Task-Independent Language Understanding

NYU
ML²
Machine Learning
for Language

Sam Bowman
@sleepinyourhat

The Goal

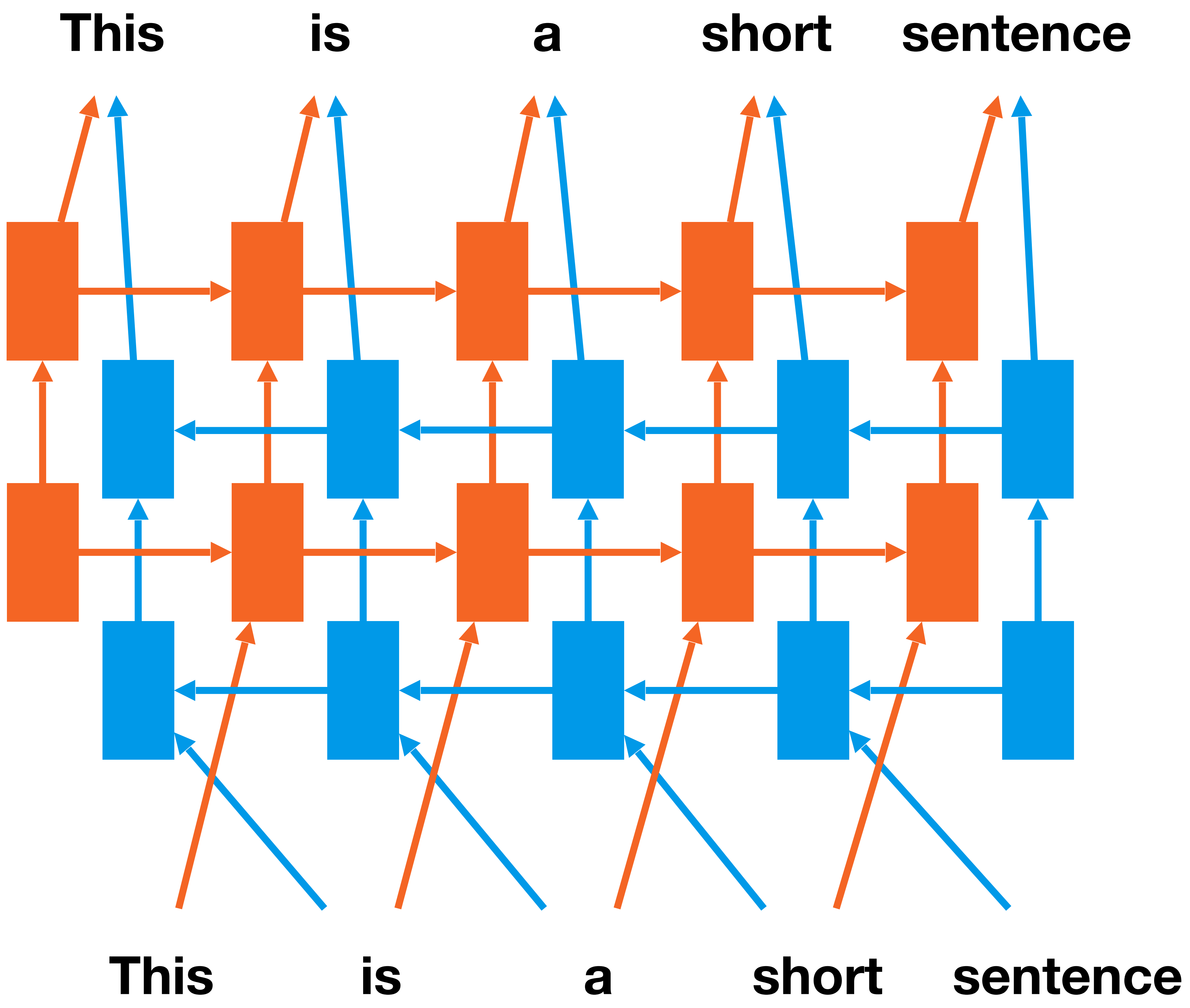
To develop a **general-purpose neural network encoder for text** which makes it possible to solve any new **language understanding task** using only enough training data to **define the possible outputs**.

The Goal

To develop a neural network model that **already understands English** when it starts learning a new task.

Case Study: ELMo

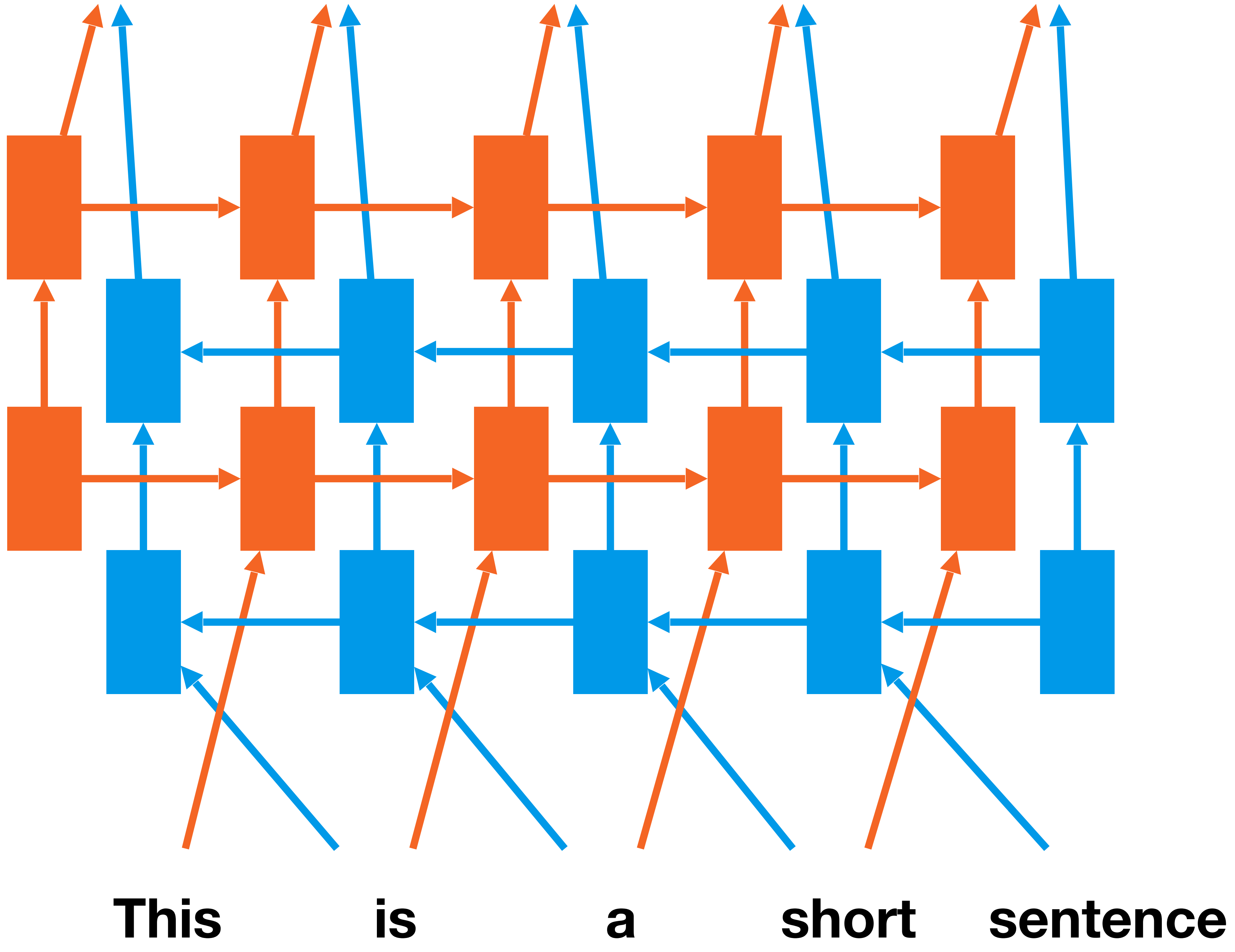
Train large forward ***and backward*** deep LSTM language models.



Peters et al. '18

Case Study: ELMo

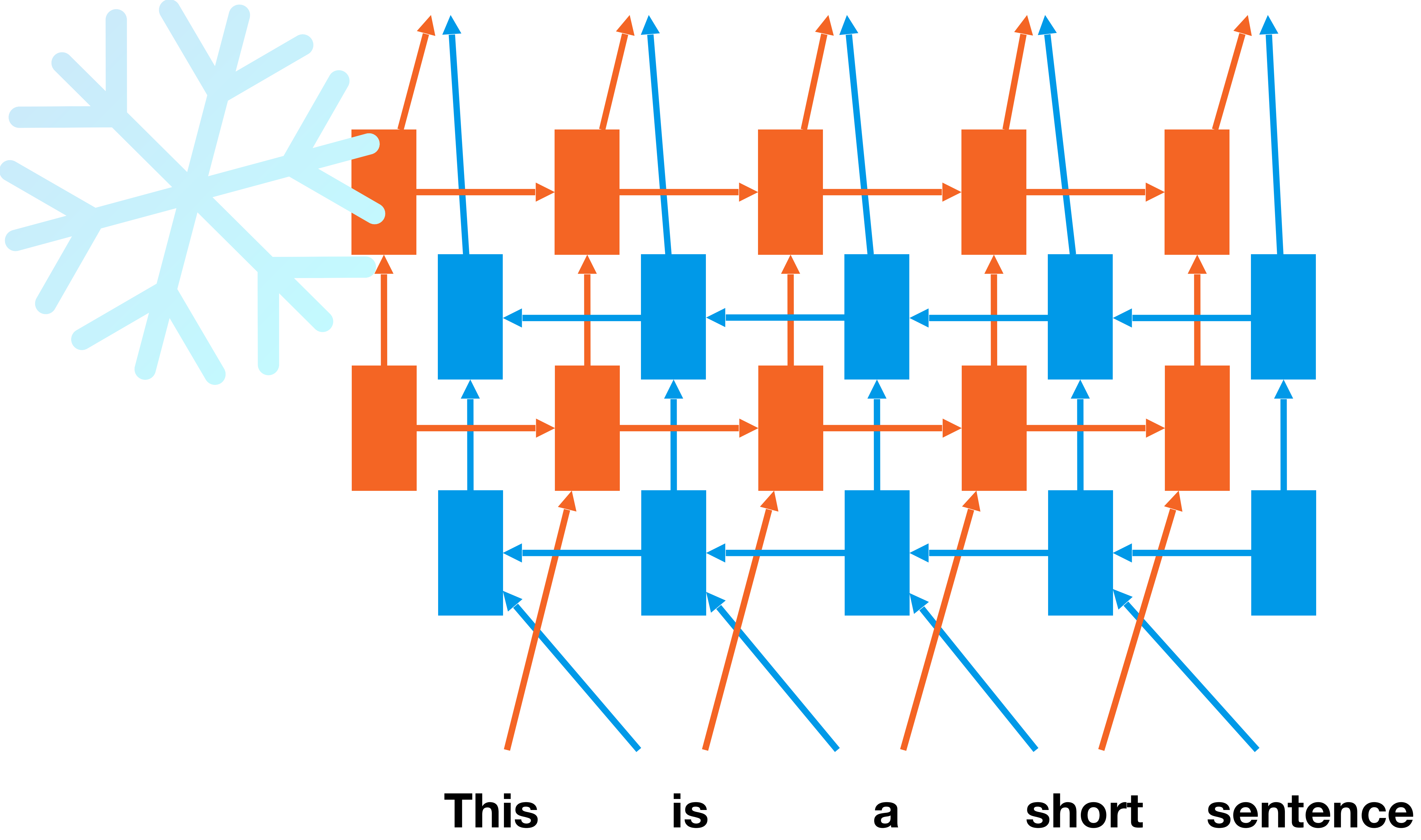
Train large (~100m-param) forward ***and backward*** deep LSTM language models.



Peters et al. '18

Case Study: ELMo

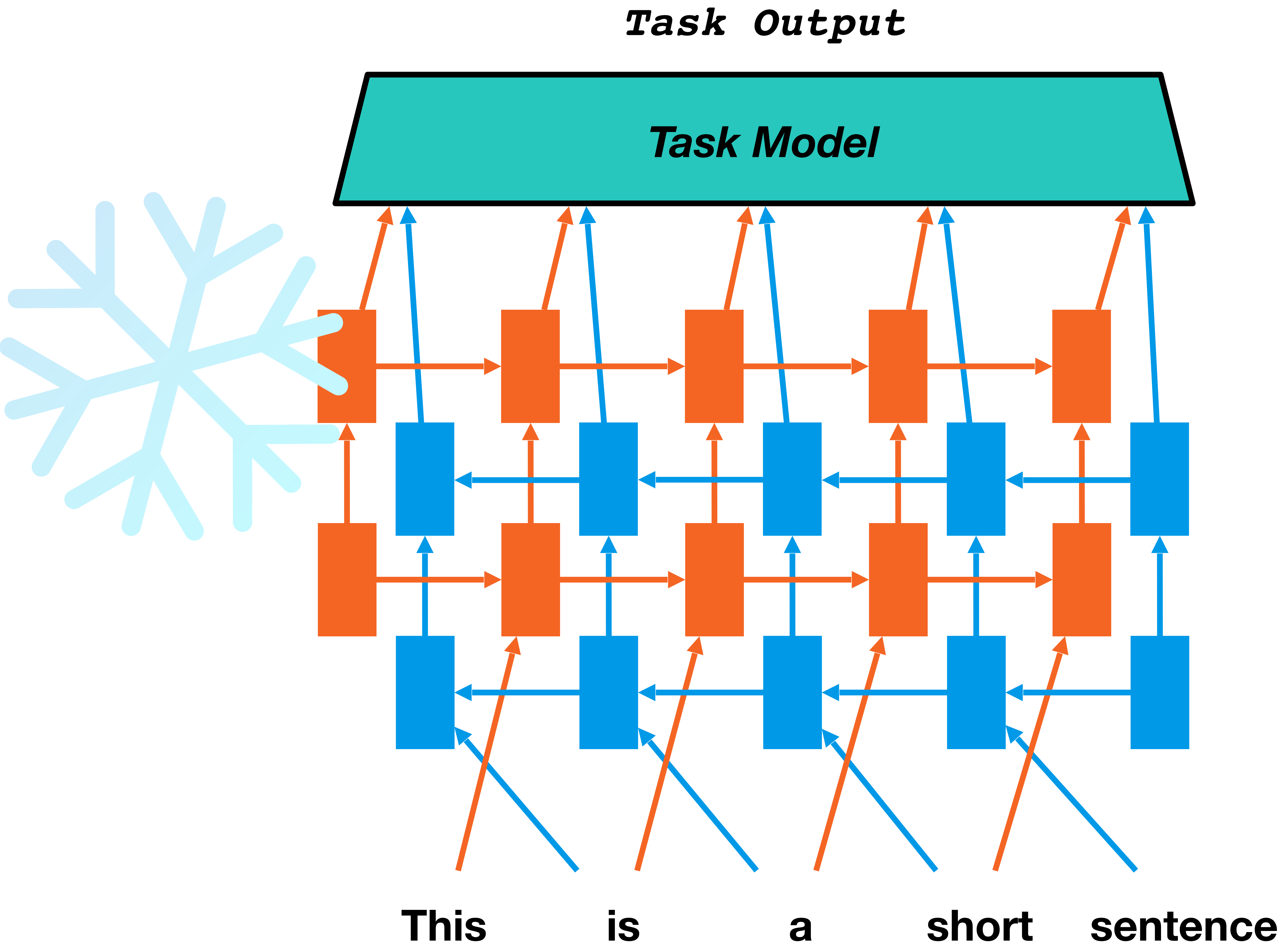
Train large (~100m-param) forward ***and backward*** deep LSTM language models.



Peters et al. '18

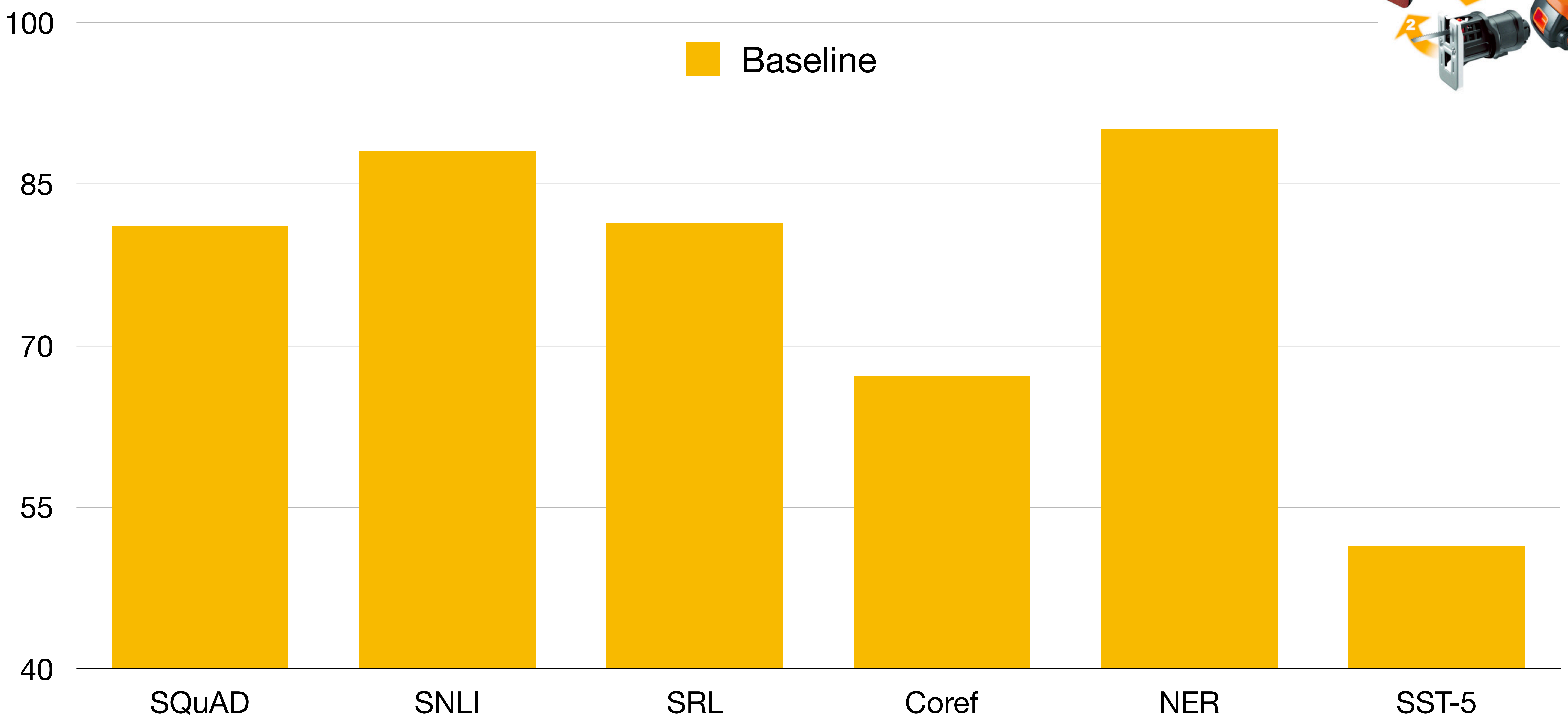
Case Study: ELMo

Train large (~100m-param) forward ***and backward*** deep LSTM language models.



Peters et al. '18

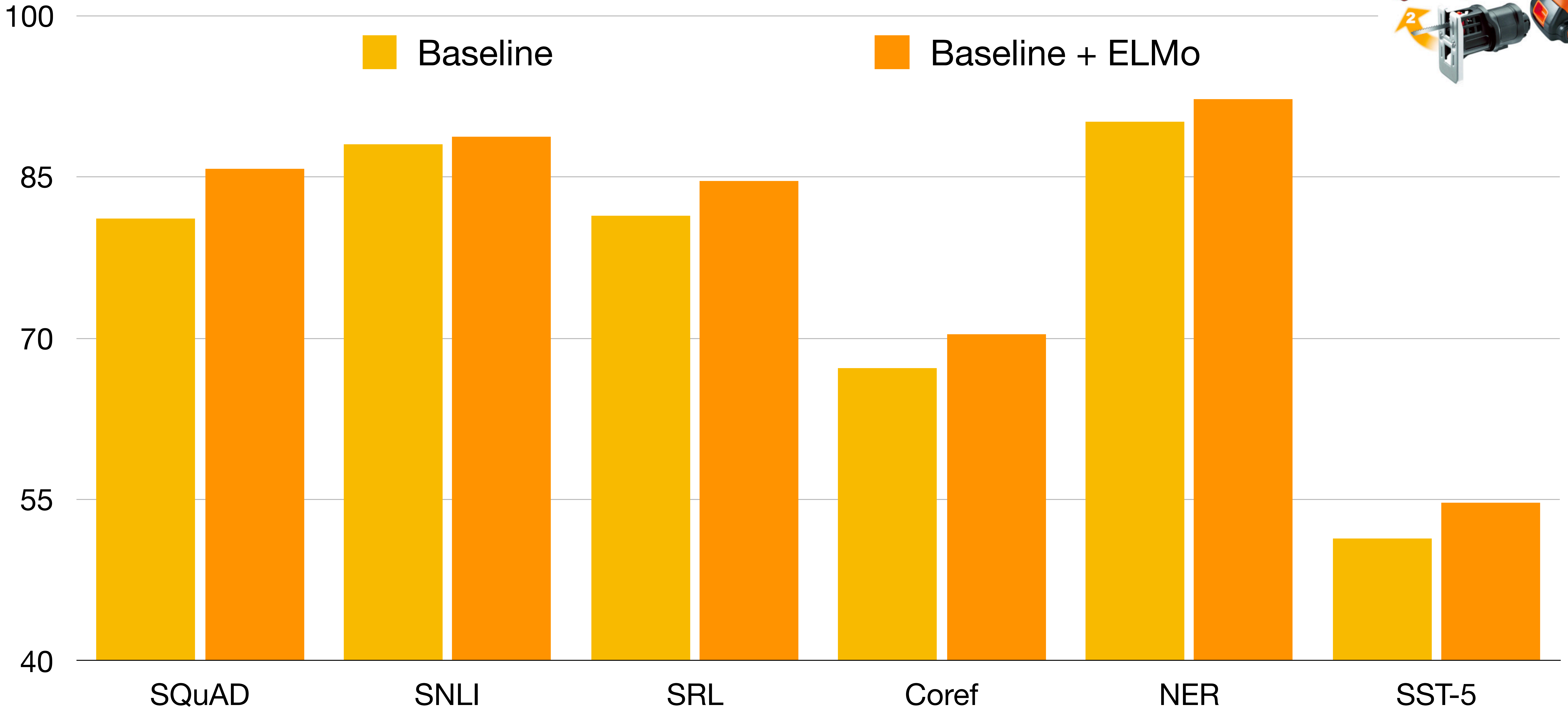
Case Study: ELMo



Peters et al. '18

Case Study: ELMo

Best paper at NAACL 2018!



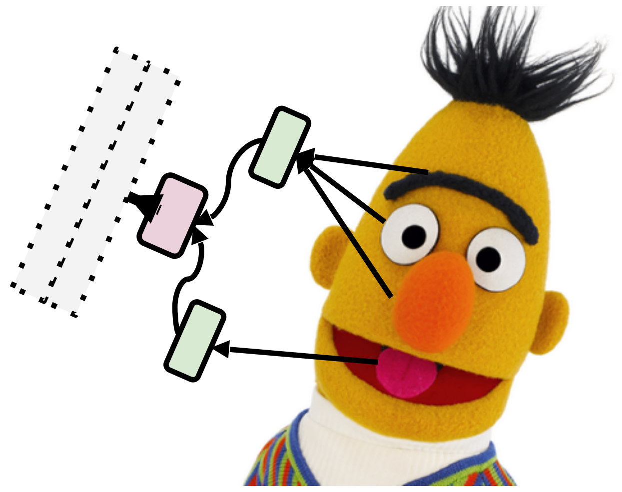
Peters et al. '18

The Rest of the Talk

- The GLUE language understanding benchmark
  Wang et al. '18

  - ...and successes with unsupervised pretraining and fine-tuning on GLUE
    Radford et al. '18 (OpenAI GPT), Devlin et al. '18 (BERT)

- A few things we've learned about modern models
  Tenney et al. '19, Warstadt et al. '19



- Recent progress and the updated SuperGLUE benchmark
  Liu et al. '19a,b, **Nangia & Bowman '19, Wang et al. '19a**

- Easy transfer learning with STILTs

  Phang et al. '19, Wang et al. '19b

GLUE: What is it?

