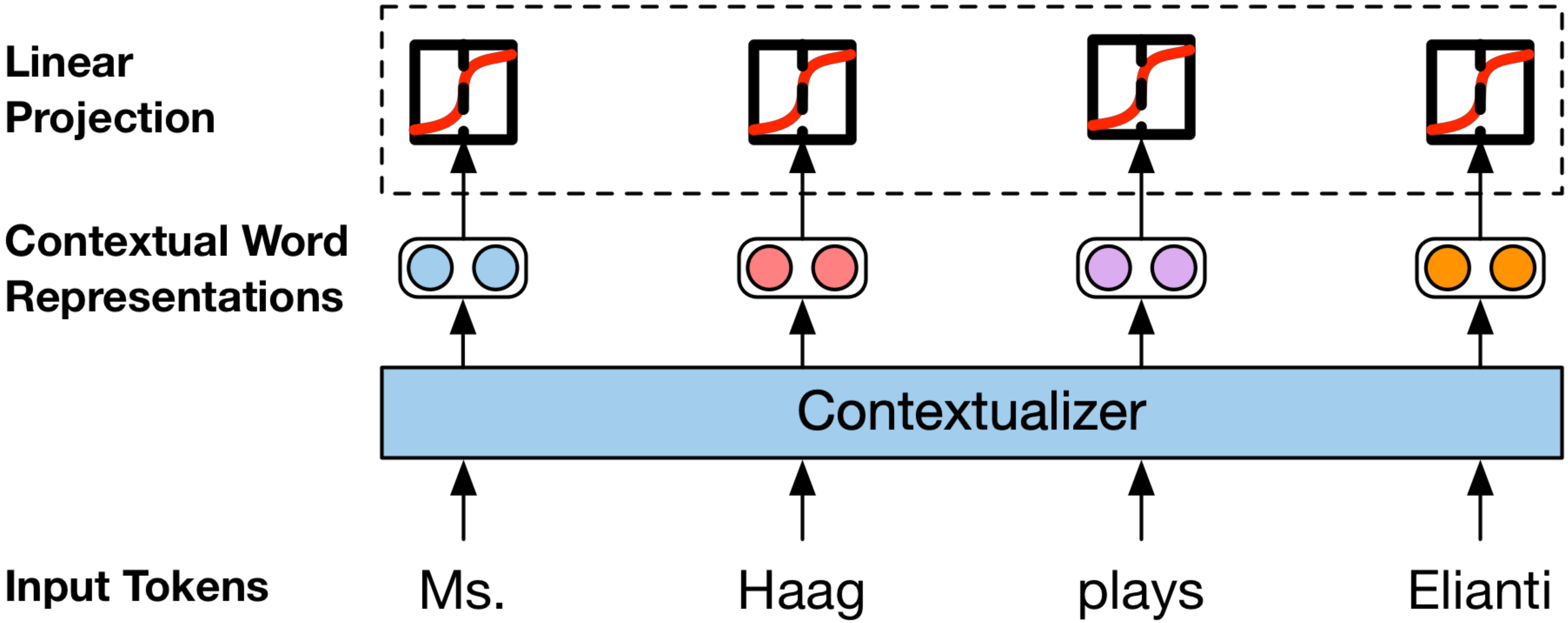
Language Model Pretraining



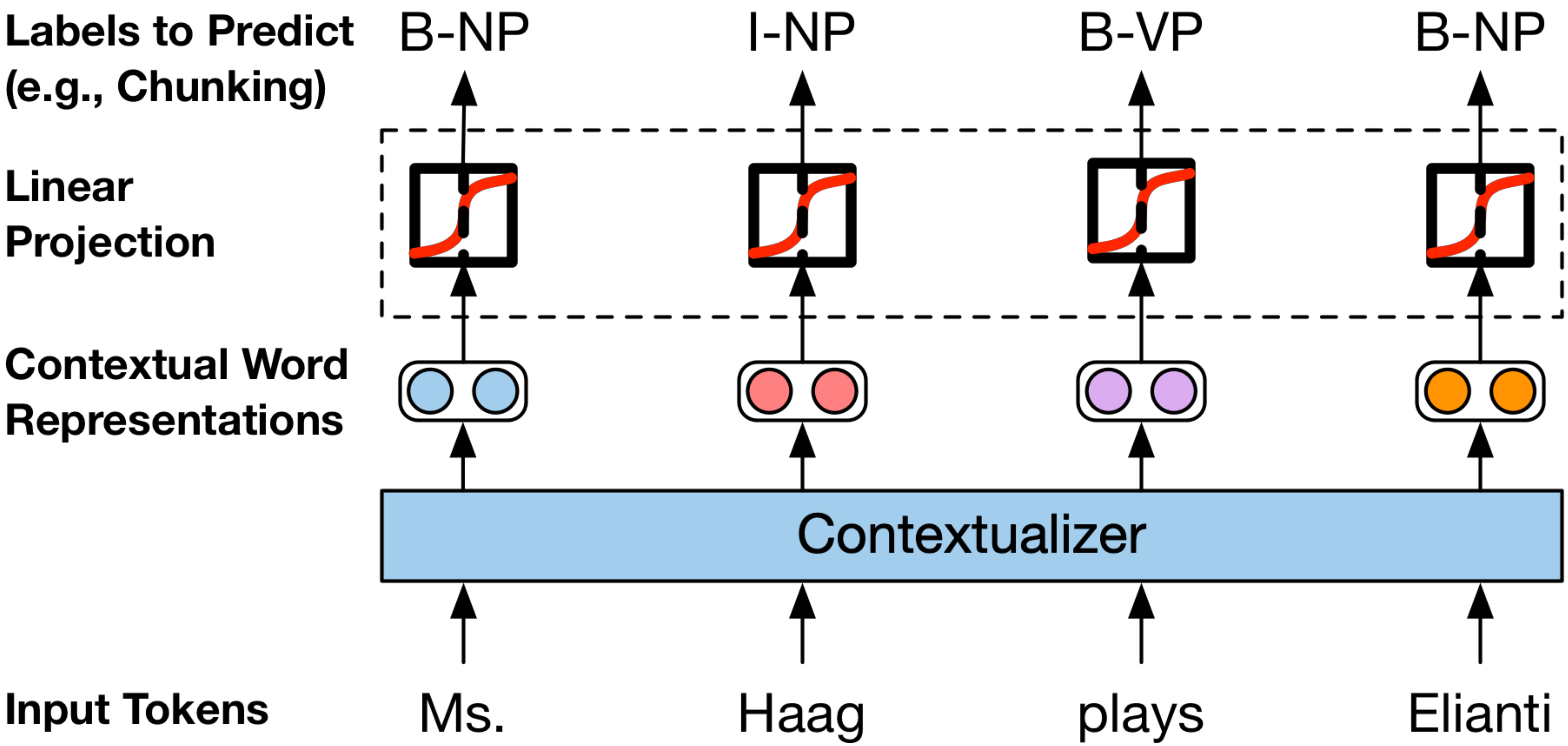
Language Model Pretraining



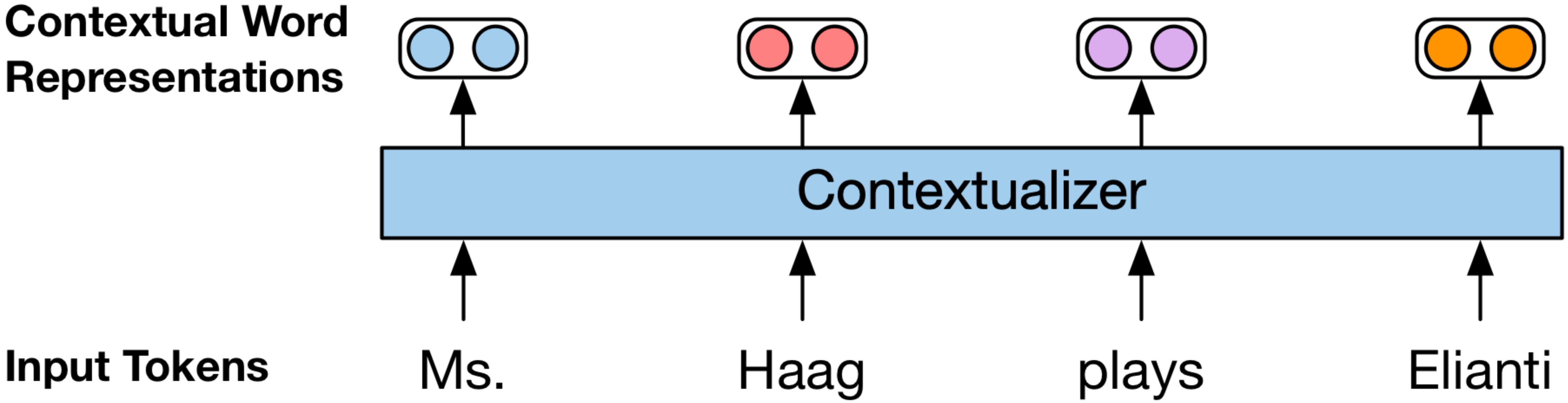