Last Spring: GLUE

The General Language Understanding Evaluation (GLUE):

An open-ended competition and evaluation platform for general-purpose sentence encoders.

Wang, Singh, Michael, Hill, Levy & Bowman '18



GLUE

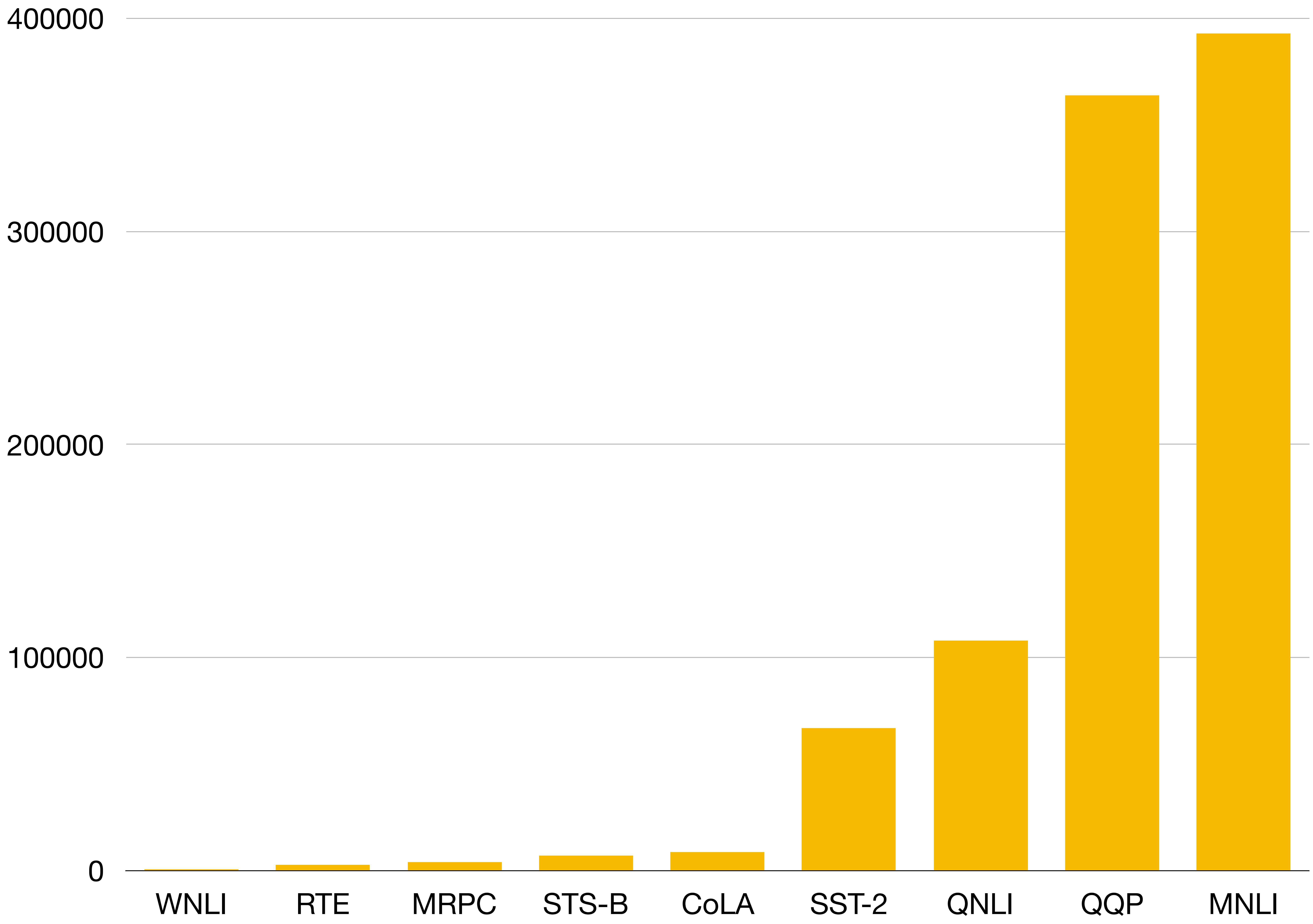

- Nine English-language sentence understanding tasks based on existing data, varying in:
  - Task difficulty
  - Training data volume and degree of training set–test set similarity
  - Language style/genre
- Simple task APIs: All sentence or sentence-pair classification.
- Simple leaderboard API: Upload predictions for a test set (Kaggle-style)
  - Usable with any kind of method/model!

Wang, Singh, Michael, Hill, Levy & Bowman '18

GLUE: The Main Tasks

Corpus	|Train|	|Dev|	|Test|	Task	Metrics	Domain
			Single-Sentence Tasks			
CoLA	8.5k	1k	**1k**	acceptability	Matthews corr.	misc.
SST-2	67k	872	1.8k	sentiment	acc.	movie reviews
			Similarity and Paraphrase Tasks			
MRPC	3.7k	408	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.5k	1.4k	sentence similarity	Pearson/Spearman corr.	misc.
QQP	364k	40k	**391k**	paraphrase	acc./F1	social QA questions
			Inference Tasks			
MNLI	393k	20k	**20k**	NLI	matched acc./mismatched acc.	misc.
QNLI	108k	5.7k	5.7k	QA/NLI	acc.	Wikipedia
RTE	2.5k	276	3k	NLI	acc.	misc.
WNLI	634	71	**146**	coreference/NLI	acc.	fiction books

Wang, Singh, Michael, Hill, Levy & Bowman '18


400000
300000
200000
100000
0
WNLI
RTE
MRPC
STS-B
CoLA
SST-2
QNLI
QQP
MNLI

GLUE: The Main Tasks

Corpus	|Train|	|Dev|	|Test|	Task	Metrics	Domain
				Single-Sentence Tasks		
CoLA	8.5k	1k	**1k**	acceptability	Matthews corr.	misc.
SST-2	67k	872	1.8k	sentiment	acc.	movie reviews
				Similarity and Paraphrase Tasks		
MRPC	3.7k	408	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.5k	1.4k	sentence similarity	Pearson/Spearman corr.	misc.
QQP	364k	40k	**391k**	paraphrase	acc./F1	social QA questions
				Inference Tasks		
MNLI	393k	20k	**20k**	NLI	matched acc./mismatched acc.	misc.
QNLI	108k	5.7k	5.7k	QA/NLI	acc.	Wikipedia
RTE	2.5k	276	3k	NLI	acc.	misc.
WNLI	634	71	**146**	coreference/NLI	acc.	fiction books

Wang, Singh, Michael, Hill, Levy & Bowman '18

The Corpus of Linguistic Acceptability (CoLA)

Warstadt et al. '18

- **Binary classification: Is some string of words a possible English sentence.**
- **Data of this form is a major source of evidence in linguistic theory. Sentences derived from books and articles on morphology, syntax, and semantics.**

* *Who do you think that will question Seamus first?*

✓ *The gardener planted roses in the garden.*

Corpus	|Train|	|Dev|	|Test|	Task	Metrics	Domain
				Single-Sentence Tasks		
CoLA	8.5k	1k	**1k**	acceptability	Matthews corr.	misc.
SST-2	67k	872	1.8k	sentiment	acc.	movie reviews
				Similarity and P[illegible]aph		

Wang, Singh, Michael, Hill, Levy & Bowman '18

Multi-Genre Natural Language Inference (MNLI)

Williams et al. '18

- **Balanced classification for pairs of sentences into *entailment*, *contradiction*, and *neutral***
- **Training set sentences drawn from five written and spoken genres. Dev/test sets divided into a matched set and a *mismatched* set with five more.**

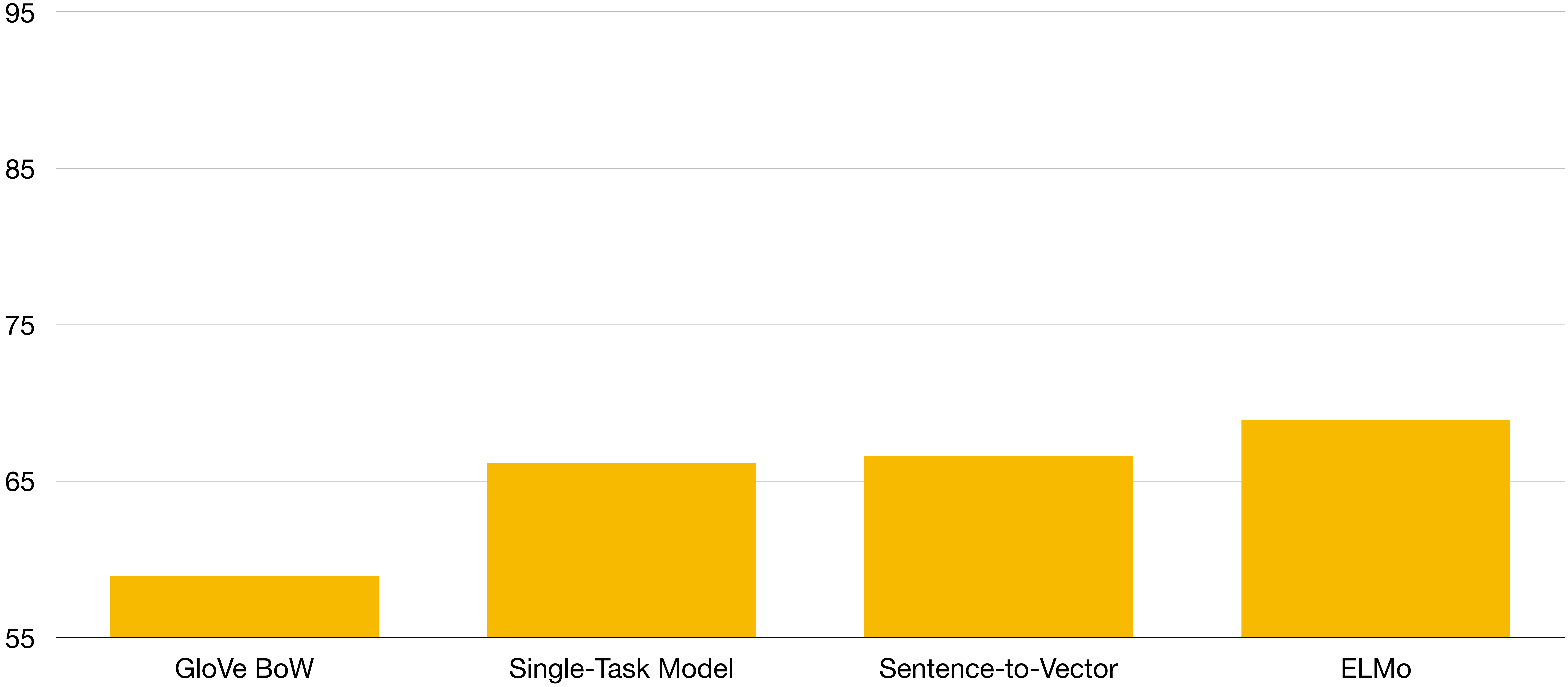
P: *The Old One always comforted Ca'daan, except today.*
H: *Ca'daan knew the Old One very well.*
neutral

Corpus						
CoLA						
SST-2						
MRPC	3.7k	408	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.5k	1.4k	sentence similarity	Pearson/Spearman corr.	misc.
QQP	364k	40k	**391k**	paraphrase	acc./F1	social QA questions
			Inference Tasks			
MNLI	393k	20k	**20k**	NLI	matched acc./mismatched acc.	misc.
QNLI	108k	5.7k	5.7k	QA/NLI	acc.	Wikipedia
RTE	2.5k	276	3k	NLI	ac	
WNLI	634	71	**146**	coreference/NLI	ac	

Wang, Singh, Michael, Hill, Levy & Bowman '18

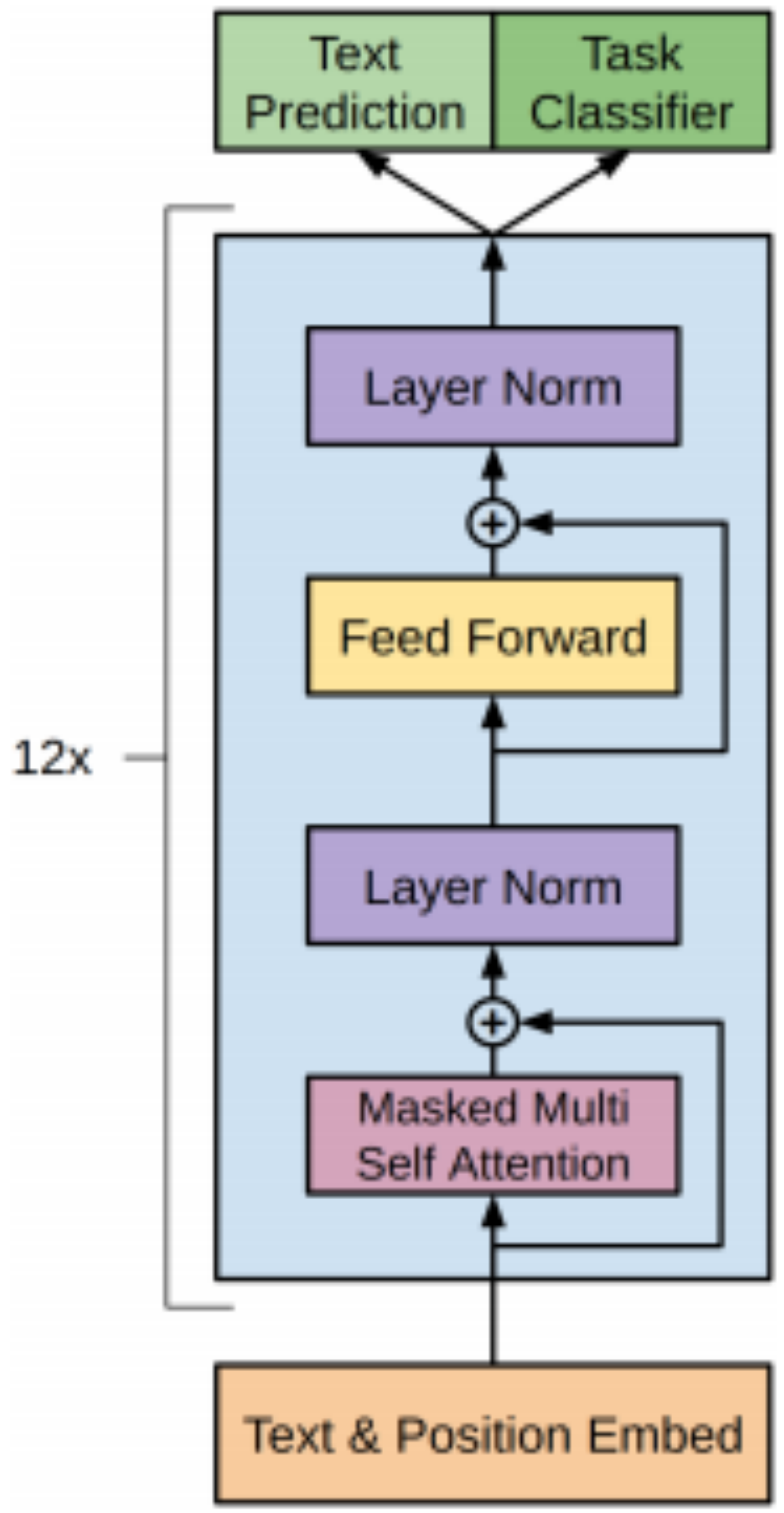
GLUE: What methods work?

Overall GLUE Score



Wang, Singh, Michael, Hill, Levy & Bowman '18

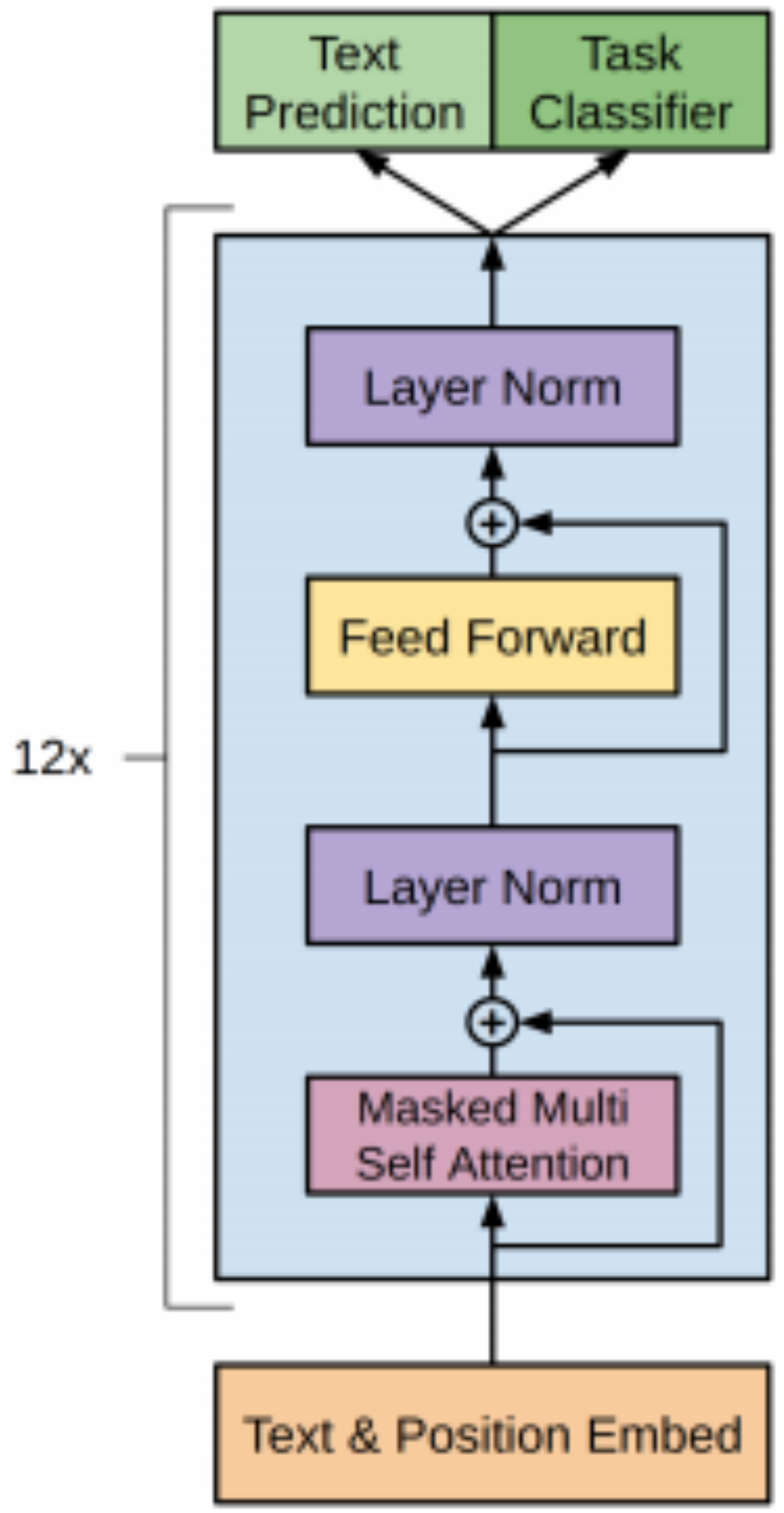
OpenAI’s GPT Language Model



- Same basic idea as ELMo, but many changes, including:
  - *Transformer* encoder architecture.
  - Entire network is *fine-tuned* for each task; few new parameters are added.

Radford et al. '18

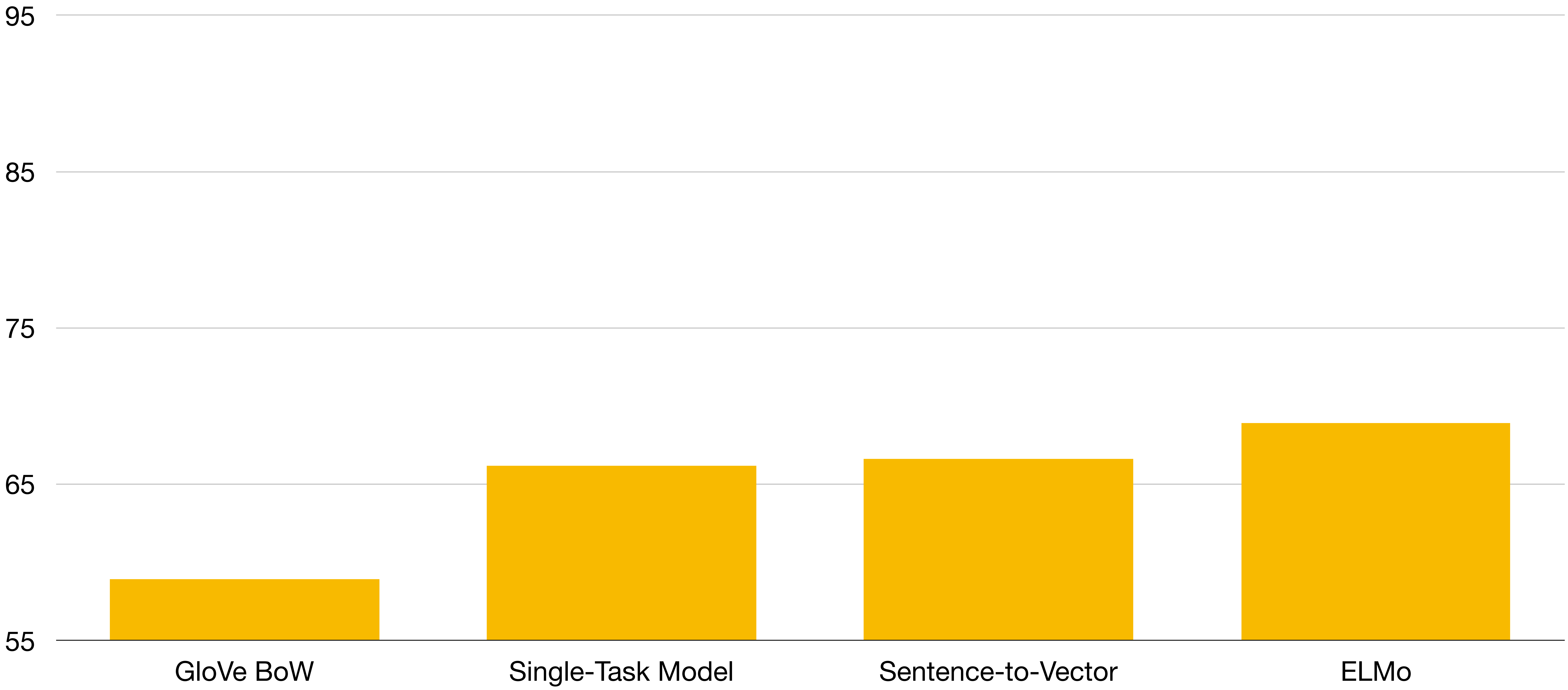
OpenAI's GPT Language Model



- Same basic idea as ELMo, but many changes, including:
  - *Transformer* encoder architecture.
  - Entire network is *fine-tuned* for each task; few new parameters are added.
  - Pretraining is on long spans of running text, not just isolated sentences.

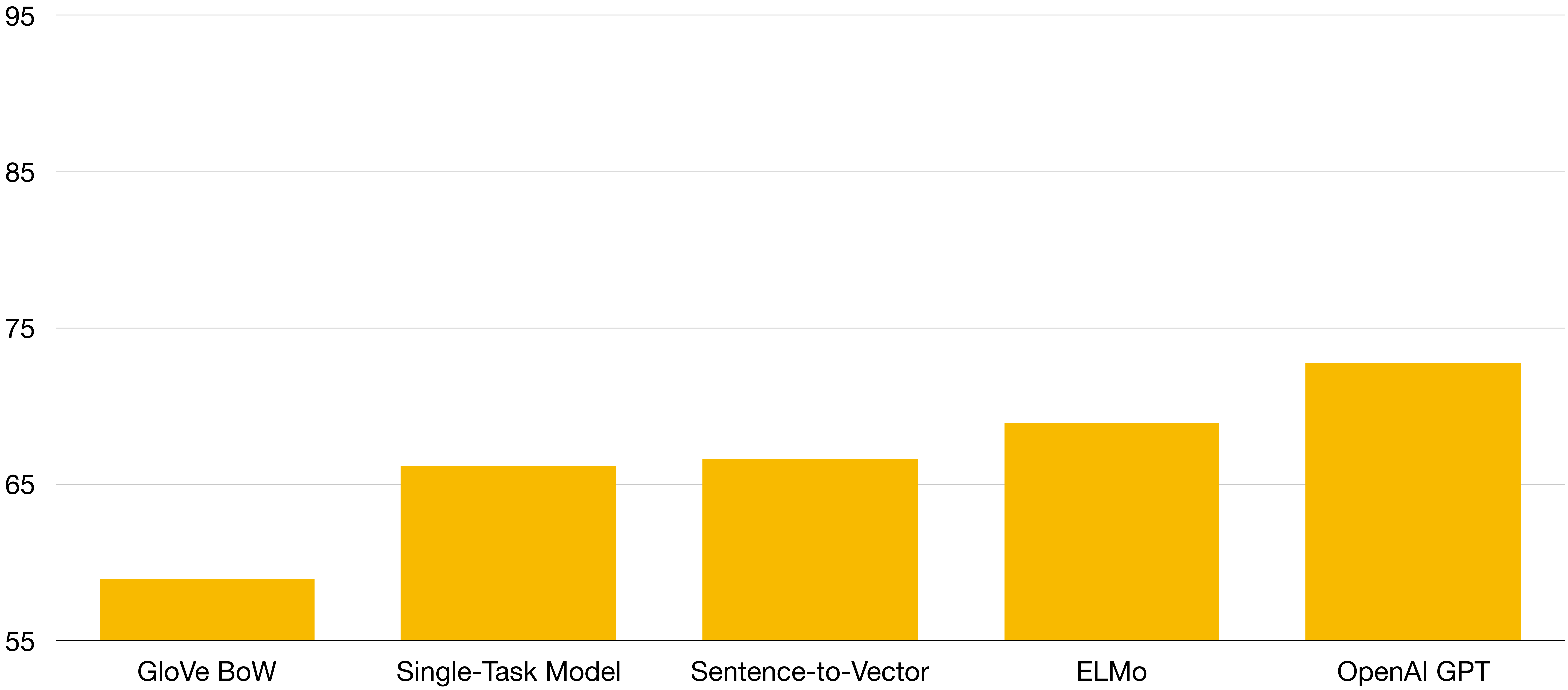
Radford et al. '18

GLUE Score



Radford et al. '18

GLUE Score



Radford et al. '18

Google's BERT

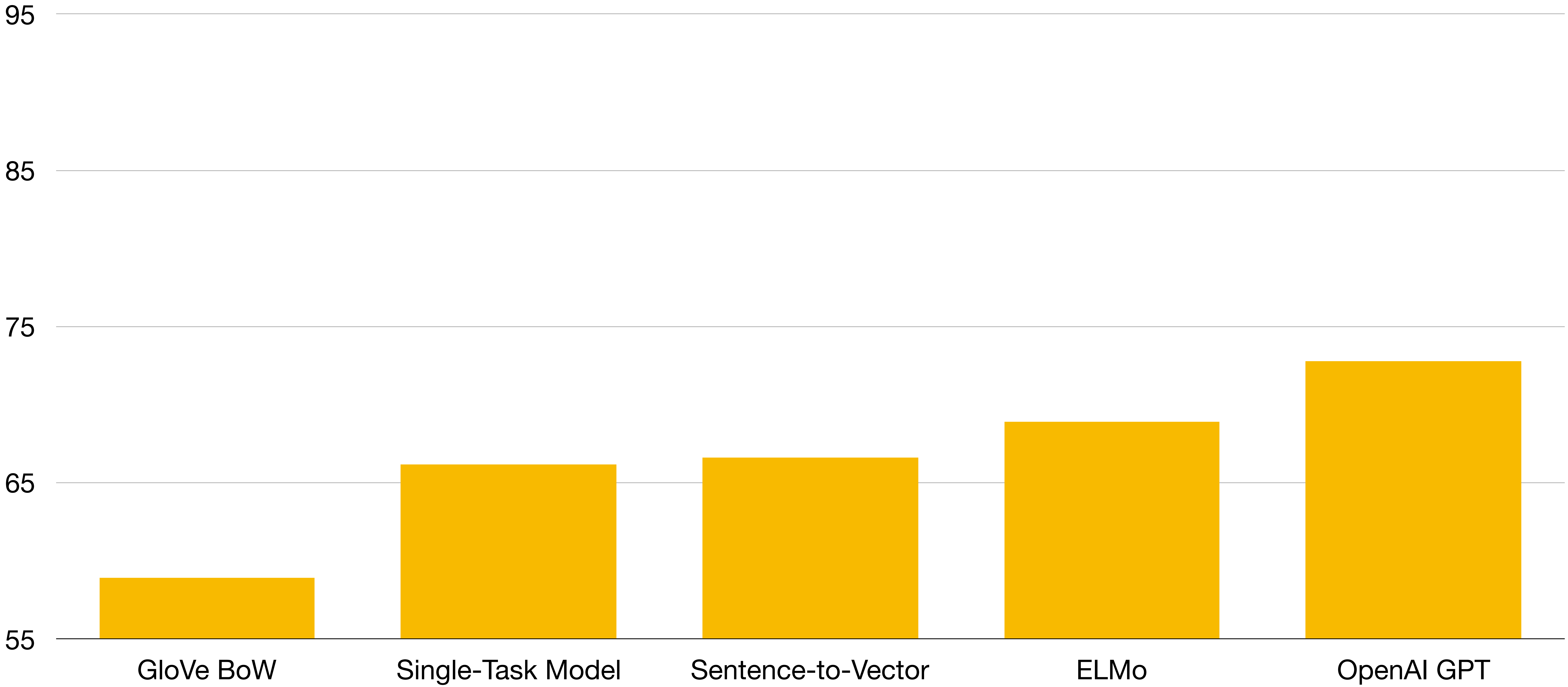
Devlin et al. '18
see Baevski et al. '19 for similar concurrent work

The BERT Model

- Same basic idea as GPT with several changes, including:
  - Two different unlabeled data tasks in place of language modeling.
    - These allow the model to process both directions together with the same network at training time.
  - Bigger (100M => 300M params).

Devlin et al. '18
see Baevski et al. '19 for similar concurrent work

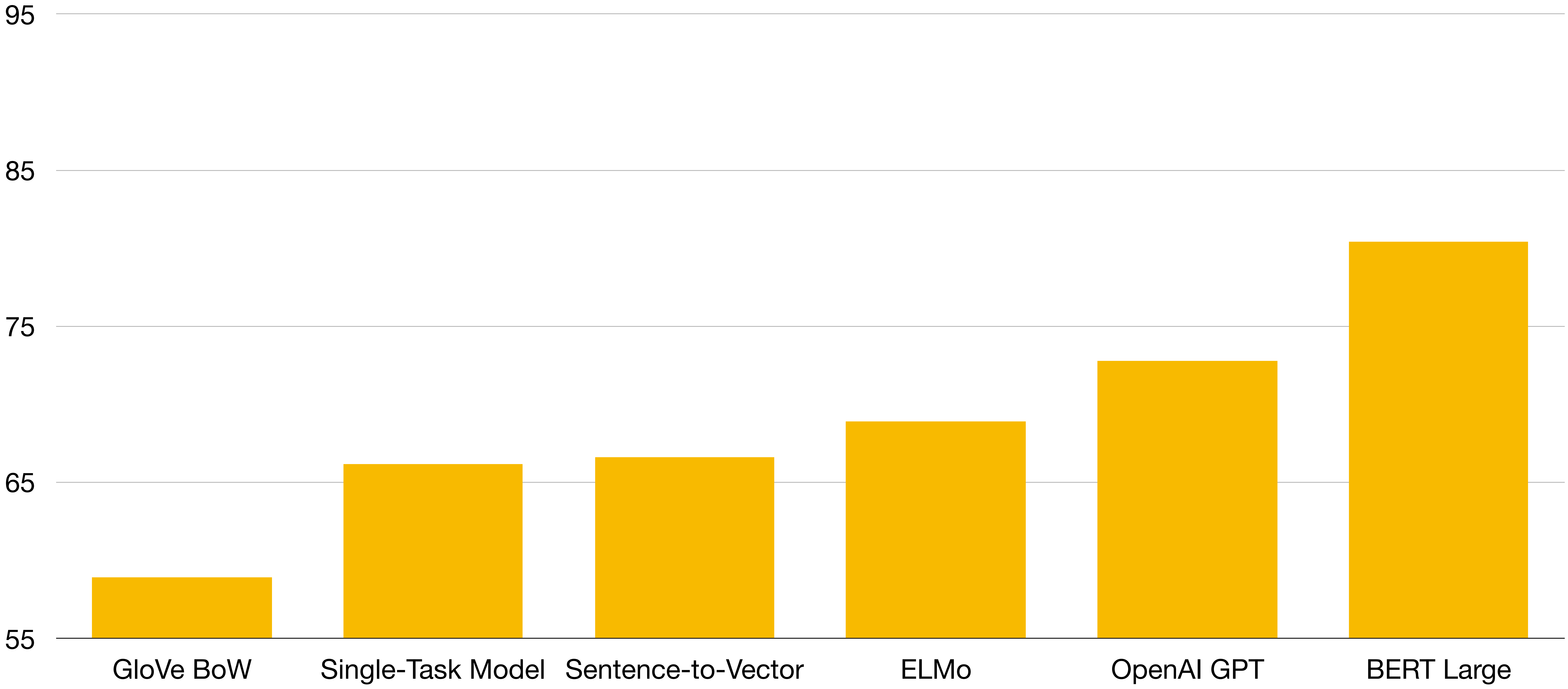
GLUE Score



Devlin et al. '18

see Baevski et al. '19 for similar concurrent work

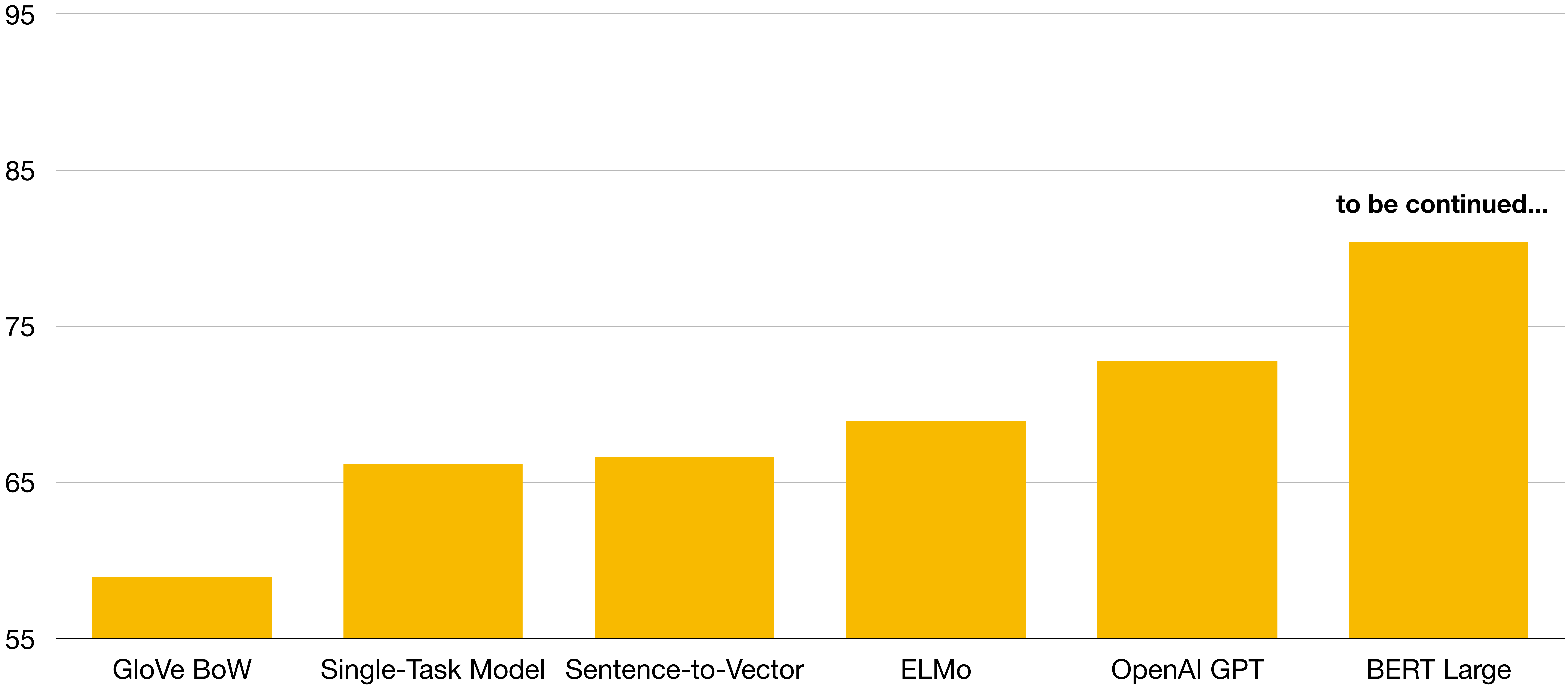
GLUE Score



Devlin et al. '18

see Baevski et al. '19 for similar concurrent work

GLUE Score

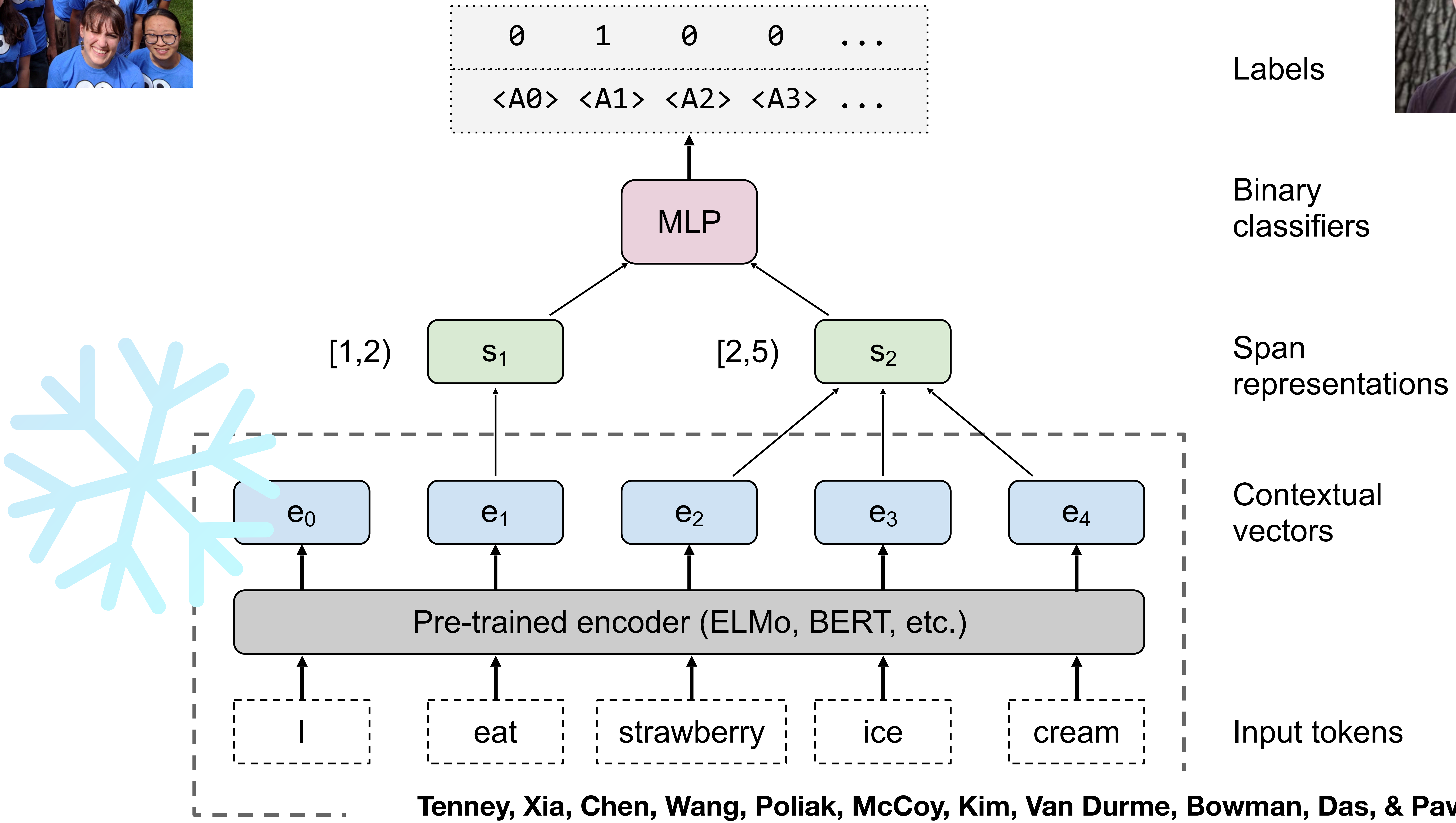


Devlin et al. '18

see Baevski et al. '19 for similar concurrent work

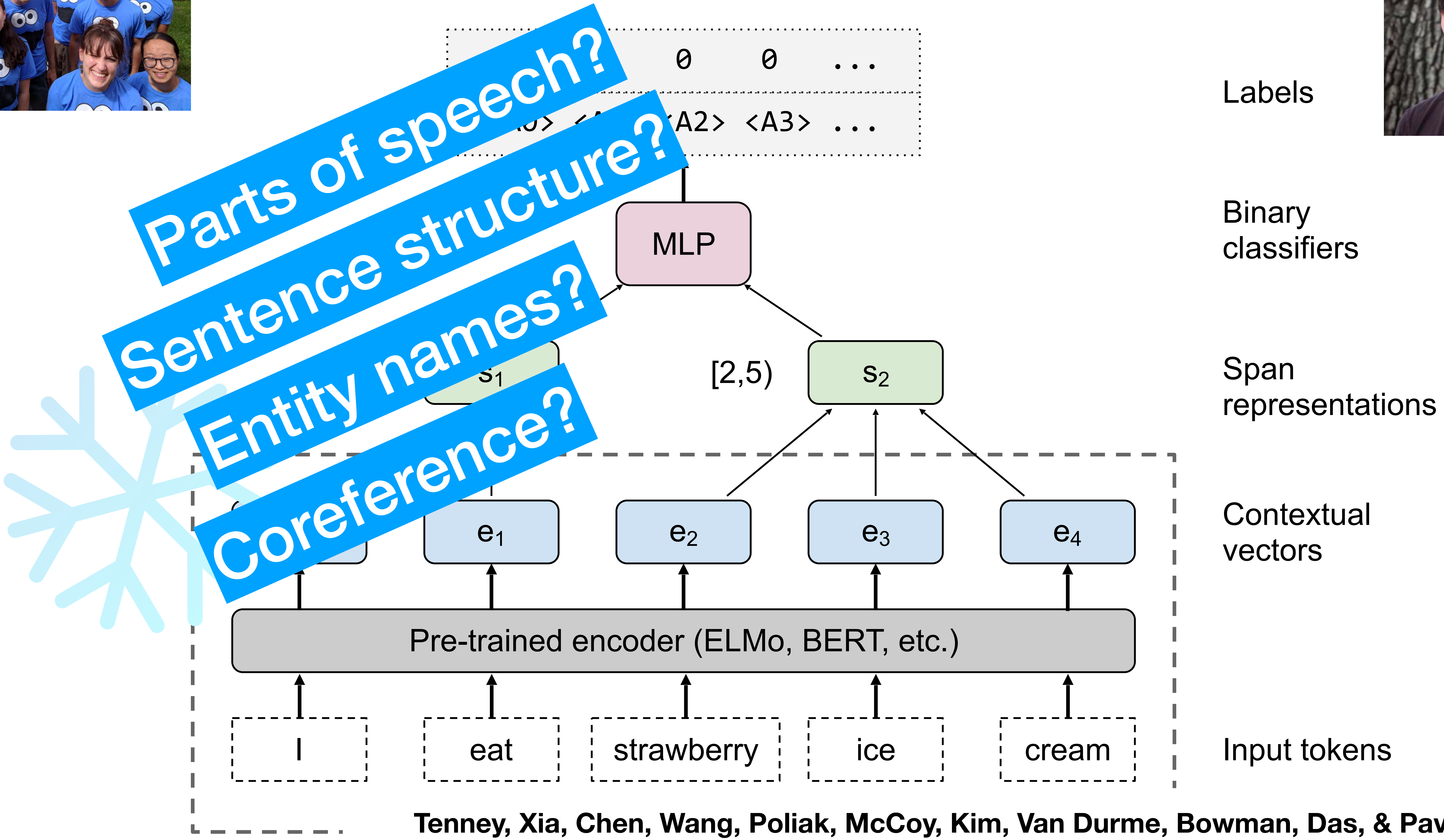
Why does BERT work so well?
What does BERT know?