Chunking Pretraining

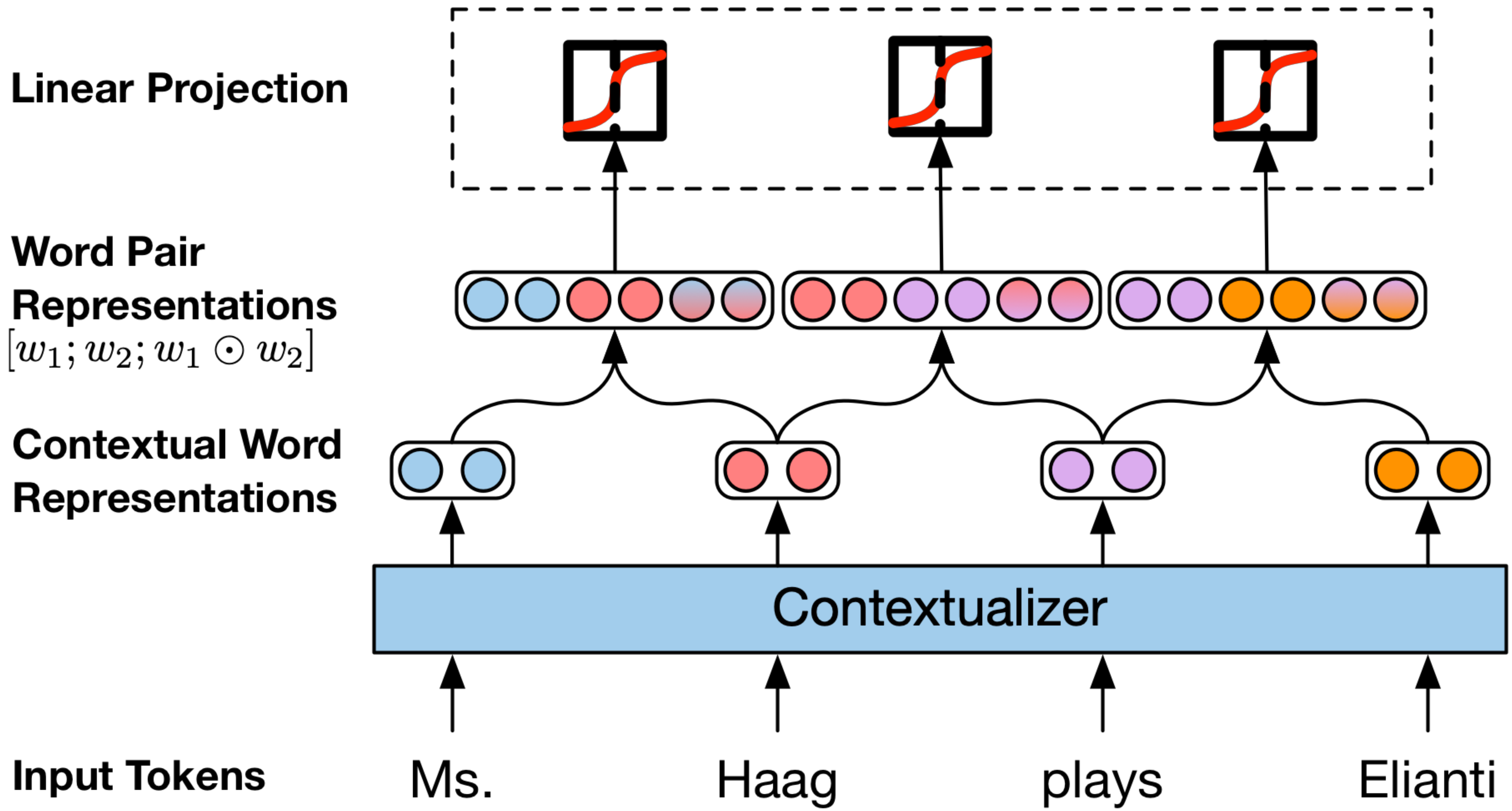


Chunking Pretraining

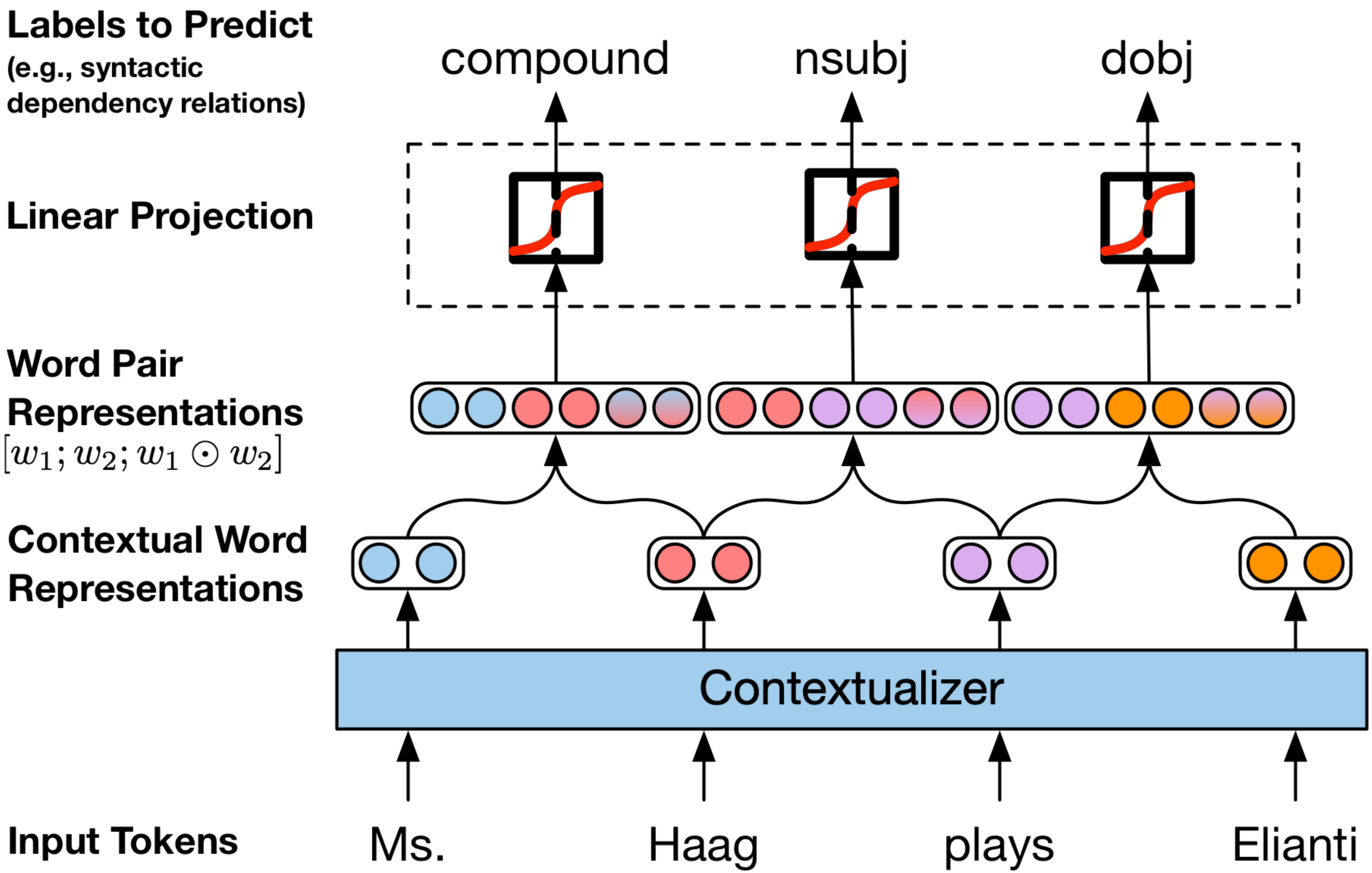


Flexible Paradigm, Use Any Task!





Linear Projection
Word Pair
Representations
$[w_1; w_2; w_1 \odot w_2]$
Contextual Word
Representations
Contextualizer
Input Tokens
Ms.
Haag
plays
Elianti


Labels to Predict
(e.g., syntactic dependency relations)
compound
nsubj
dobj
Linear Projection
Word Pair Representations
$[w_1; w_2; w_1 \odot w_2]$
Contextual Word Representations
Contextualizer
Input Tokens
Ms.
Haag
plays
Elianti