Edge Probing



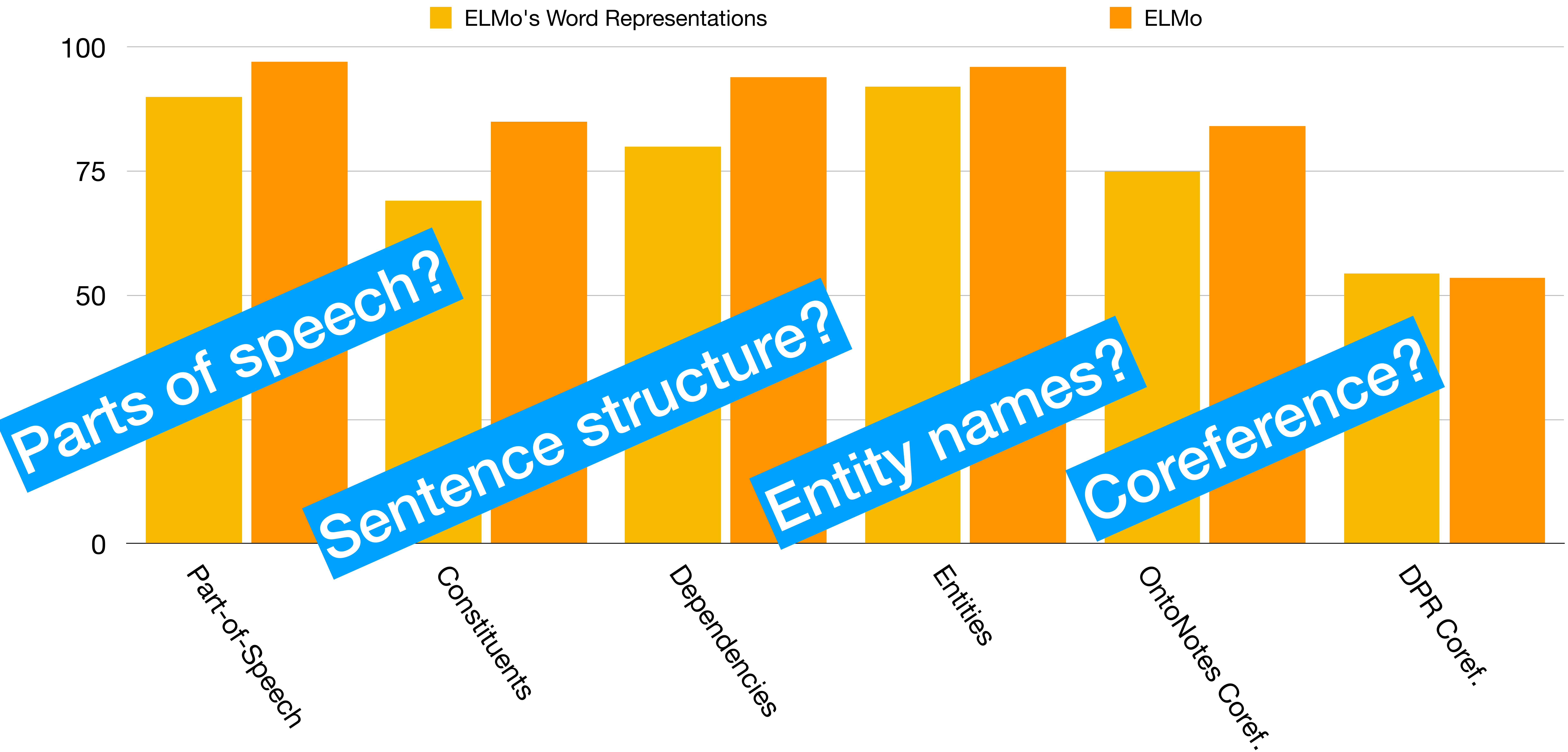
Tenney, Xia, Chen, Wang, Poliak, McCoy, Kim, Van Durme, Bowman, Das, & Pavlick '19

Edge Probing

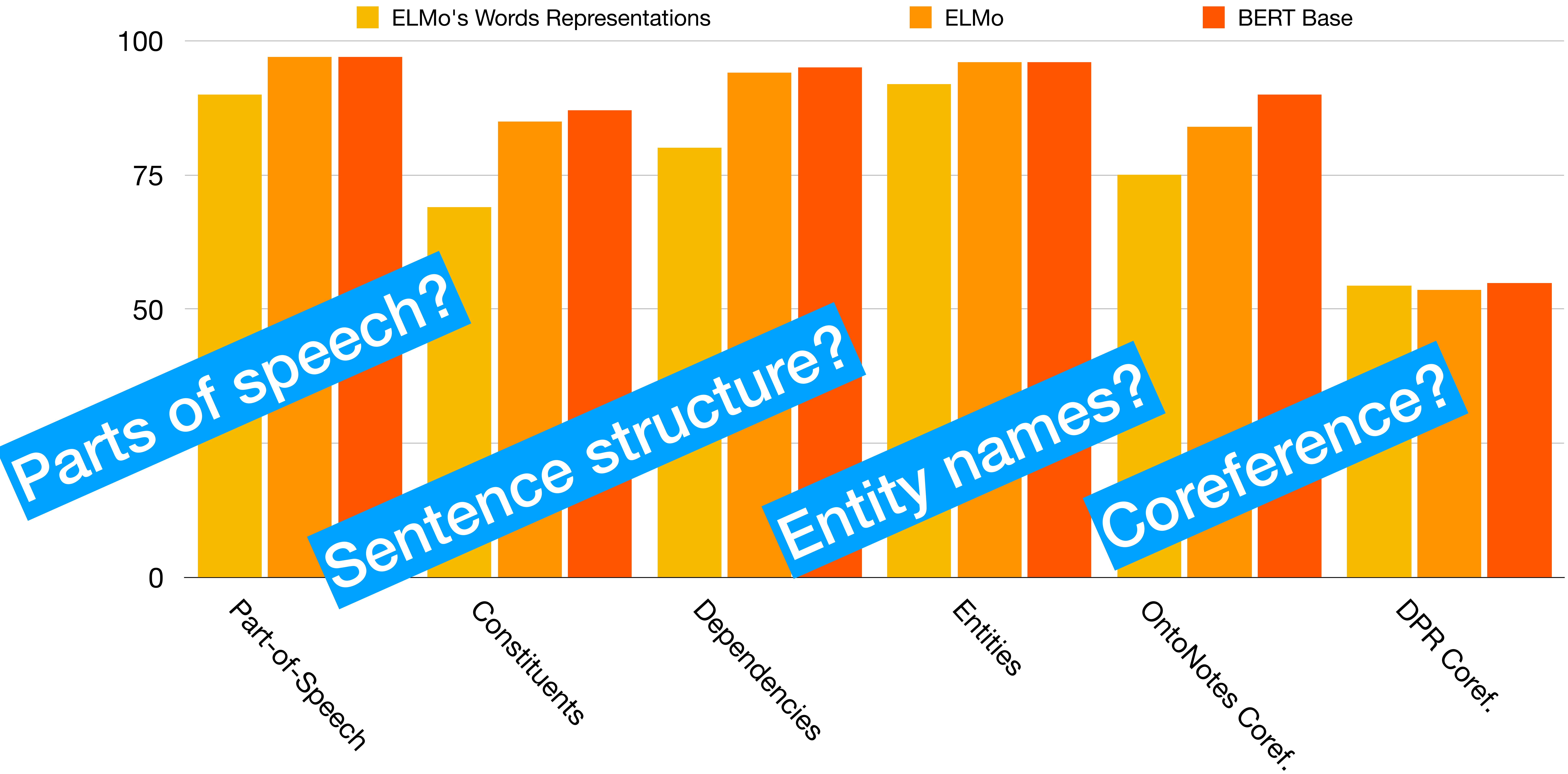


Tenney, Xia, Chen, Wang, Poliak, McCoy, Kim, Van Durme, Bowman, Das, & Pavlick '19

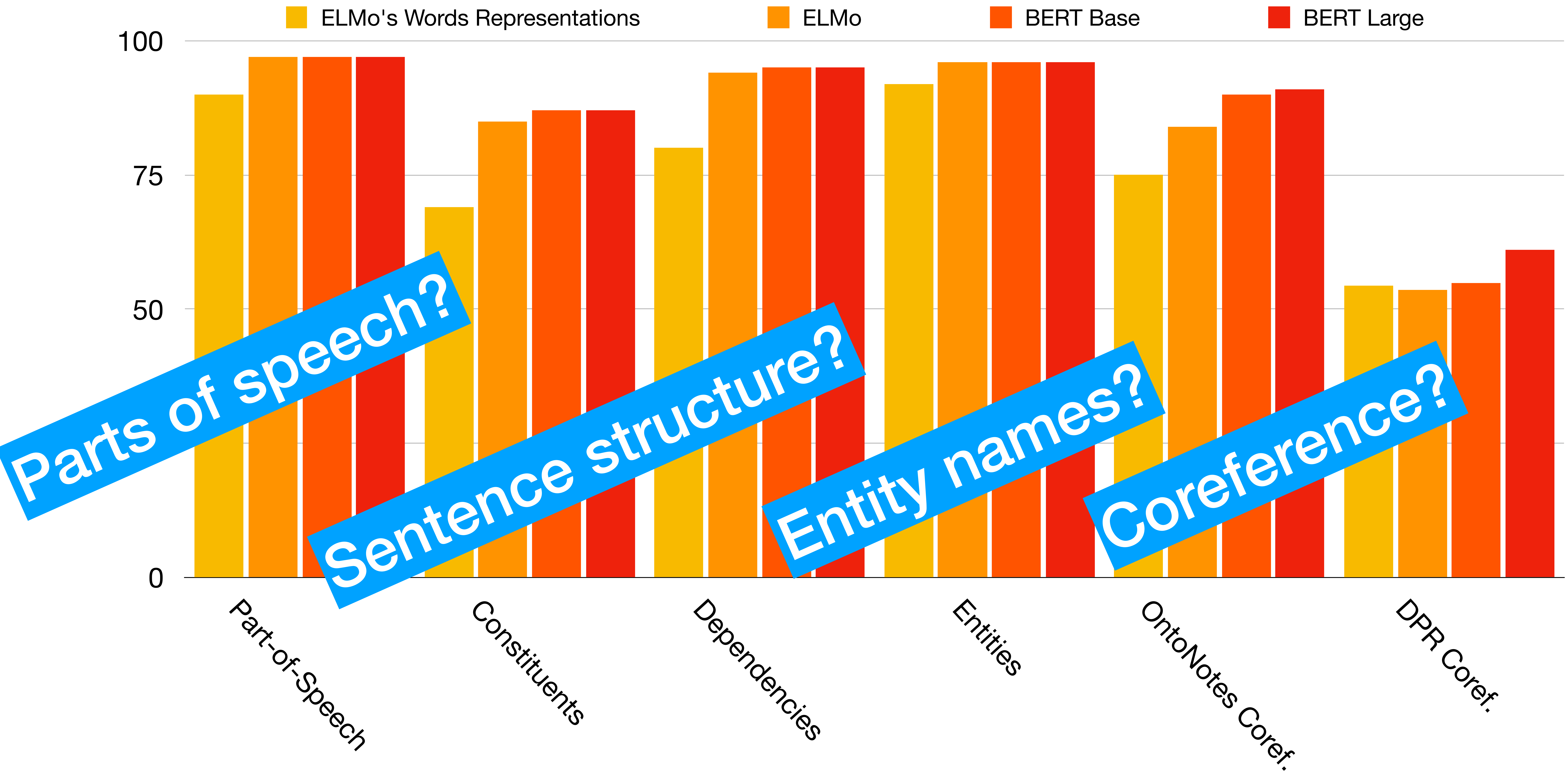
Edge Probing with ELMo



Edge Probing with ELMo and BERT



Edge Probing with ELMo and BERT



How much can we trust our conclusions?

How much can we trust these conclusions?

- Studies like ours that use auxiliary analysis datasets are a <u>common tool</u> for trying to understand what models like BERT know.
- There are many ways to design such a study, and each bakes in a few substantial assumptions.
    - Edge probing assumes that if a model *knows* about coreference, then it should be possible to extract that information with a simple MLP model.
- *Do different probing methods give us the same answer?*

{Warstadt, Cao, Grosu, Peng, Blix, Nie, Alsop, Bordia, Liu, Parrish, Wang, Phang, Mohananey, Htut, Jeretič} & Bowman ‘19

Case Study: NPI Licensing

- *NPI* words like *any* or *ever* can only occur in the scope of specific linguistic *licensing environments* like negations or conditionals,
  - Common in natural data.
  - Well-characterized in the linguistics literature.
  - Depends on long-distance dependencies and complex structures, rather than local co-occurrence.
  - Should be learnable from raw text alone.
- *Does BERT know when NPIs are licensed?*

(1) Mary has**n't** eaten *any* cookies.

(2) *Mary has eaten *any* cookies.

{Warstadt, Cao, Grosu, Peng, Blix, Nie, Alsop, Bordia, Liu, Parrish, Wang, Phang, Mohananey, Htut, Jeretič} & Bowman '19

Case Study: NPI Licensing

- *NPI* words like *any* or *ever* can only occur in the scope of specific linguistic *licensing environments* like negations or conditionals,
  - Common in natural data.
  - Well-characterized in the linguistics literature.
  - Depends on long-distance dependencies and complex structures, rather than local co-occu
  - Should be learnable from
- *Does BERT know when NPIs are licensed?*

(1) Mary has**n't** eaten *any* cookies.

(2) *Mary has eaten *any* cookies.

Let's ask this as many ways as we can!

{Warstadt, Cao, Grosu, Peng, Blix, Nie, Alsop, Bordia, Liu, Parrish, Wang, Phang, Mohananey, Htut, Jeretič} & Bowman '19

Case Study: NPI Licensing

- Evaluation data: Nine custom NPI test sets isolating different NPI licensors:

*Those boys say **that** [the doctors *ever* went to an art gallery.]
*Those boys *ever* say **that** [the doctors went to an art gallery.]
Those boys say **that** [the doctors *often* went to an art gallery.]
Those boys *often* say **that** [the doctors went to an art gallery.]

{Warstadt, Cao, Grosu, Peng, Blix, Nie, Alsop, Bordia, Liu, Parrish, Wang, Phang, Mohananey, Htut, Jeretič} & Bowman '19

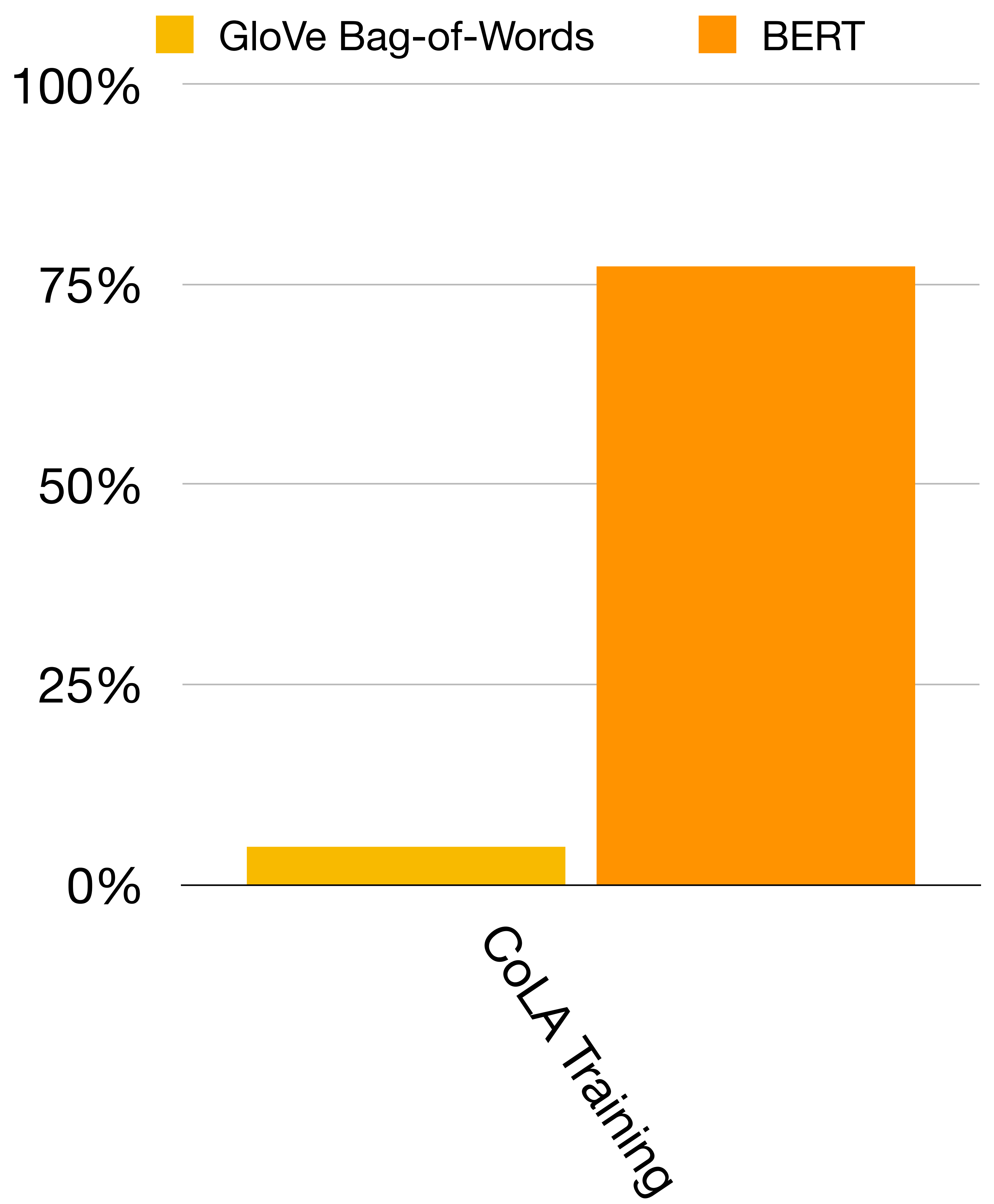
Let's teach the model to judge acceptability.

*Who do you think that will question Seamus first?
*Usually, any lion is majestic.
The gardener planted roses in the garden.
I wrote Blair a letter, but I tore it up before I sent it.

Train:
The CoLA general acceptability corpus

Test:
NPI environment test sets

Metric:
Matthews Correlation (MCC) for acceptability



Let's teach the model to judge acceptability.

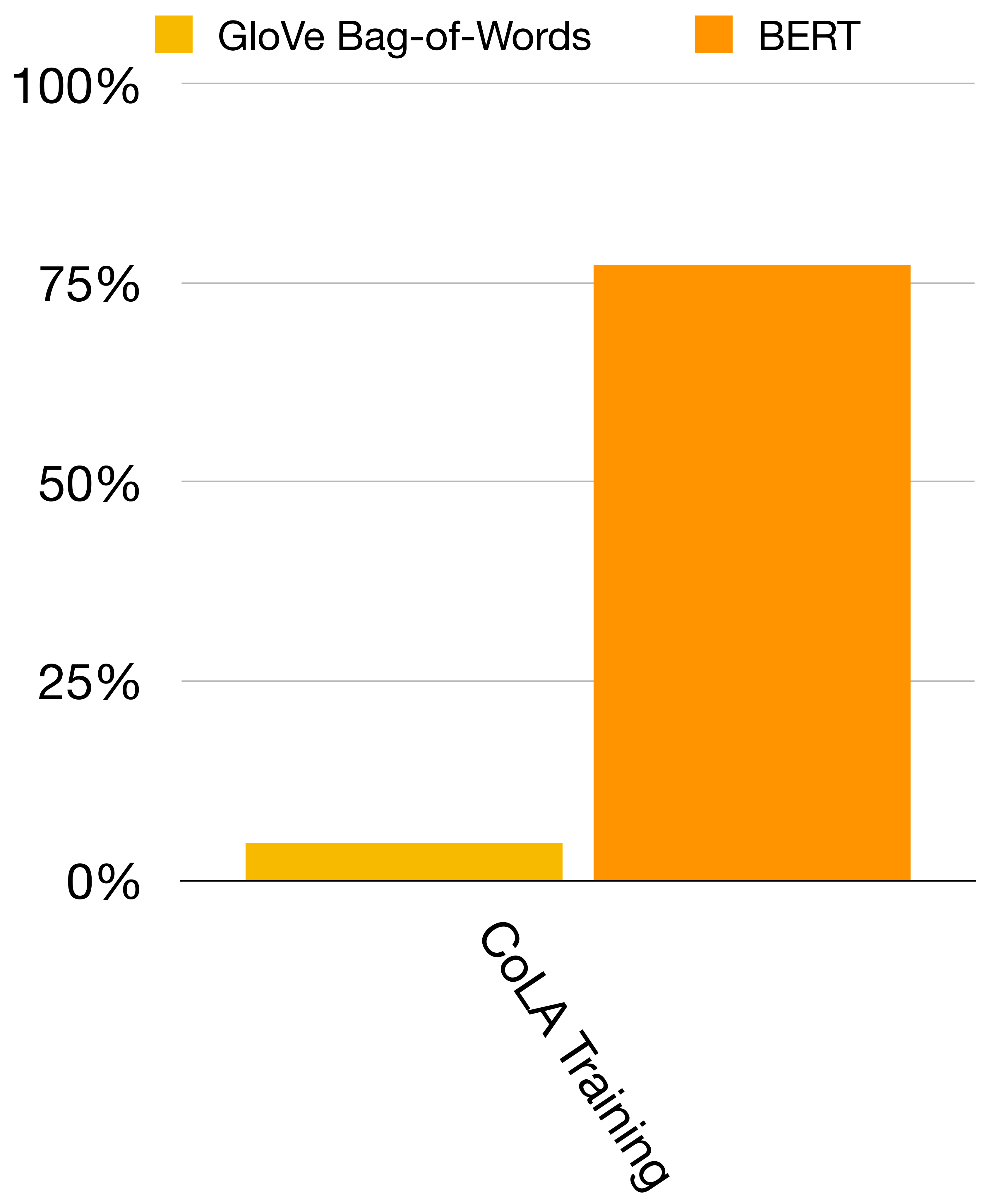
BERT knows a bit about NPIs,
but its not perfect.

*Wh…
*Usually, any lion is majest…
The gardener planted r… the garden.
I wrote Blair a letter … tore it up before I sent it.

Train:
The CoLA general acceptability corpus

Test:
NPI environment test sets

Metric:
Matthews Correlation (MCC) for acceptability



What if we train on NPI data directly?

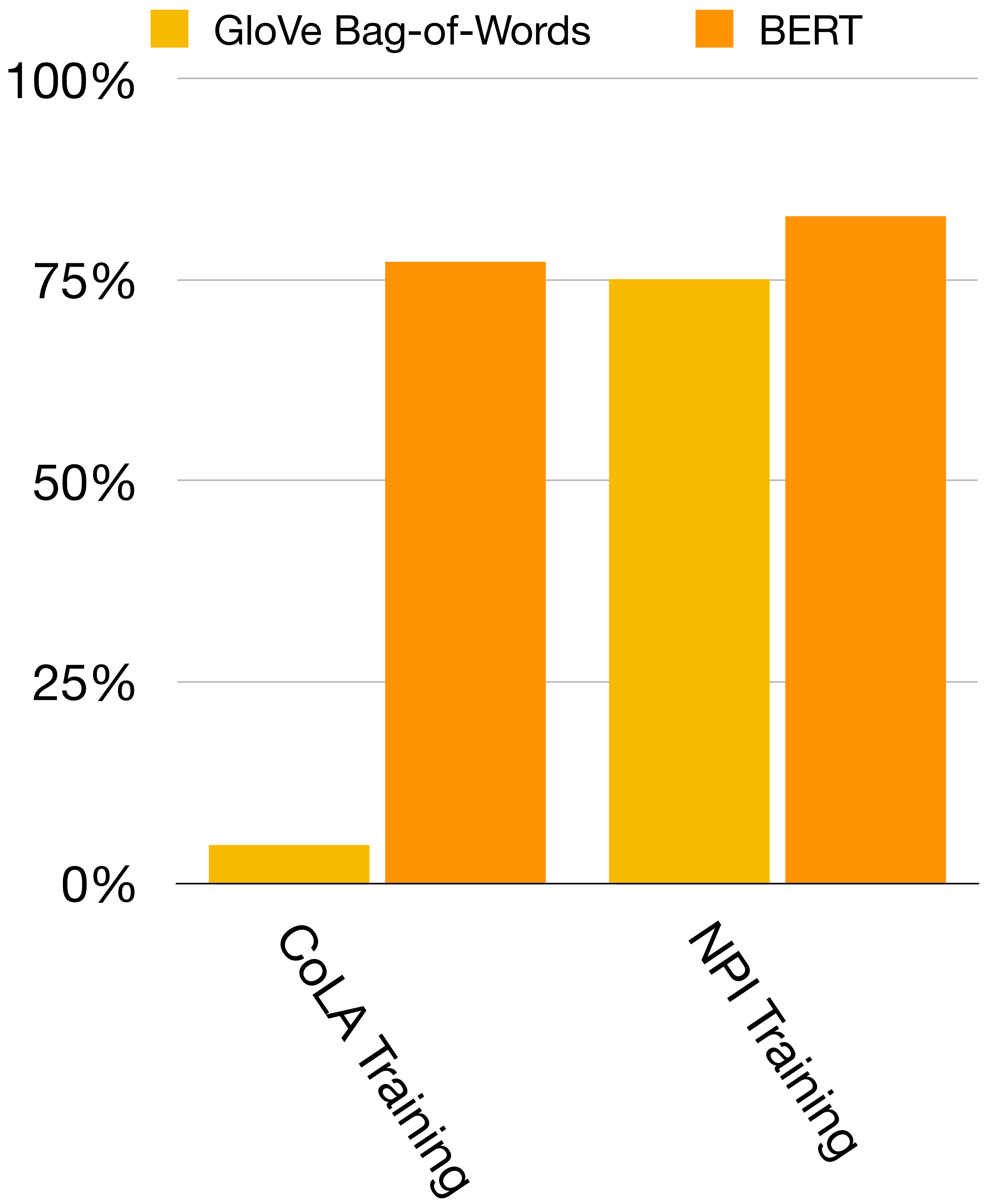
*Those boys say **that** [the doctors *ever* went to an art gallery.]
*Those boys *ever* say **that** [the doctors went to an art gallery.]
Those boys say **that** [the doctors *often* went to an art gallery.]
Those boys *often* say **that** [the doctors went to an art gallery.]

*Who do you think that will question Seamus first?
*Usually, any lion is majestic.
The gardener planted roses in the garden.
I wrote Blair a letter, but I tore it up before I sent it.

Train:
NPI training set (hold-one-out by environment) or the CoLA general acceptability corpus

Test:
NPI environment test sets

Metric:
Matthews Correlation (MCC) for acceptability



What if we train on NPI data directly?

*Those boys say **that** [the doctors *ever* went to an art gallery.]
gallery.]
allery.]
allery.]

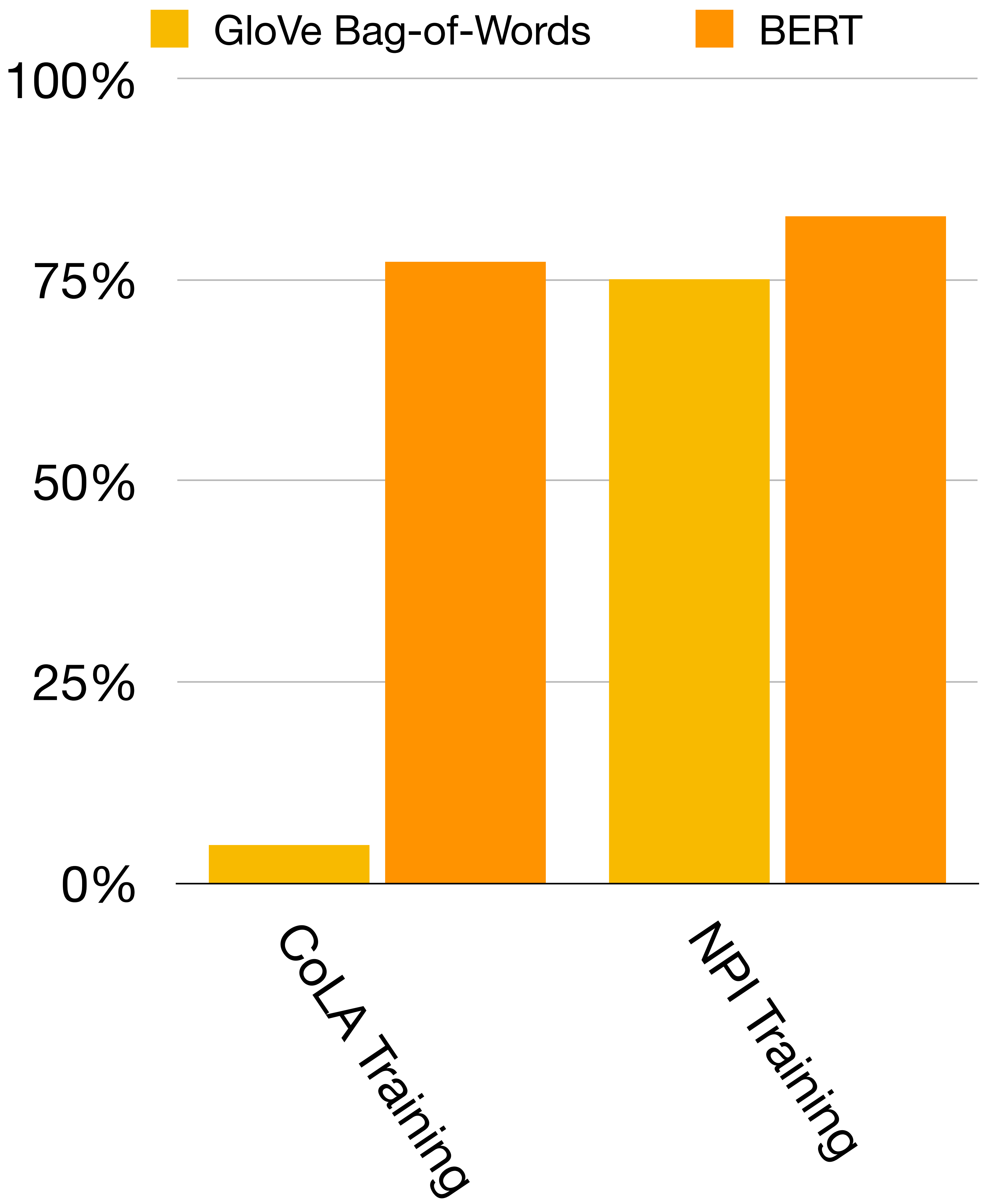
*Wh
*Usually, any lion is majest
The gardener planted r the garden.
I wrote Blair a letter tore it up before I sent it.

BERT knows something about NPIs, but not all that much.

Train:
NPI training set (hold-one-out by environment) or the CoLA general acceptability corpus

Test:
NPI environment test sets

Metric:
Matthews Correlation (MCC) for acceptability



Let's re-structure our data to isolate BERT's knowledge of NPIs...

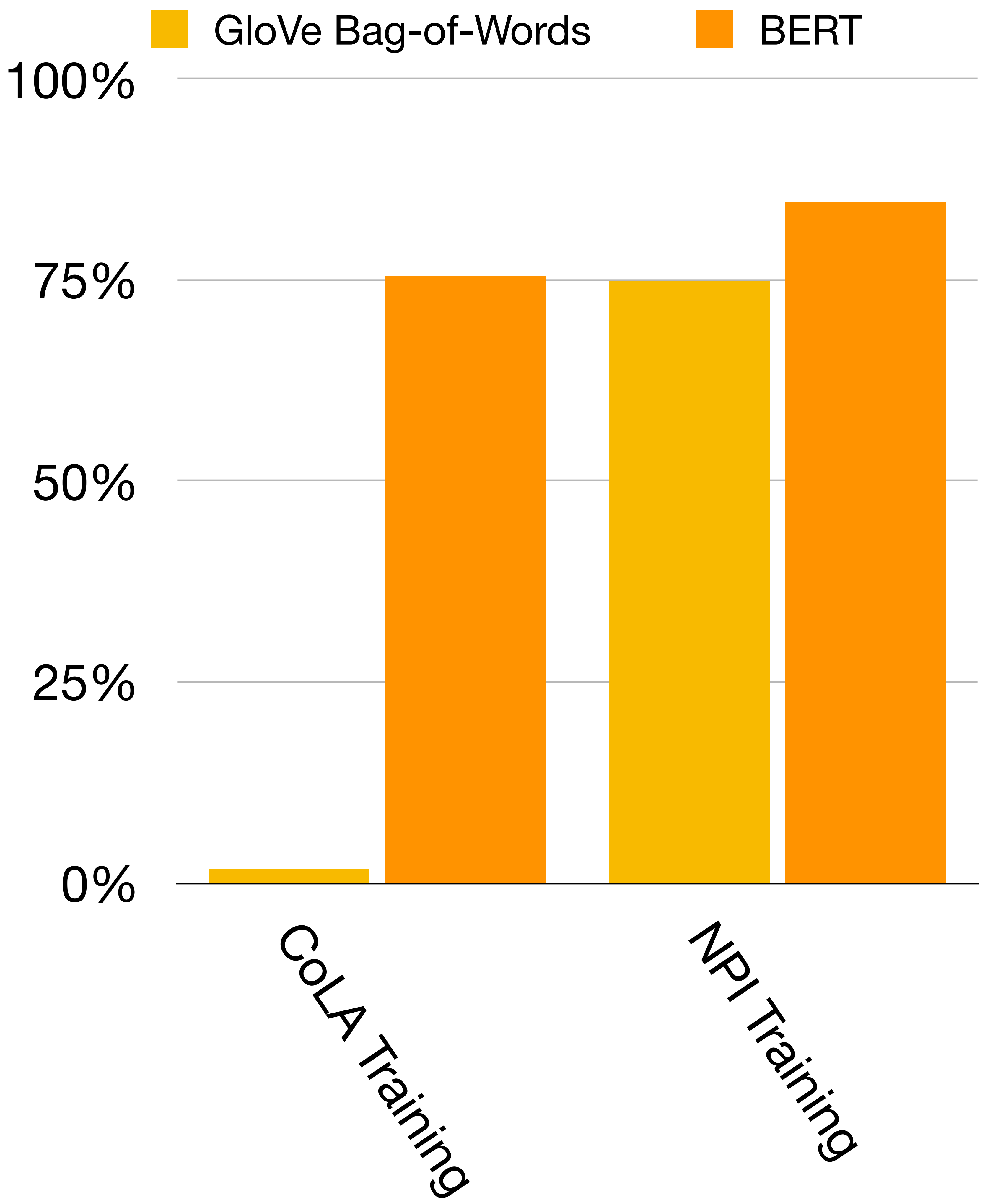
(1) Mary has**n't** eaten *any* cookies.

(2) *Mary has eaten *any* cookies.

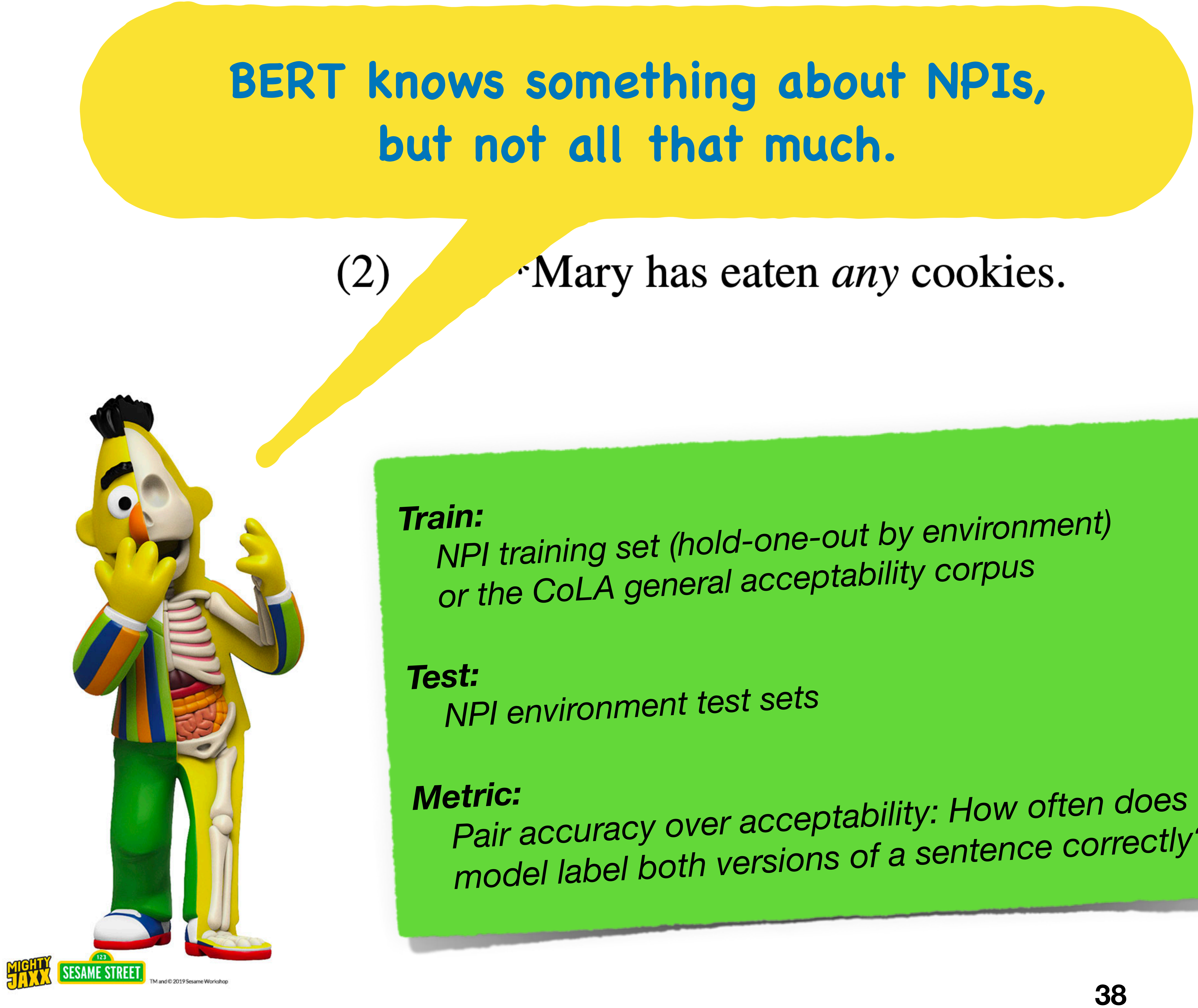
Train:
NPI training set (hold-one-out by environment) or the CoLA general acceptability corpus

Test:
NPI environment test sets

Metric:
Pair accuracy over acceptability: How often does the model label both versions of a sentence correctly?



Let's re-structure our data to isolate BERT's knowledge of NPIs...

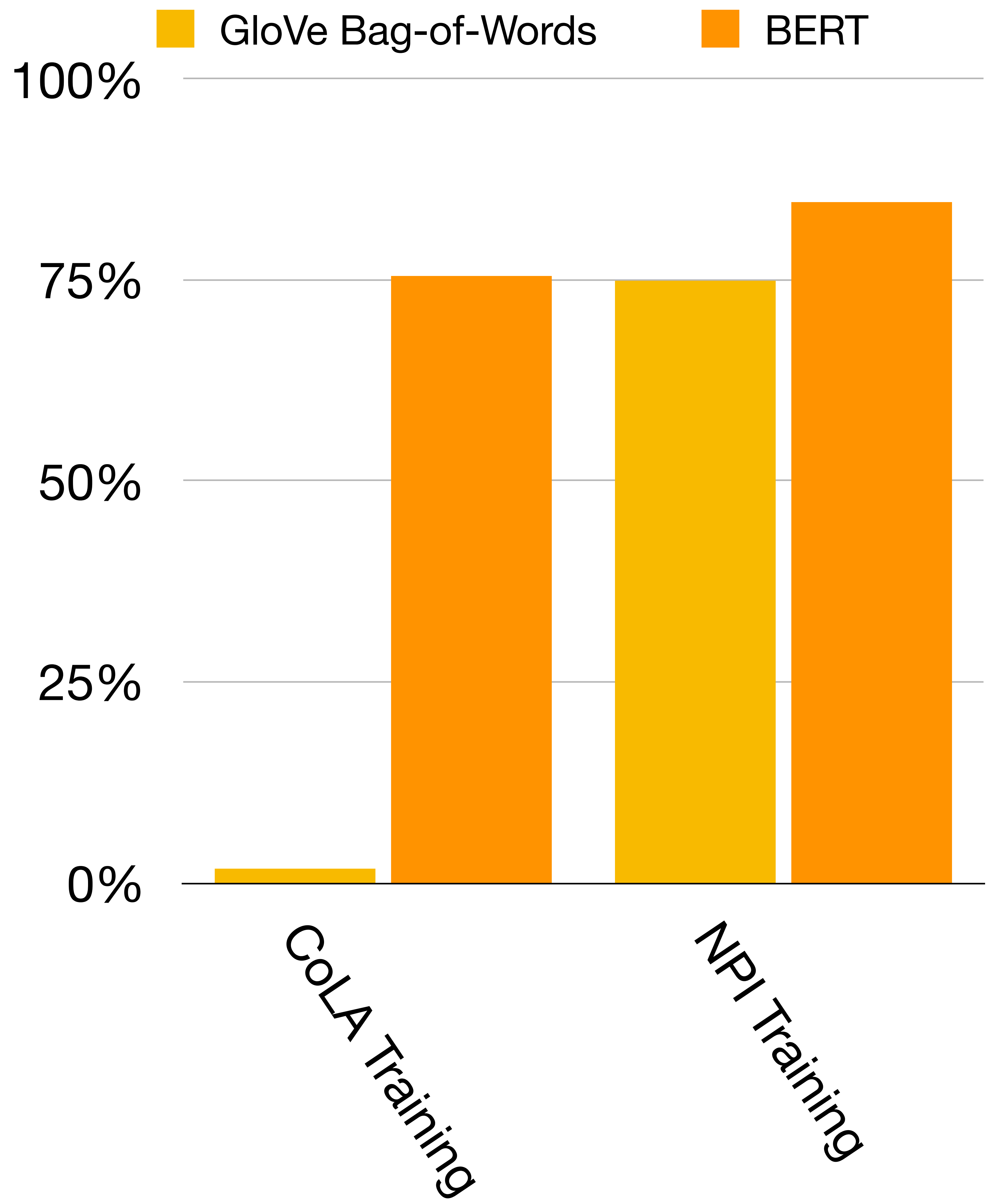


(2) *Mary has eaten *any* cookies.

Train:
NPI training set (hold-one-out by environment) or the CoLA general acceptability corpus

Test:
NPI environment test sets

Metric:
Pair accuracy over acceptability: How often does the model label both versions of a sentence correctly?



Let's re-structure our data to isolate BERT's knowledge of NPIs…

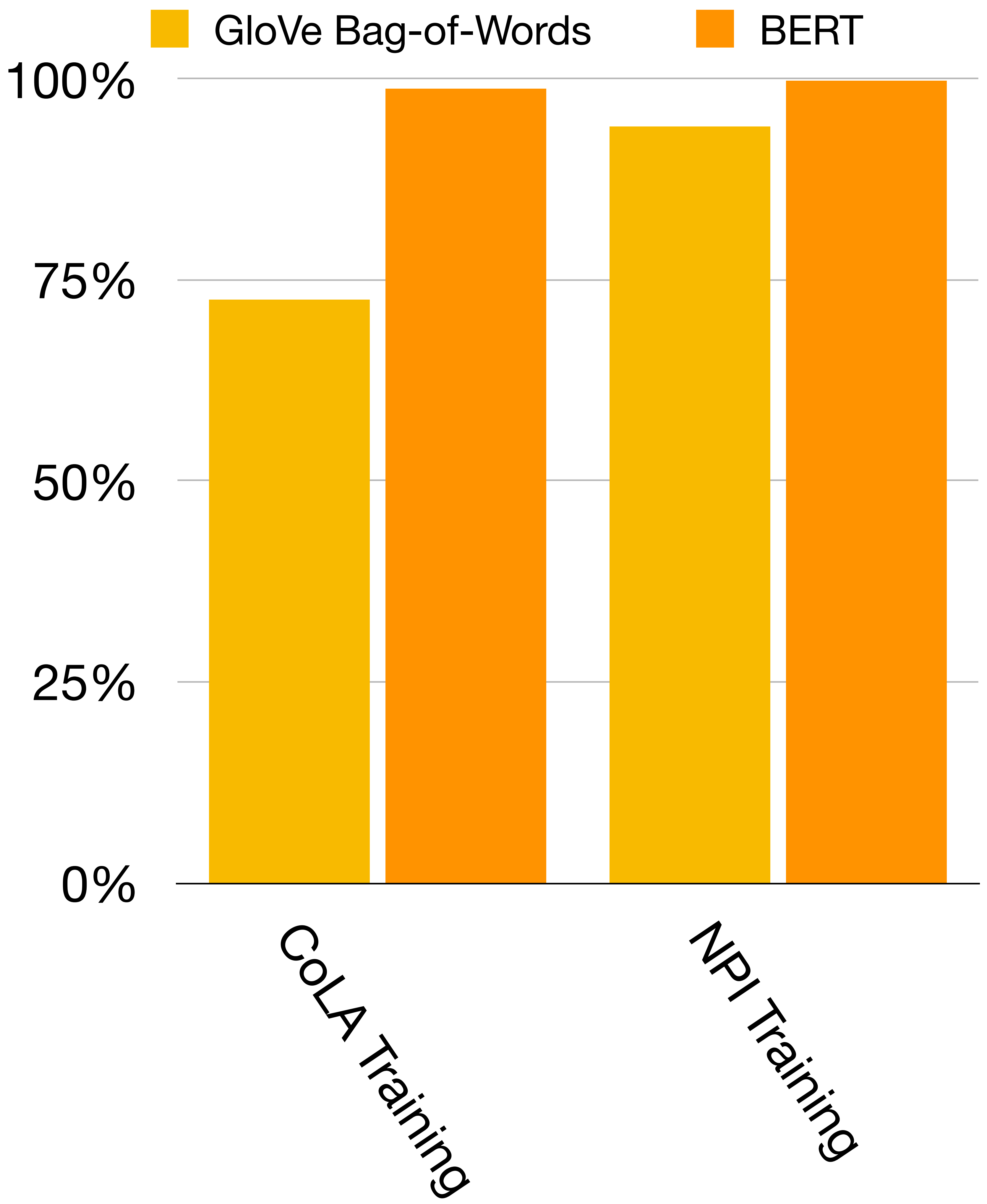
(1) Mary hasn**'t** eaten *any* cookies.

(2) *Mary has eaten *any* cookies.

Train:
NPI training set (hold-one-out by environment) or the CoLA general acceptability corpus

Test:
NPI environment test sets

Metric:
Pair preference accuracy: How often does the model assign a higher probability of acceptability to the correct sentence?



Let’s re-structure our data to isolate BERT’s knowledge of NPIs...

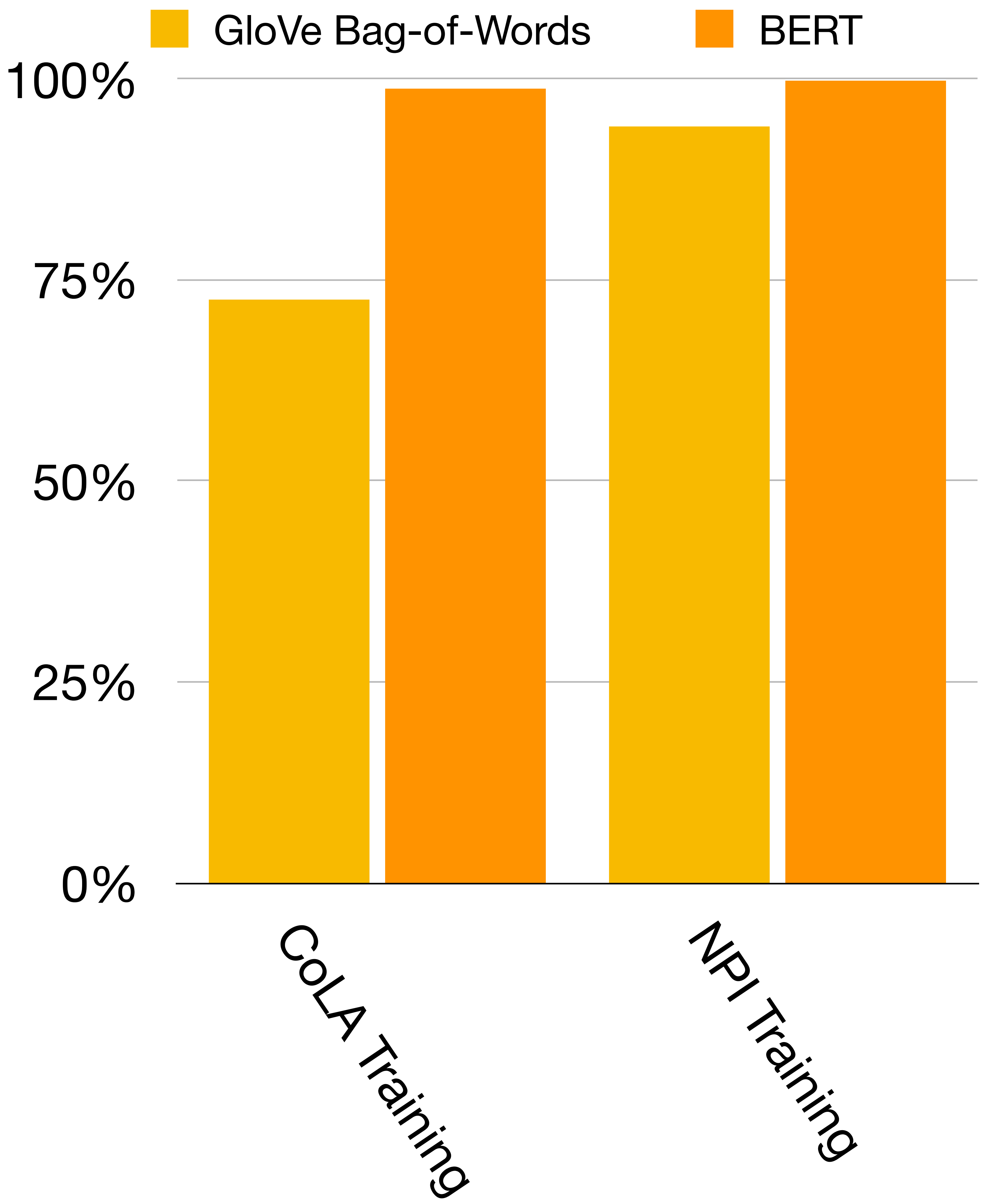
BERT has complete and perfect knowledge of NPI licensing.

(2) *Mary has eaten *any* cookies.

Train:
NPI training set (hold-one-out by environment) or the CoLA general acceptability corpus

Test:
NPI environment test sets

Metric:
Pair preference accuracy: How often does the model assign a higher probability of acceptability to the correct sentence?



What if we ask BERT directly?

(1) Mary hasn**'t** eaten *any* cookies.

(2) *Mary has eaten *any* cookies.

Train:

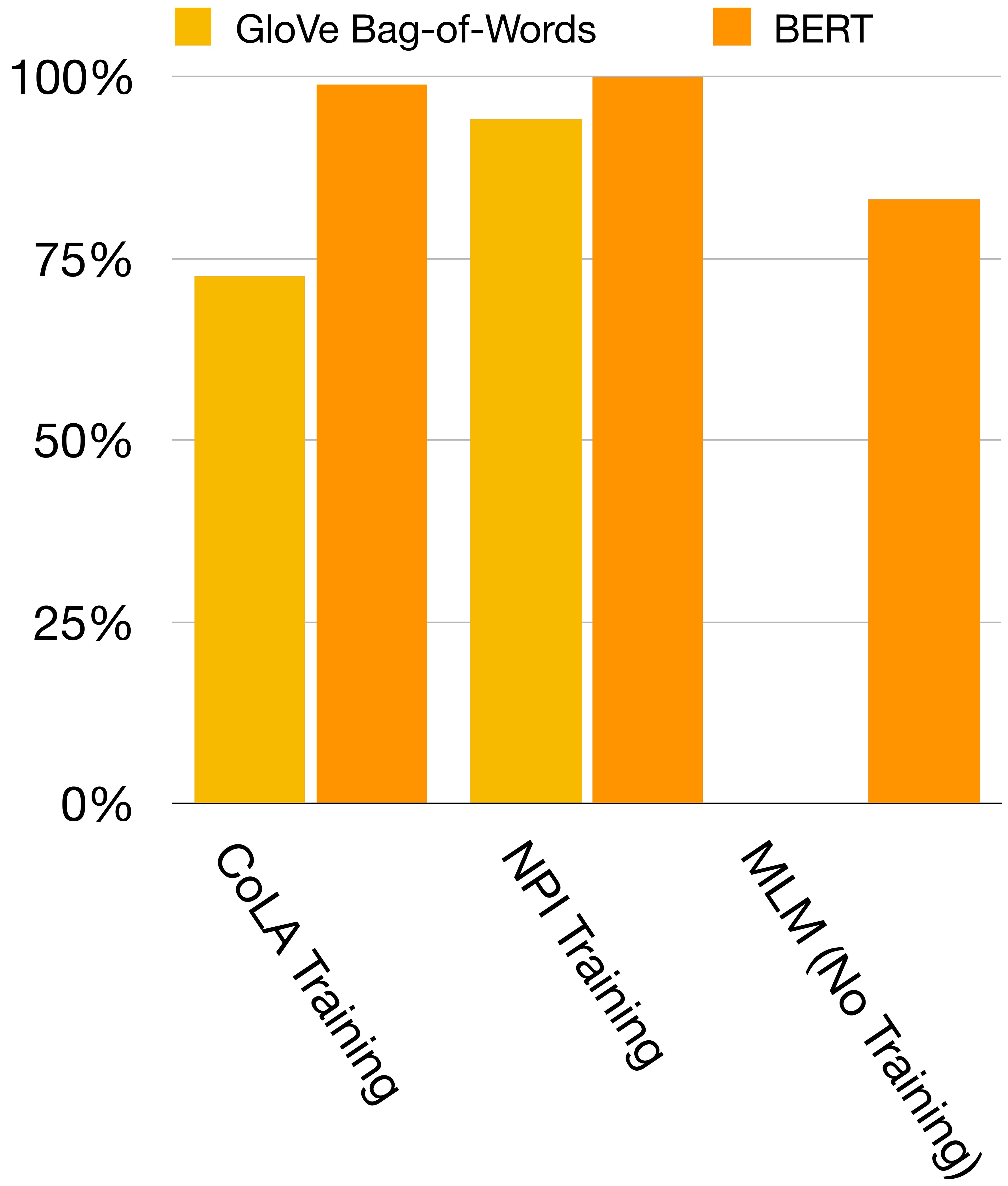
NPI training set (hold-one-out by environment)
or the CoLA general acceptability corpus
or use BERT's language modeling head directly

Test:

NPI environment test sets

Metric:

Pair preference accuracy: How often does the model assign a higher probability of acceptability to the correct sentence?



What if we ask BERT directly?

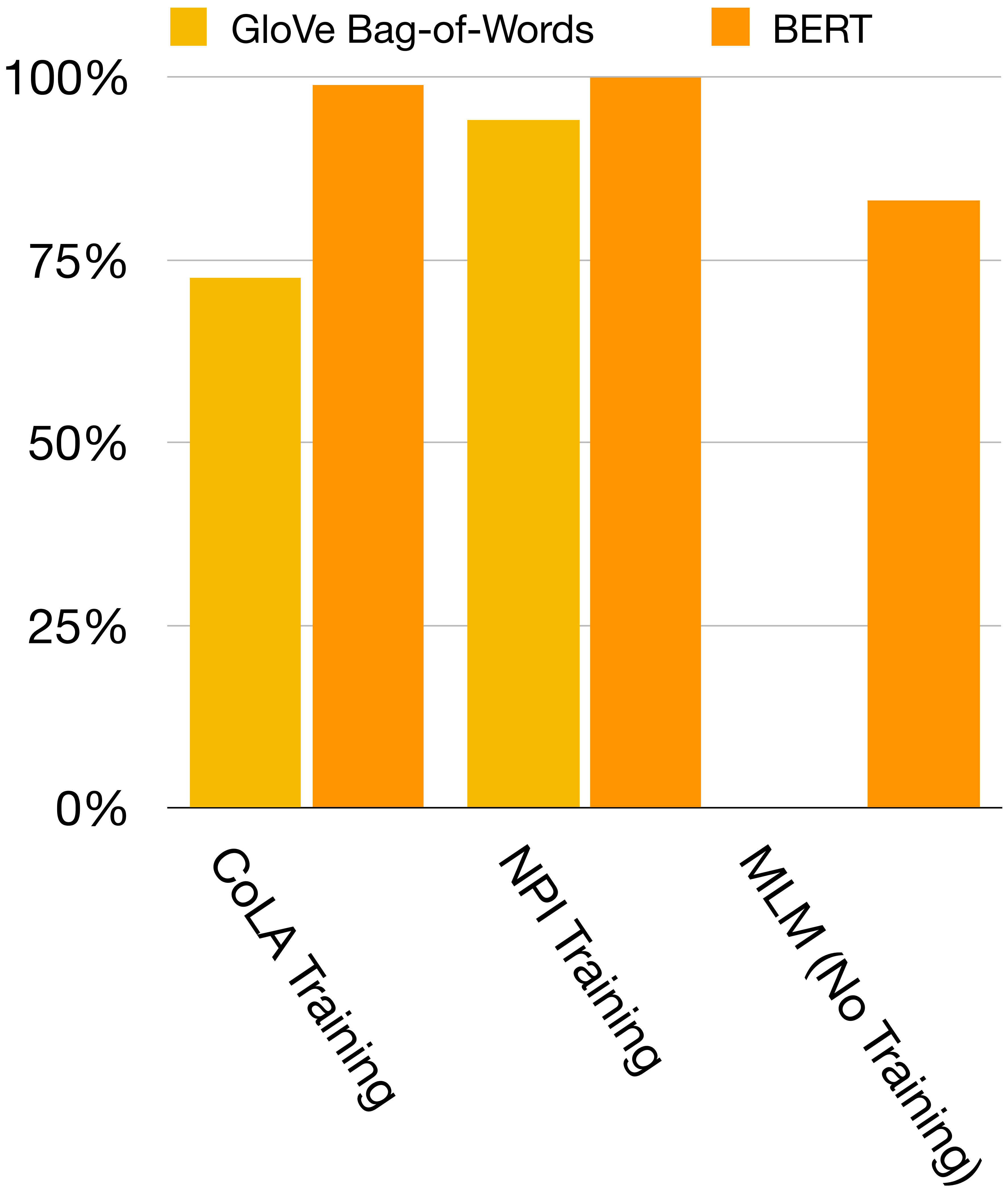
BERT does better than chance (50%), but not especially well.

(2) *Mary has eaten *any* cookies.

Train:
NPI training set (hold-one-out by environment)
or the CoLA general acceptability corpus
or use BERT's language modeling head directly

Test:
NPI environment test sets

Metric:
Pair preference accuracy: How often does the model assign a higher probability of acceptability to the correct sentence?



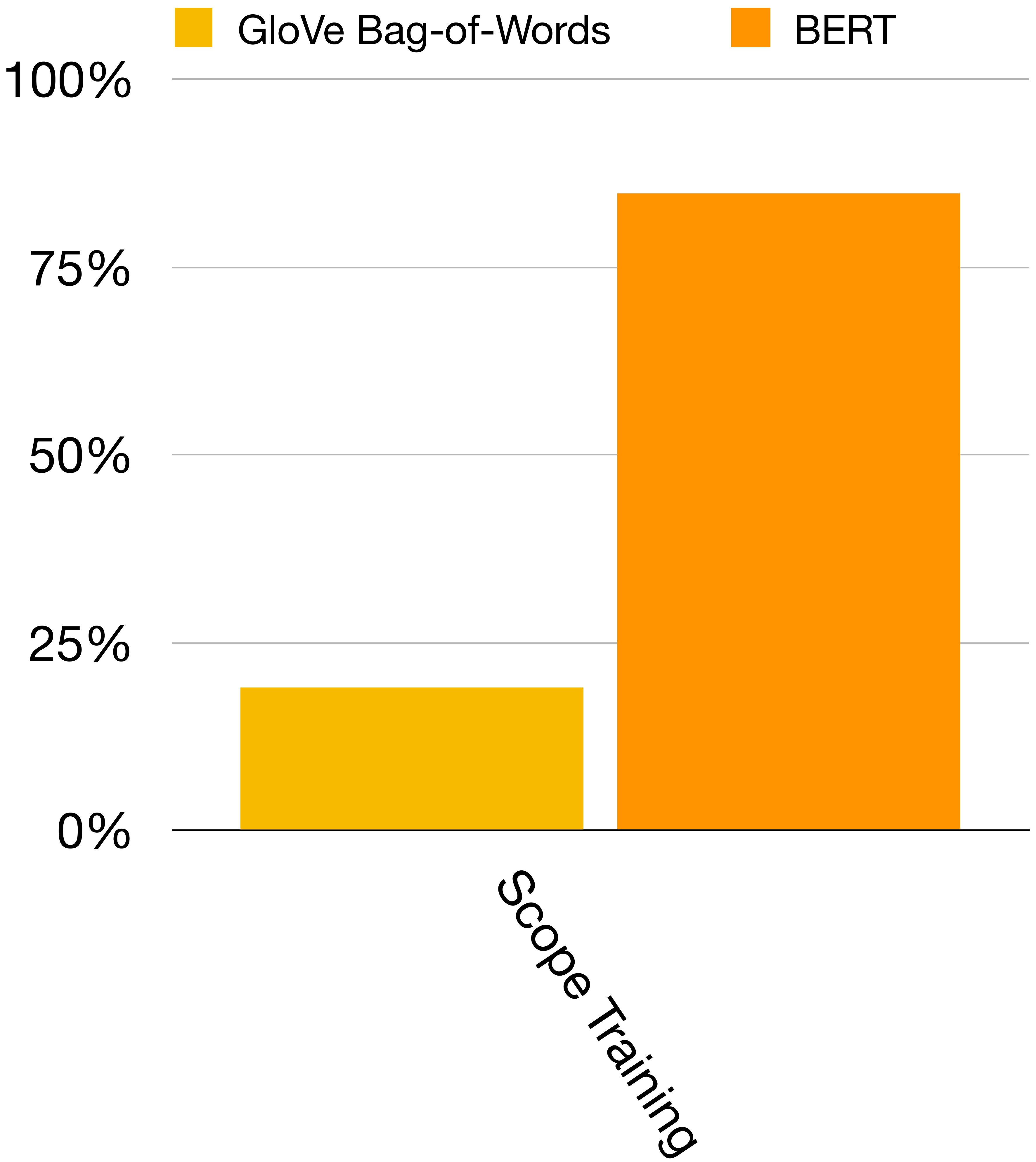
What if we use probing classifiers?

1 Those boys wonder **whether** [the doctors *ever* went to an art gallery.]
0 *Those boys *ever* wonder **whether** [the doctors went to an art gallery.]
1 Those boys wonder **whether** [the doctors *often* went to an art gallery.]
0 Those boys *often* wonder **whether** [the doctors went to an art gallery.]
1 *Those boys say **that** [the doctors *ever* went to an art gallery.]
0 *Those boys *ever* say **that** [the doctors went to an art gallery.]
1 Those boys say **that** [the doctors *often* went to an art gallery.]
0 Those boys *often* say **that** [the doctors went to an art gallery.]

Train:
Scope prediction task, training only a small classifier without fine-tuning BERT (hold-one-out over environments)

Test:
Scope prediction task

Metric:
Matthews Correlation (MCC) for scope judgment



What if we use probing classifiers?

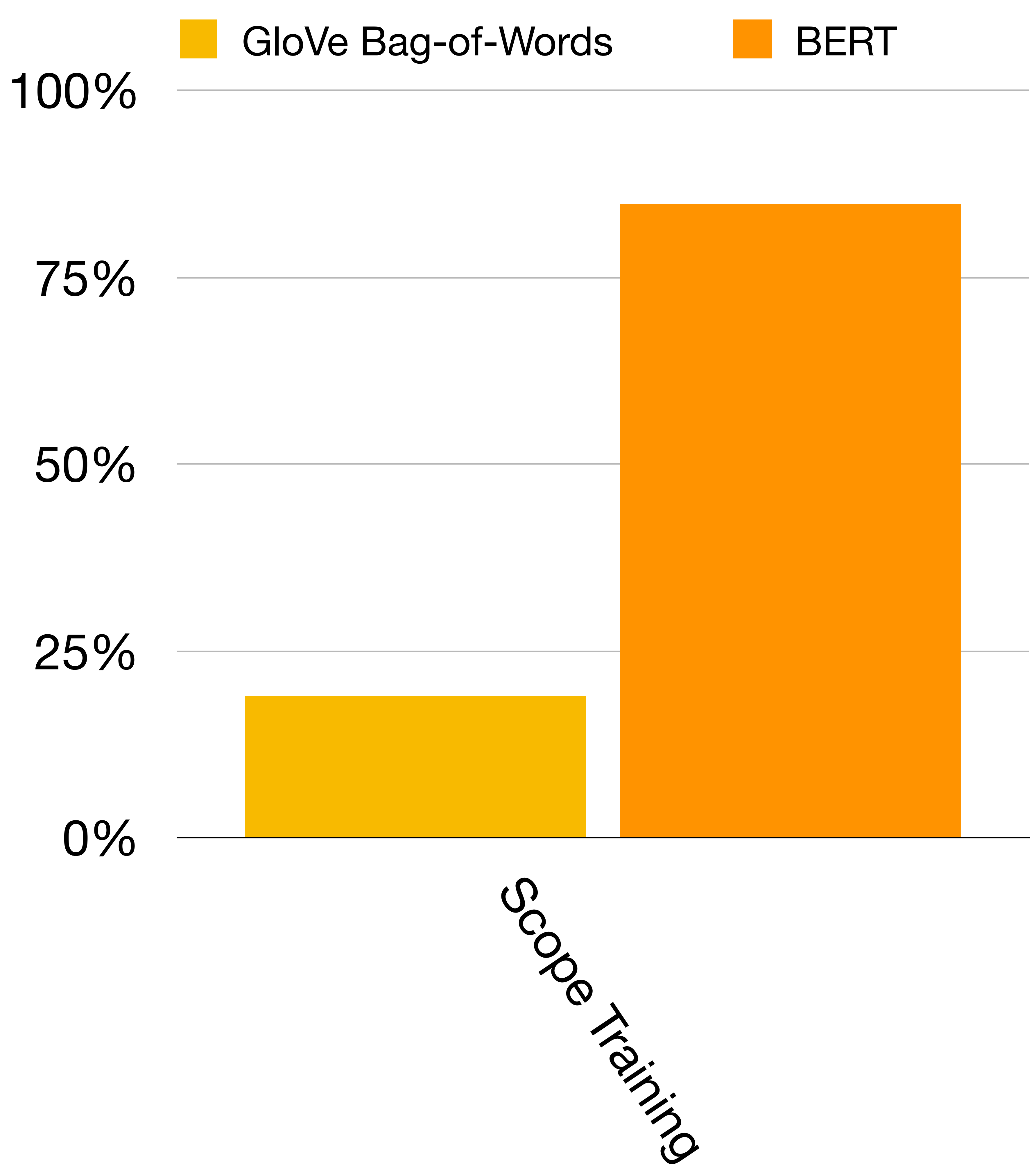
1 ... ery.]
0 ... ry.]
1 ... y.]
0 ... y.]
1
0 *Those boys *ever* ... at [the doctors went to an art gallery.]
1 Those boys say ... [the doctors *often* went to an art gallery.]
0 Those boys *of*... say **that** [the doctors went to an art gallery.]

BERT knows a bit about NPIs, but its not perfect.

Train:
Scope prediction task, training only a small classifier without fine-tuning BERT (hold-one-out over environments)

Test:
Scope prediction task

Metric:
Matthews Correlation (MCC) for scope judgment



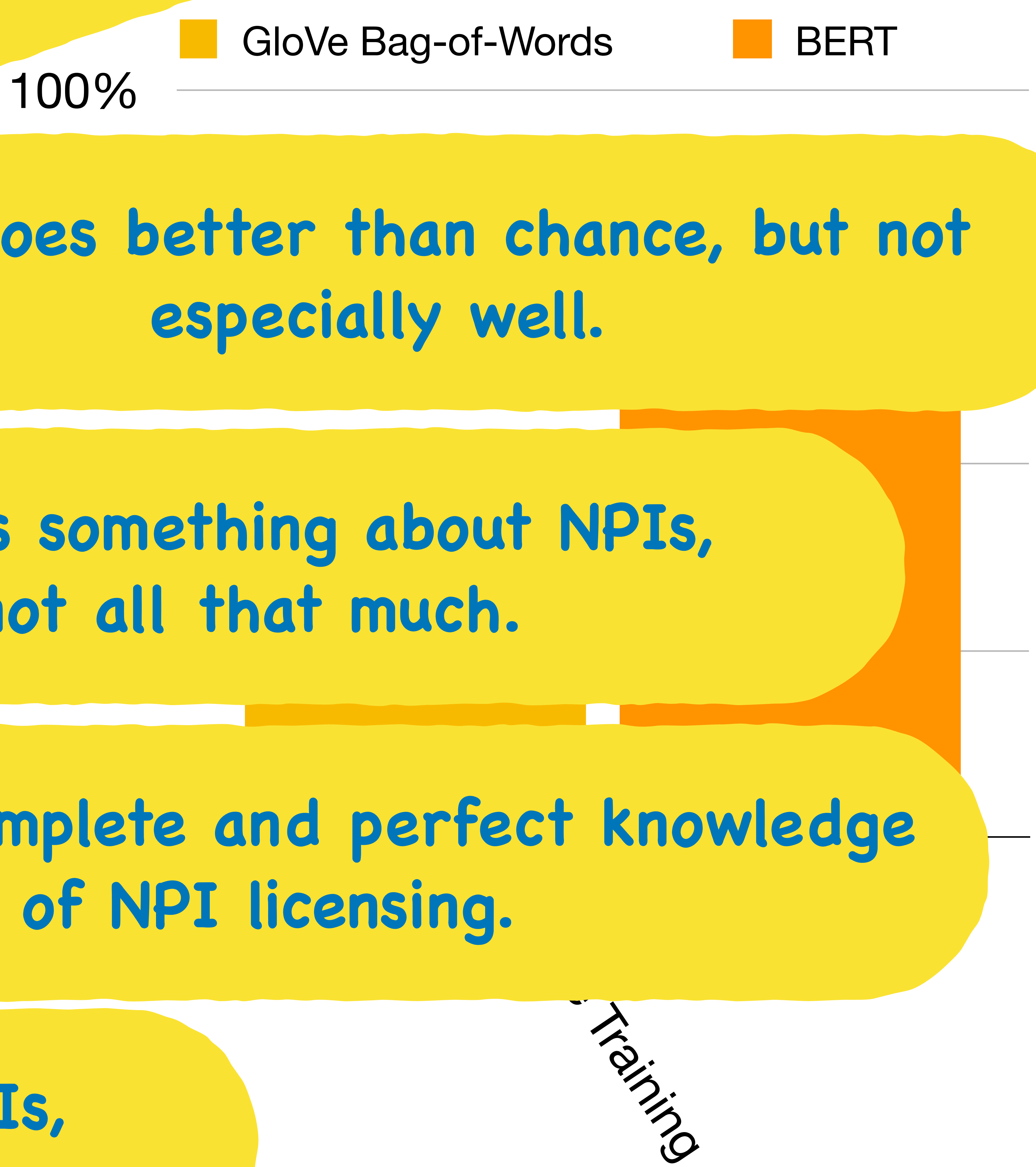
What if we use pr

1 ...ery.]
0 ...ry.]
0
1
0 *Those boys *ever* ... [the doctors went ... art gallery
1 Those boys say ... [the doctors *ofte*... to an art gallery.
0 Those boys *of*... say **that** [the ... tors went to an art gallery]

Train:
Scop... *diction task, training o*... *hold-one*
without fine-tu...
en... *nments)*

Test:
Scope predic... *sk*

Met...



BERT knows a bit about NPIs, but its not perfect.

BERT knows a bit about NPIs, but its not perfect.

BERT does better than chance, but not especially well.

BERT knows something about NPIs, but not all that much.

BERT has complete and perfect knowledge of NPI licensing.

BERT knows something about NPIs, but not all that much.

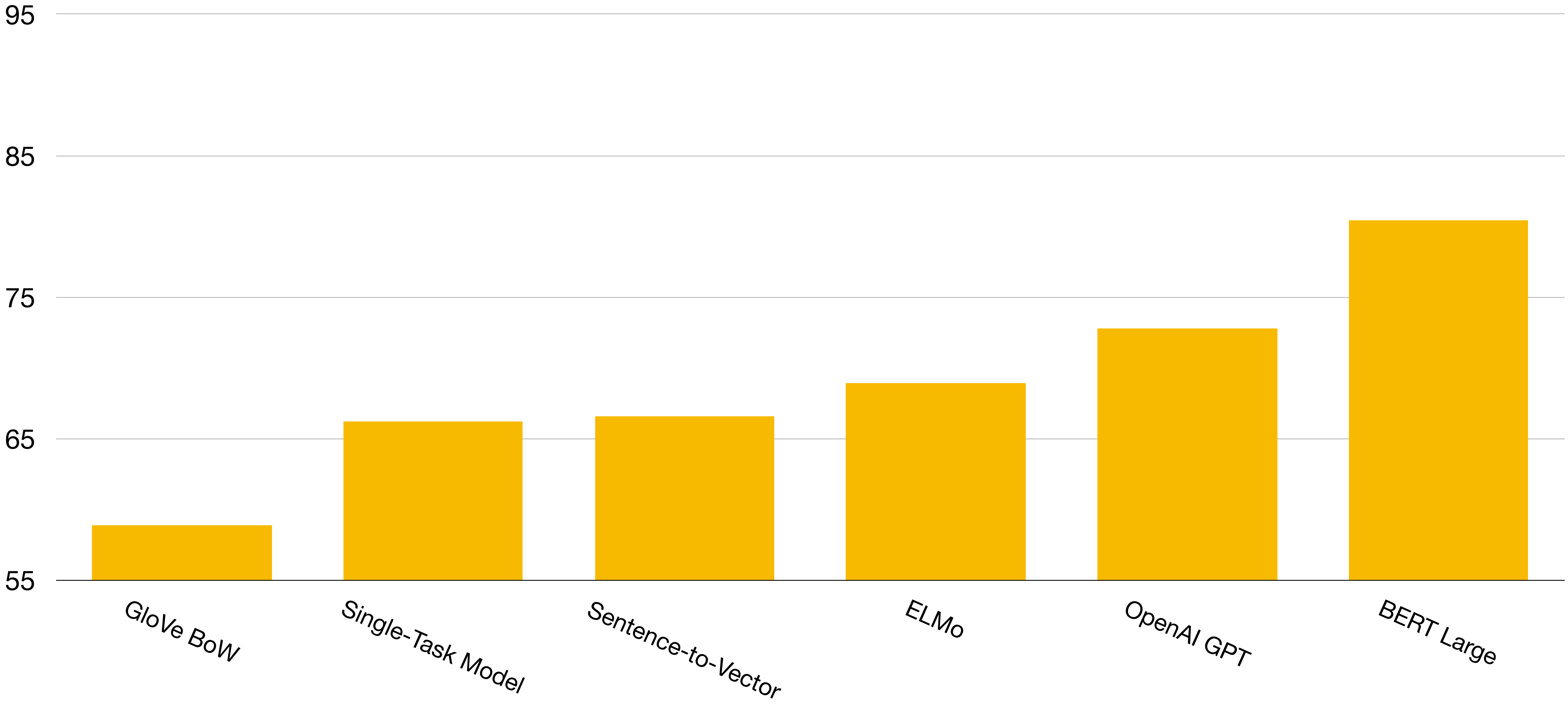
Recent Progress on GLUE

Building a Better Muppet

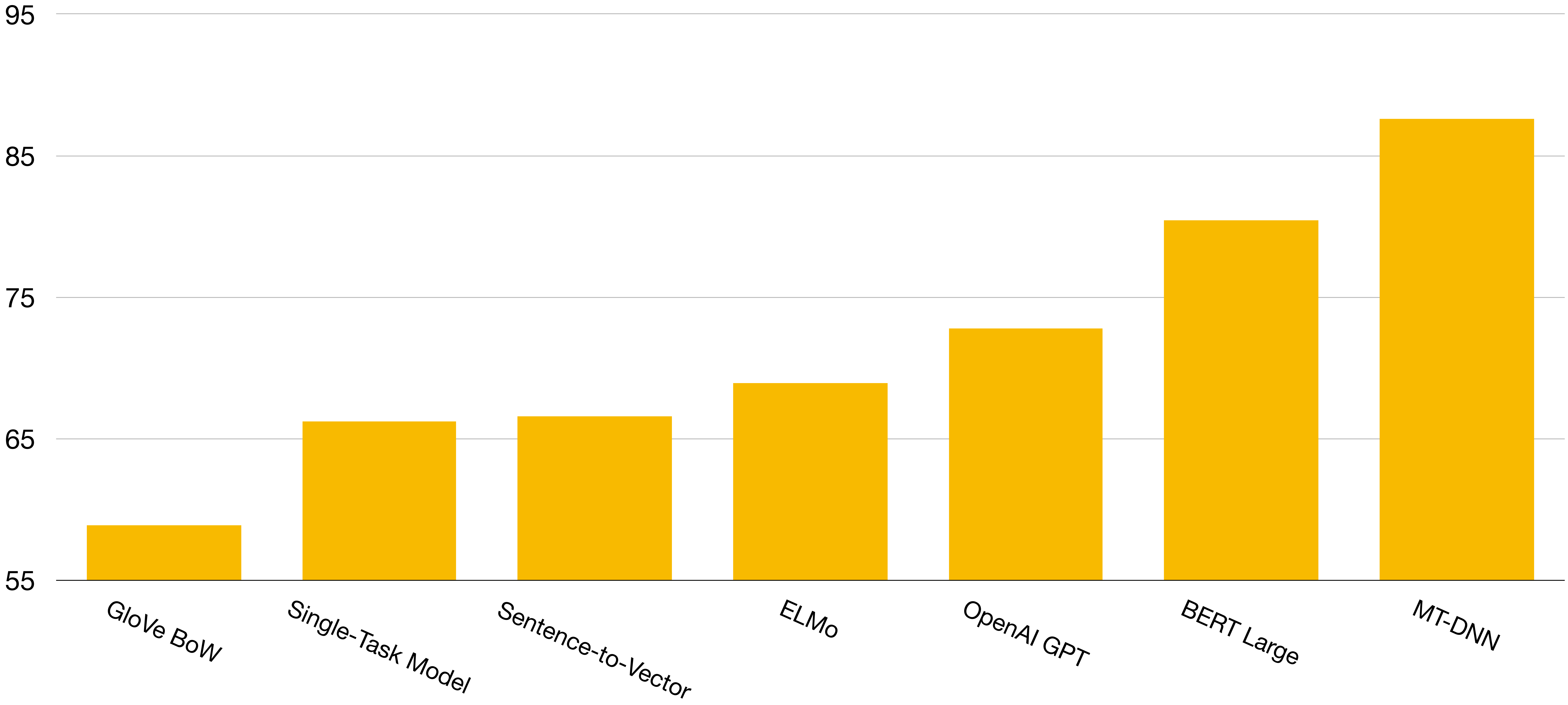
- Lots of follow-up work, including:
  - MT-DNN/ALICE: Multi-task fine-tuning; ensembling
  - RoBERTa: Simplified objective; more training data
  - ALBERT: Modified objective; parameter sharing across layers

Liu et al. '19a, Wang et al. '19, Liu et al. '19b, Anonymous '19

GLUE Score

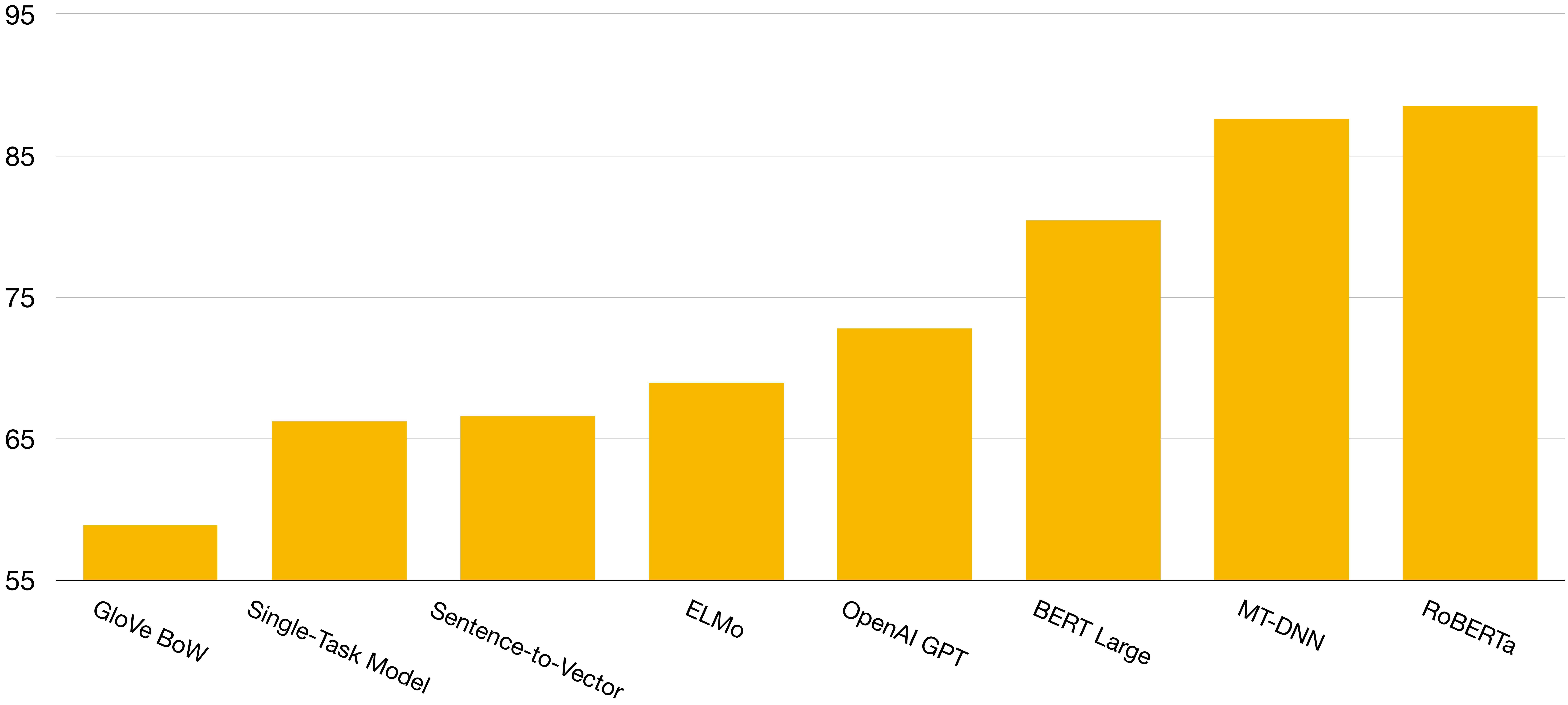

95
85
75
65
55
GloVe BoW
Single-Task Model
Sentence-to-Vector
ELMo
OpenAI GPT
BERT Large

GLUE Score



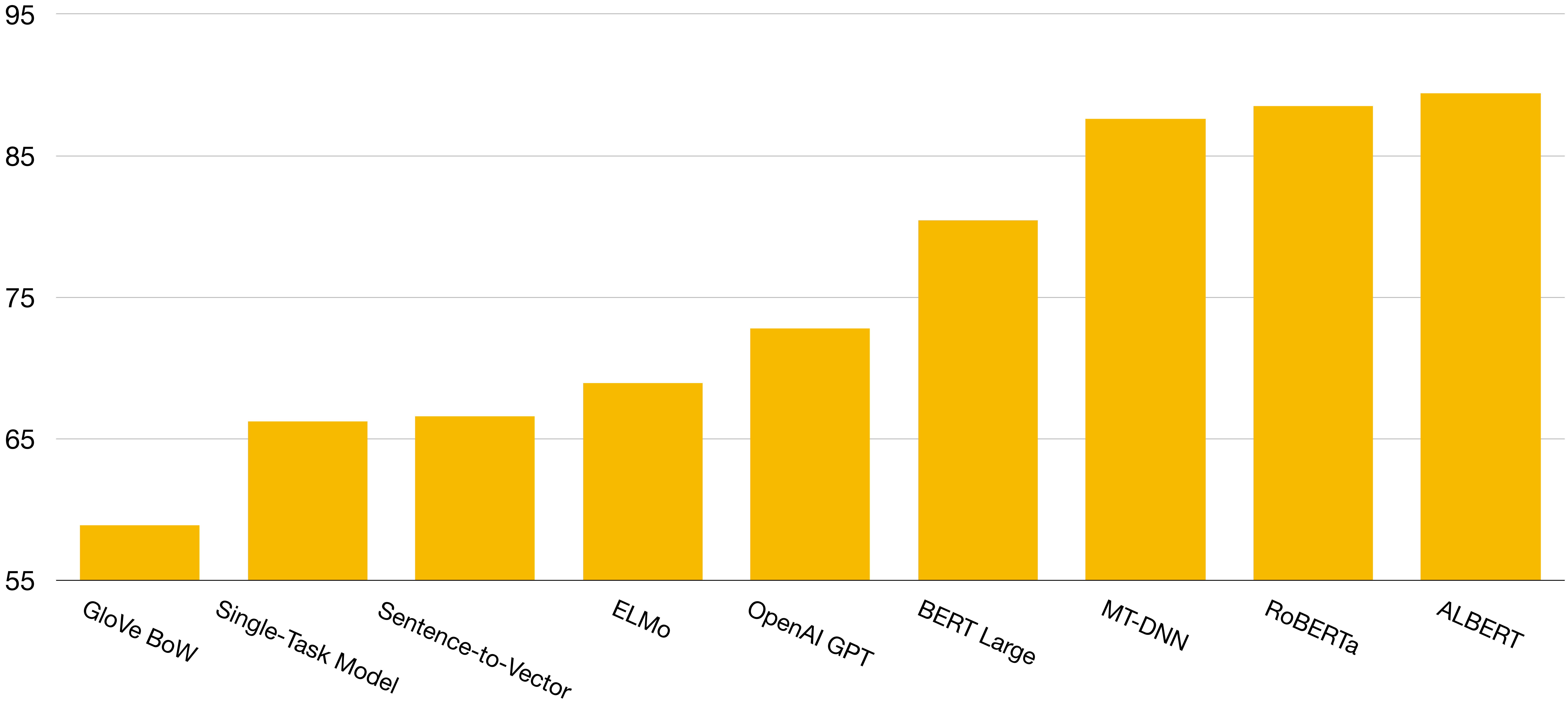
Liu et al. '19

GLUE Score



Liu et al. '19

GLUE Score



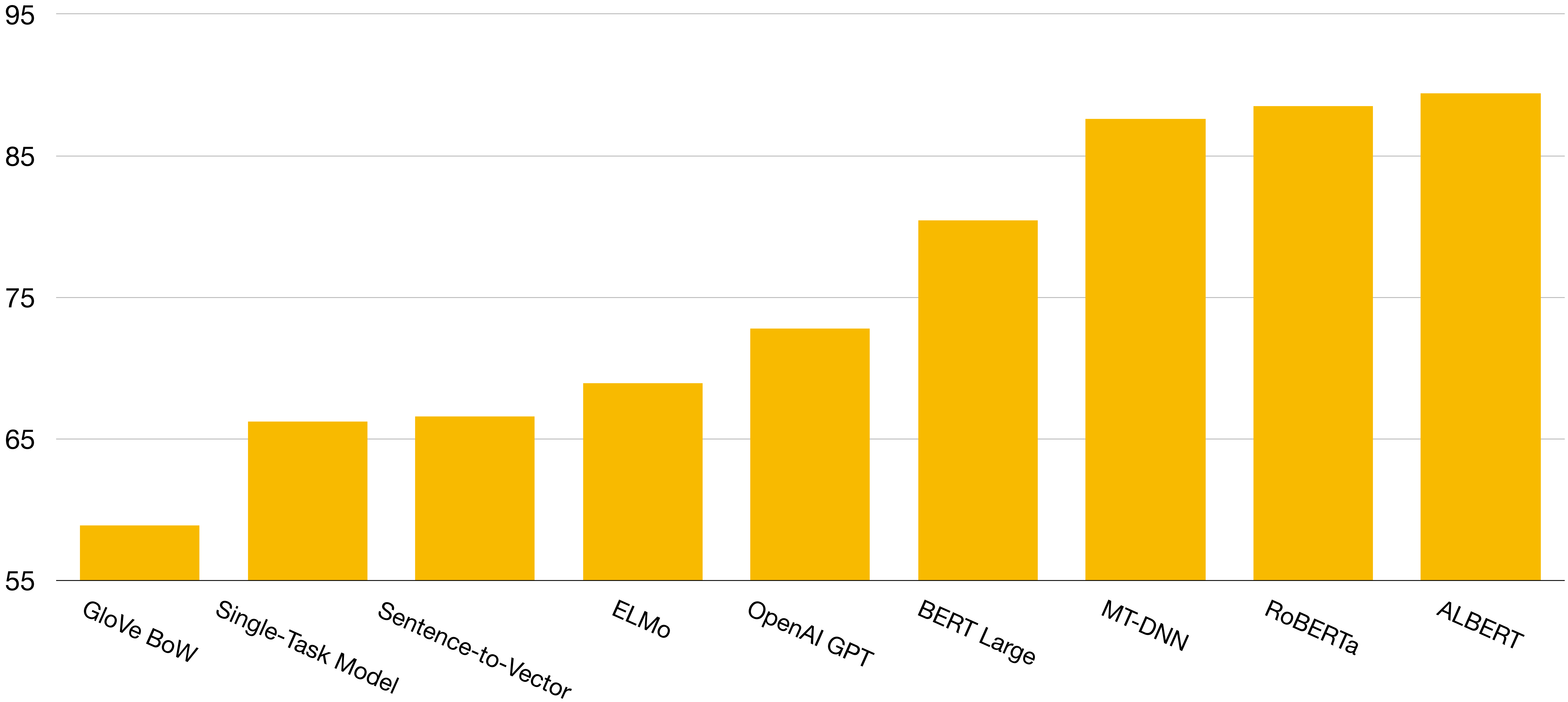
Anonymous '19 (ICLR)

Human Baseline

- **How much headroom does GLUE have left?**
- To compute a conservative estimate for each task:
  - Train crowdworkers with instructions, plus twenty labeled *development set* examples in an interactive training mode.
  - Collect five labels per example for 500 *test set* examples.

Nangia & Bowman '19

GLUE Score



Nangia & Bowman '19

GLUE Score

Nangia & Bowman '19


OVERHEIGHT
MUST TURN
11'-8"
11foot8.com

SuperGLUE

A revised version of GLUE with:

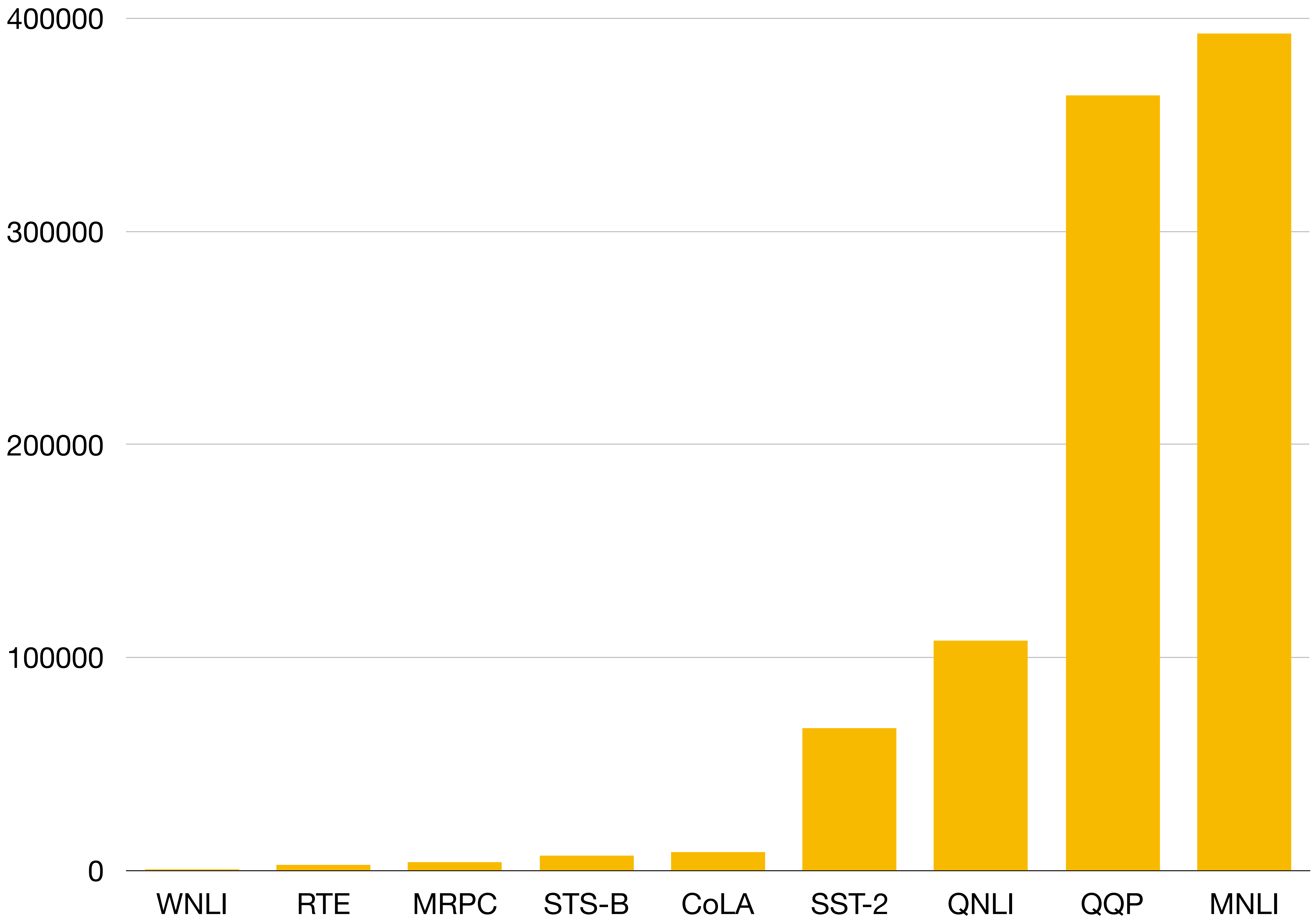
- A new set of eight target tasks...
- ...selected from 30+ submissions to an open call for participation to be easy for humans and hard for BERT.
- A slightly expanded set of task APIs (including multiple-choice QA, word-in-context classification, and more)

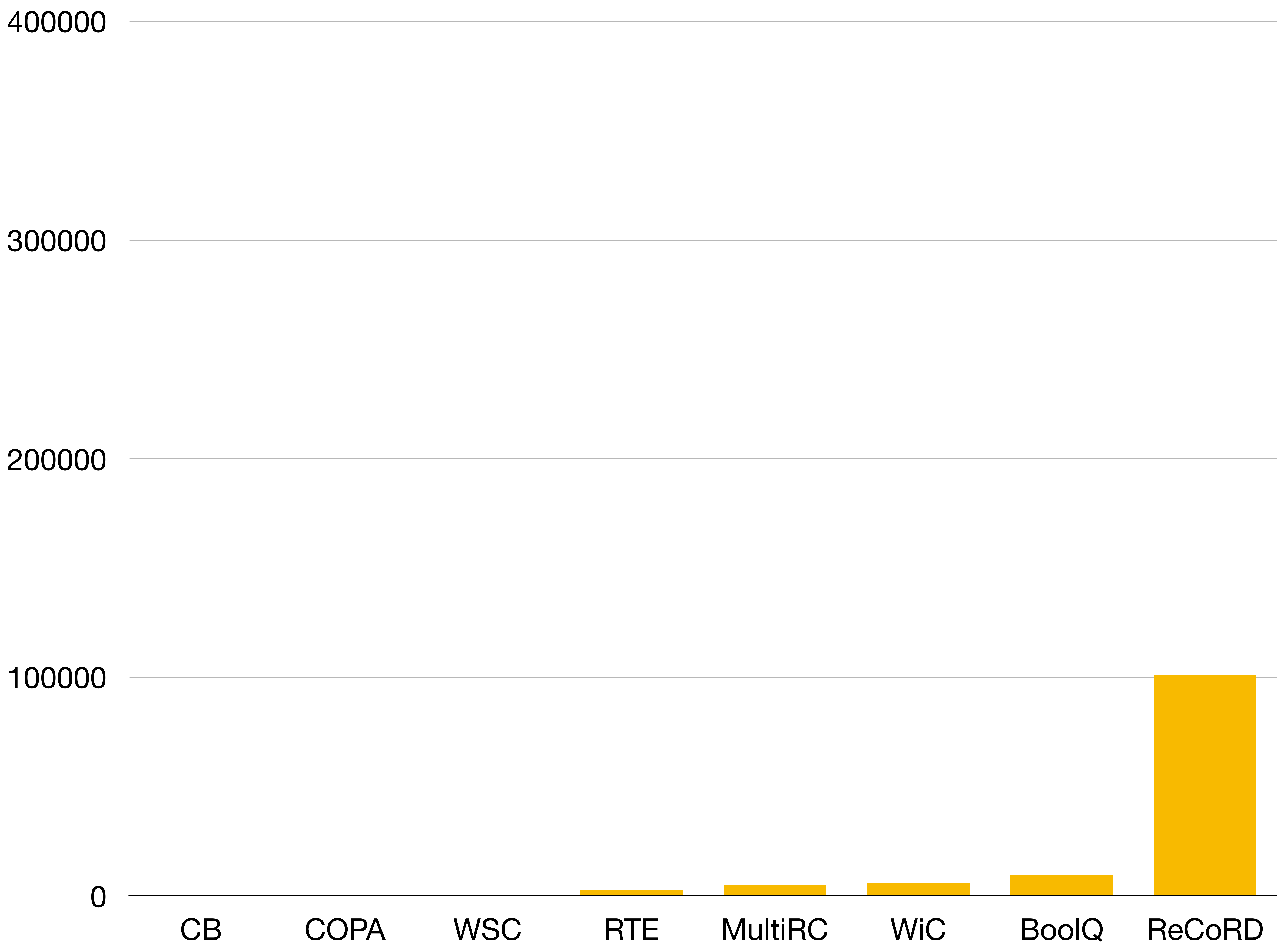
{Wang, Pruksachatkun, Nangia, Singh},
Michael, Hill, Levy & Bowman '19

SuperGLUE: The Main Tasks

Corpus	|Train|	|Dev|	|Test|	Task	Metrics	Text Sources
BoolQ	9427	3270	3245	QA	acc.	Google queries, Wikipedia
CB	250	57	250	NLI	acc./F1	various
COPA	400	100	500	QA	acc.	blogs, photography encyclopedia
MultiRC	5100	953	1800	QA	$F1_a$/EM	various
ReCoRD	101k	10k	10k	QA	F1/EM	news (CNN, Daily Mail)
RTE	2500	278	300	NLI	acc.	news, Wikipedia
WiC	6000	638	1400	WSD	acc.	WordNet, VerbNet, Wiktionary
WSC	554	104	146	coref.	acc.	fiction books

{Wang, Pruksachatkun, Nangia, Singh},
Michael, Hill, Levy & Bowman '19


400000
300000
200000
100000
0
WNLI
RTE
MRPC
STS-B
CoLA
SST-2
QNLI
QQP
MNLI


400000
300000
200000
100000
0
CB
COPA
WSC
RTE
MultiRC
WiC
BoolQ
ReCoRD

The Commitment Bank

de Marneffe et al. '19

- **Three-way NLI classification: Does a speaker utterance entail some embedded clause within that utterance?**

Text: *B: And yet, uh, I we-, I hope to see employer based, you know, helping out. You know, child, uh, care centers at the place of employment and things like that, that will help out. A: Uh-huh. B: What do you think, do you think we are, setting a trend?* **Hypothesis:** *they are setting a trend* **Entailment:** `Unknown`

Corpus	|Train|	|Dev|	|Test|	Task	Metrics	Text Sources
BoolQ	9427	3270	3245	QA	acc.	Google queries, Wikipedia
CB	250	57	250	NLI	acc./F1	various
COPA	400	100	500	QA	acc.	blogs, photography encyclopedia
MultiRC	5100	953	1800	QA	[illegible]	[illegible]
ReCoRD	101k	10k	10k	QA	[illegible]	[illegible]

{Wang, Pruksachatkun, Nangia, Singh}, Michael, Hill, Levy & Bowman '19

MultiRC

Khashabi et al. '18

- **Multiple choice reading comprehension QA over paragraphs.**

Paragraph: *(CNN) – Gabriel García Márquez, widely regarded as one of the most important contemporary Latin American authors, was admitted to a hospital in Mexico earlier this week, according to the Ministry of Health. The Nobel Prize recipient, known as "Gabo," had infections in his lungs and his urinary tract. He was suffering from dehydration, the ministry said. García Márquez, 87, is responding well to antibiotics, but his release date is still to be determined. "I wish him a speedy recovery." Mexican President Enrique Peña wrote on Twitter. García Márquez was born in the northern Colombian town of Aracataca, the inspiration for the fictional town of Macondo, the setting of the 1967 novel "One Hundred Years of Solitude." He won the Nobel Prize for literature in 1982 "for his novels and short stories, in which the fantastic and the realistic are combined in a richly composed world of imagination, reflecting a continent's life and conflicts," according to the Nobel Prize website. García Márquez has spent many years in Mexico and has a huge following there. Colombian President Juan Manuel Santos said his country is thinking of the author. "All of Colombia wishes a speedy recovery to the greatest of all time: Gabriel García Márquez," he tweeted. CNN en Español's Fidel Gutierrez contributed to this story.*

Question: *Whose speedy recover did Mexican President Enrique Peña wish on Twitter?*

Candidate answers: *Enrique Peña* (F), *Gabriel Garcia Marquez* (T), *Gabo* (T), *Gabriel Mata* (F), *Fidel Gutierrez* (F), *87* (F), *The Nobel Prize recipient* (T)

MultiRC	5100	953	1800	QA	$F1_a$/EM	various
ReCoRD	101k	10k	10k	QA	[illegible]	

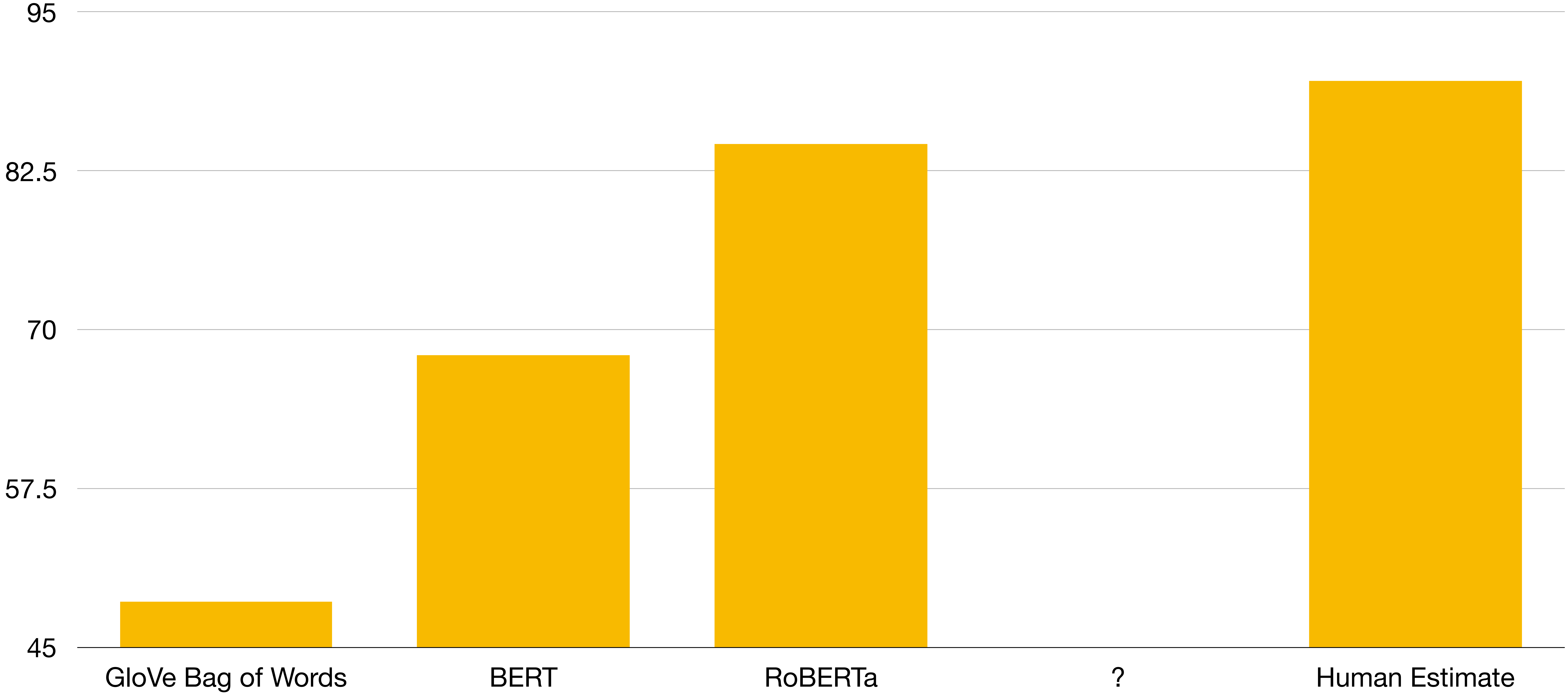
{Wang, Pruksachatkun, Nangia, Singh}, Michael, Hill, Levy & Bowman '19

SuperGLUE: The Main Tasks

Corpus	|Train|	|Dev|	|Test|	Task	Metrics	Text Sources
BoolQ	9427	3270	3245	QA	acc.	Google queries, Wikipedia
CB	250	57	250	NLI	acc./F1	various
COPA	400	100	500	QA	acc.	blogs, photography encyclopedia
MultiRC	5100	953	1800	QA	$F1_a$/EM	various
ReCoRD	101k	10k	10k	QA	F1/EM	news (CNN, Daily Mail)
RTE	2500	278	300	NLI	acc.	news, Wikipedia
WiC	6000	638	1400	WSD	acc.	WordNet, VerbNet, Wiktionary
WSC	554	104	146	coref.	acc.	fiction books

{Wang, Pruksachatkun, Nangia, Singh},
Michael, Hill, Levy & Bowman '19

SuperGLUE Score



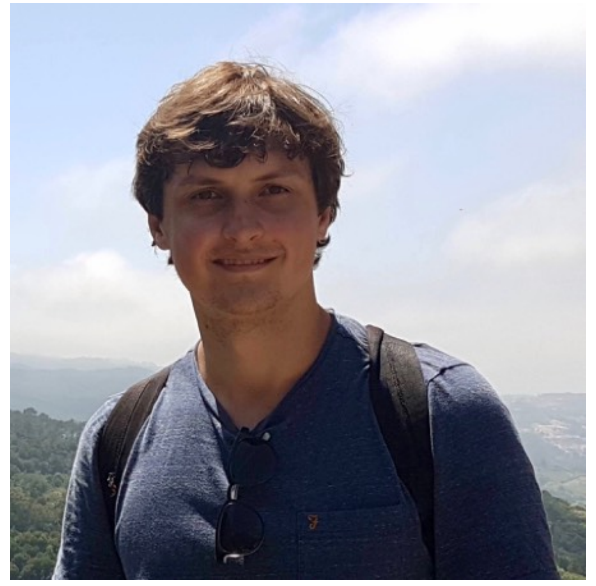
{Wang, Pruksachatkun, Nangia, Singh}, Michael, Hill, Levy & Bowman '19

⚠ GLUE and SuperGLUE: Limitations

- GLUE and SuperGLUE are built only on English data.
    - General-purpose pretraining may look quite different in lower-resource languages!
- GLUE and SuperGLUE use some naturally occurring and crowdsourced data.
    - Therefore safe to presume that these datasets contain evidence of social bias (see Rudinger et al., EthNLP '17).
    - All else being equal, models that learn and use these biases will ***do better on these benchmarks***.
    - In SuperGLUE's WinoGender Schema evaluation (Rudinger et al. ’18), RoBERTa ~9x more sensitive to irrelevant gender information than humans.

A Handy Trick

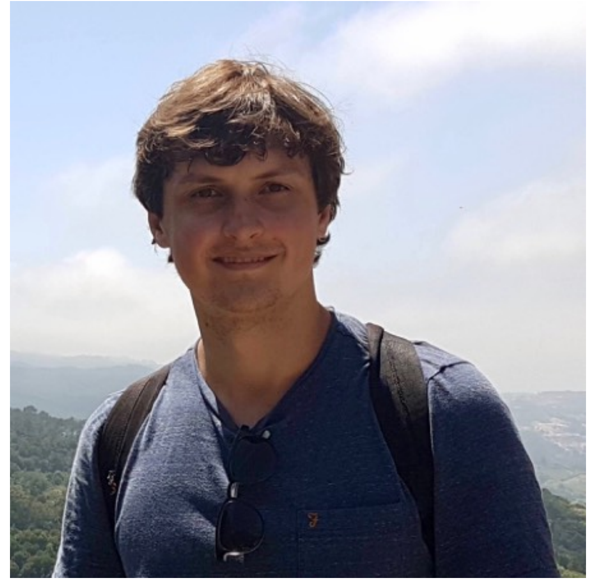
Muppets on STILTs?



- What if you want to solve a hard task with limited training data, but have access to abundant data for another task with that uses similar skills?
- Example: Commitment Bank (250) with MNLI (393k)
- Supplementary Training on Intermediate Labeled-data Tasks (*STILTs*) is an **easy but very robust** solution:
    - Download a large model like BERT that was pretrained on unlabeled data.
    - Fine tune that model on the *intermediate* labeled-data task.
    - Fine tune the same model further on the target task.

Phang, Févry & Bowman '18

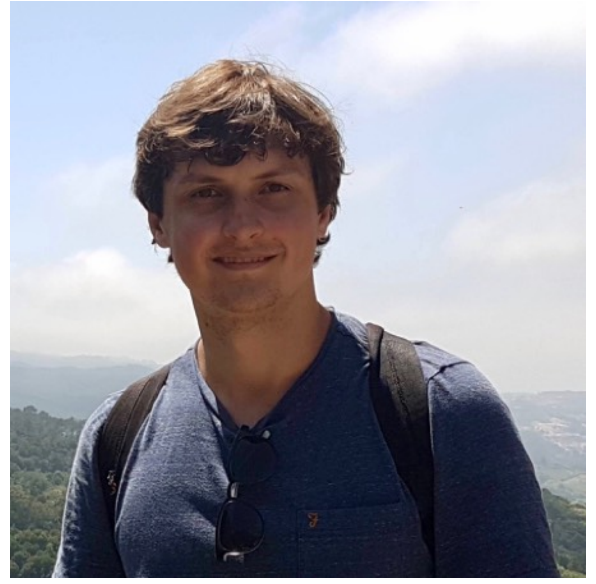
BERT on STILTs



- +1.5 on GLUE w/ MNLI and QQP
- +2.5 on SuperGLUE w/ MNLI
- *Clark et al. '19*: +3.7 on BoolQ w/ MNLI
- *Sap et al. '18*: +4 to +8 on commonsense tasks w/ SocialIQA
- MNLI+STILTs built into RoBERTa and ALBERT

Phang, Févry & Bowman '18

BERT on STILTs



- +1.5 on GLUE w/ MNLI and QQP
- +2.5 on SuperGLUE w/ MNLI
- *Clark et al. '19*: +3.7 on BoolQ w/ MNLI
- *Sap et al. '19*: +4 to +8 on commonsense tasks w/ SocialIQA
- MNLI+STILTs built into RoBERTa and ALBERT

Tuning Not Required!

Phang, Févry & Bowman '18

ELMo and BERT Base on STILTs

Intermediate Task	Avg	CoLA	SST							
			ELMo							
Random^E	70.5	38.5	87.7							
Single-Task^E	71.2	39.4	**90.6**							
CoLA^E	71.1	39.4	87.3							
SST^E	71.2	38.8	**90.6**							
MRPC^E	71.3	40.0	88.4							
QQP^E	70.8	34.3	88.6							
STS^E	71.6	39.9	88.4							
MNLI^E	72.1	38.9	89.0							
QNLI^E	71.2	37.2	88.3	81.1/86.9	85.5/81.7	78.9/80.1	74.7	78.0	58.8	22.5*
RTE^E	71.2	38.5	87.7	81.1/87.3	86.6/83.2	80.1/81.1	74.6	78.0	55.6	32.4*
WNLI^E	70.9	38.4	88.6	78.4/85.9	86.3/82.8	79.1/80.0	73.9	77.9	57.0	11.3*
DisSent WP^E	71.9	39.9	87.6	**81.9/87.2**	85.8/82.3	79.0/80.7	74.6	79.1	61.4	23.9*
MT En-De^E	72.1	40.1	87.8	79.9/86.6	86.4/83.2	81.8/82.4	75.9	79.4	58.8	31.0*
MT En-Ru^E	70.4	**41.0**	86.8	76.5/85.0	82.5/76.3	81.4/81.5	70.1	77.3	60.3	45.1*
Reddit^E	71.0	38.5	87.7	77.2/85.0	85.4/82.1	80.9/81.7	74.2	79.3	56.7	21.1*
SkipThought^E	71.7	40.6	87.7	79.7/86.5	85.2/82.1	81.0/81.7	75.0	79.1	58.1	52.1*
MTL GLUE^E	72.1	33.8	90.5	81.1/87.4	86.6/83.0	82.1/83.3	**76.2**	79.2	61.4	42.3*
MTL Non-GLUE^E	72.4	39.4	88.8	80.6/86.8	**87.1/84.1**	**83.2/83.9**	75.9	**80.9**	57.8	22.5*
MTL All^E	72.2	37.9	89.6	79.2/86.4	86.0/82.8	81.6/82.5	76.1	80.2	60.3	31.0*
			BERT with Intermediate Task Training							
Single-Task^B	78.8	56.6	90.9	88.5/91.8	89.9/86.4	86.1/86.0	83.5	**87.9**	69.7	**56.3**
CoLA^B	78.3	**61.3**	91.1	87.7/91.4	89.7/86.3	85.0/85.0	83.3	85.9	64.3	43.7*
SST^B	78.4	57.4	**92.2**	86.3/90.0	89.6/86.1	85.3/85.1	83.2	87.4	67.5	43.7*
MRPC^B	78.3	60.3	90.8	87.0/91.1	89.7/86.3	86.6/86.4	**83.8**	83.9	66.4	**56.3**
QQP^B	79.1	56.8	91.3	88.5/91.7	**90.5/87.3**	88.1/87.8	83.4	87.2	69.7	**56.3**
STS^B	79.4	61.1	92.3	88.0/91.5	89.3/85.5	86.2/86.0	82.9	87.0	71.5	50.7*
MNLI^B	**79.6**	56.0	91.3	88.0/91.3	90.0/86.7	87.8/87.7	82.9	87.0	**76.9**	**56.3**
QNLI^B	78.4	55.4	91.2	**88.7/92.1**	89.9/86.4	86.5/86.3	82.9	86.8	68.2	**56.3**
RTE^B	77.7	59.3	91.2	86.0/90.4	89.2/85.9	85.9/85.7	82.0	83.3	65.3	**56.3**
WNLI^B	76.2	53.2	92.1	85.5/90.0	89.1/85.5	85.6/85.4	82.4	82.5	58.5	**56.3**
DisSent WP^B	78.1	58.1	91.9	87.7/91.2	89.2/85.9	84.2/84.1	82.5	85.5	67.5	43.7*
MT En-De^B	73.9	47.0	90.5	75.0/83.4	89.6/86.1	84.1/83.9	81.8	83.8	54.9	**56.3**
MT En-Ru^B	74.3	52.4	89.9	71.8/81.3	89.4/85.6	82.8/82.8	81.5	83.1	58.5	43.7*
Reddit^B	75.6	49.5	91.7	84.6/89.2	89.4/85.8	83.8/83.6	81.8	84.4	58.1	**56.3**
SkipThought^B	75.2	53.9	90.8	78.7/85.2	89.7/86.3	81.2/81.5	82.2	84.6	57.4	43.7*
MTL GLUE^B	**79.6**	56.8	91.3	88.0/91.4	90.3/86.9	**89.2/89.0**	83.0	86.8	74.7	43.7*
MTL Non-GLUE^B	76.7	54.8	91.1	83.6/88.7	89.2/85.6	83.2/83.2	82.4	84.4	64.3	43.7*
MTL All^B	79.3	53.1	91.7	88.0/91.3	90.4/87.0	88.1/87.9	83.5	87.6	75.1	45.1*
			Test Set Results							
Non-GLUE^E	69.7	34.5	89.5	78.2/84.8	83.6/64.3	77.5/76.0	75.4	74.8	55.6	65.1
MNLI^B	77.1	49.6	93.2	88.5/84.7	70.6/88.3	86.0/85.5	82.7	78.7	72.6	65.1
GLUE^B	77.3	49.0	93.5	89.0/85.3	70.6/88.6	85.8/84.9	82.9	81.0	71.7	34.9
BERT Base	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

- Most intermediate tasks *harm* performance, especially with BERT.
  - This includes most of the GLUE tasks, MT, Reddit prediction, DisSent, and several more!
- BERT with MNLI *or* BERT with GLUE (multi-task) work best, and show consistent improvements.

Wang, Hula, Xia, Pappagari, McCoy, Patel, Kim, Tenney, Huang, Yu, Jin, Chen, Van Durme, Grave, Pavlick and Bowman '19

Practical Conclusions

- If you're building a language understanding model now, you have at least a few thousand training examples, and you need the best performance you can get:
  - Use **RoBERTa**.
  - If you're aware of a big dataset for some related task, or if you're working with very limited training data, use **STILTs**, too!
- **Don't be too quick to trust** any one analysis study that claims to tell you what NLP models *know*.
- Keep an eye on **super.gluebenchmark.com** for future developments in this area.
- For a toolkit that implements everything I've spoken about, try **jiant.info**.

Open Questions

Plenty of open questions!

- How far can we push plain unsupervised pretraining with bigger models?
- What makes a task suitable for use as as intermediate task in STILTs?
- Are we nearing the end of the line for evaluation with IID test sets?
- How can we mitigate the social biases that these models learn during pretraining and fine-tuning?

Thanks!

Questions:

bowman@nyu.edu

@sleepinyourhat

NYU

ML 2 Machine Learning for Language

Try SuperGLUE: super.gluebenchmark.com

Sponsors

SAMSUNG Research

NSF

NVIDIA

See cited papers for full project details.

But wait! There's more!

Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	AX
1	ALBERT-Team Google Language	ALBERT (Ensemble)		89.4	69.1	97.1	93.4/91.2	92.5/92.0	74.2/90.5	91.3	91.0	99.2	89.2	91.8	50.2
2	Microsoft D365 AI & UMD	Adv-RoBERTa (ensemble)		88.8	68.0	96.8	93.1/90.8	92.4/92.2	74.8/90.3	91.1	90.7	98.8	88.7	89.0	50.1
3	Facebook AI	RoBERTa		88.5	67.8	96.7	92.3/89.8	92.2/91.9	74.3/90.2	90.8	90.2	98.9	88.2	89.0	48.7
4	XLNet Team	XLNet-Large (ensemble)		88.4	67.8	96.8	93.0/90.7	91.6/91.1	74.2/90.3	90.2	89.8	98.6	86.3	90.4	47.5
5	Microsoft D365 AI & MSR AI	MT-DNN-ensemble		87.6	68.4	96.5	92.7/90.3	91.1/90.7	73.7/89.9	87.9	87.4	96.0	86.3	89.0	42.8
6	GLUE Human Baselines	GLUE Human Baselines		87.1	66.4	97.8	86.3/80.8	92.7/92.6	59.5/80.4	92.0	92.8	91.2	93.6	95.9	-

Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-b	AX-g
1	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0	76.6	99.3/99.7
2	Facebook AI	RoBERTa		84.6	87.1	90.5/95.2	90.6	84.4/52.5	90.6/90.0	88.2	69.9	89.0	57.9	91.0/78.1
3	SuperGLUE Baselines	BERT++		71.5	79.0	84.8/90.4	73.8	70.0/24.1	72.0/71.3	79.0	69.6	64.4	38.0	99.4/51.4
		BERT		69.0	77.4	75.7/83.6	70.6	70.0/24.1	72.0/71.3	71.7	69.6	64.4	23.0	97.8/51.7