Evaluating Recent Progress Toward General-Purpose Language Understanding Models

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.

The Technique: Muppets

Large-scale pretrained language models like **ELMo**, GPT, **BERT**, XLNet, **RoBERTa**, and T5 have offered a recent surge of progress toward this goal.

This Talk

- The GLUE language understanding benchmark
  Wang et al. '19a

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

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

- What's next for evaluation?
  Idle speculation '19

GLUE: What is it?



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 ICLR '19

Why GLUE?

Increasingly common for researchers outside NLP to evaluate new techniques on language understanding tasks.

- We can learn a lot this way...
- ...if these researchers evaluate on significant open problems...
- ...which doesn't always happen.

Wang, Singh, Michael, Hill, Levy & Bowman ICLR '19

Why GLUE?

GLUE for non-NLP-specialist researchers:

- We provide tasks, metrics, baselines, and code that represent open problems of interest to researchers in NLU.
- We don't enforce any particular experimental design—that's up to the (expert) users.

Wang, Singh, Michael, Hill, Levy & Bowman ICLR '19

GLUE

Nine English-language sentence understanding tasks based on existing data:

- Unsolved
- Varied training data volume
- Varied language style/genre

Wang, Singh, Michael, Hill, Levy & Bowman ICLR '19

GLUE

Simple task APIs:

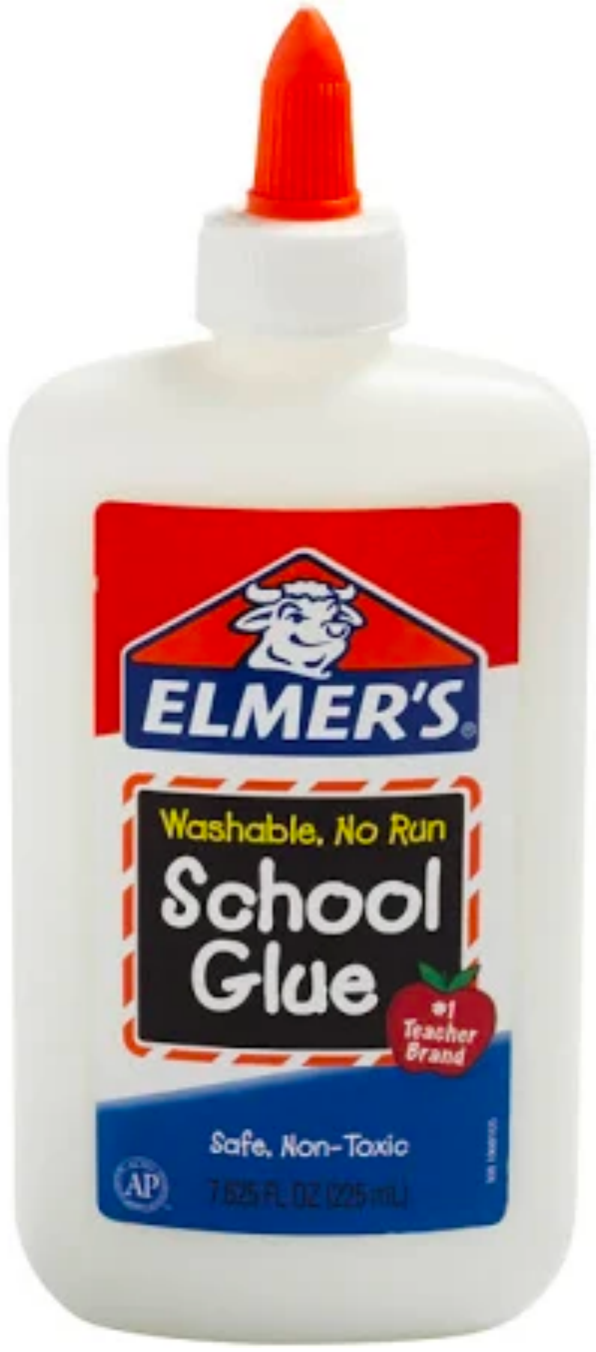
- Only sentence or sentence pair inputs.
- Only classification or regression outputs.
- *No generation or structured prediction.*

Wang, Singh, Michael, Hill, Levy & Bowman ICLR '19

GLUE

Simple leaderboard API: Upload predictions for a test set (like Kaggle/SemEval)

- Usable with any software infrastructure.
- Usable with any kind of method/model!
- Allows us to limit use of the test sets.



Wang, Singh, Michael, Hill, Levy & Bowman ICLR '19

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

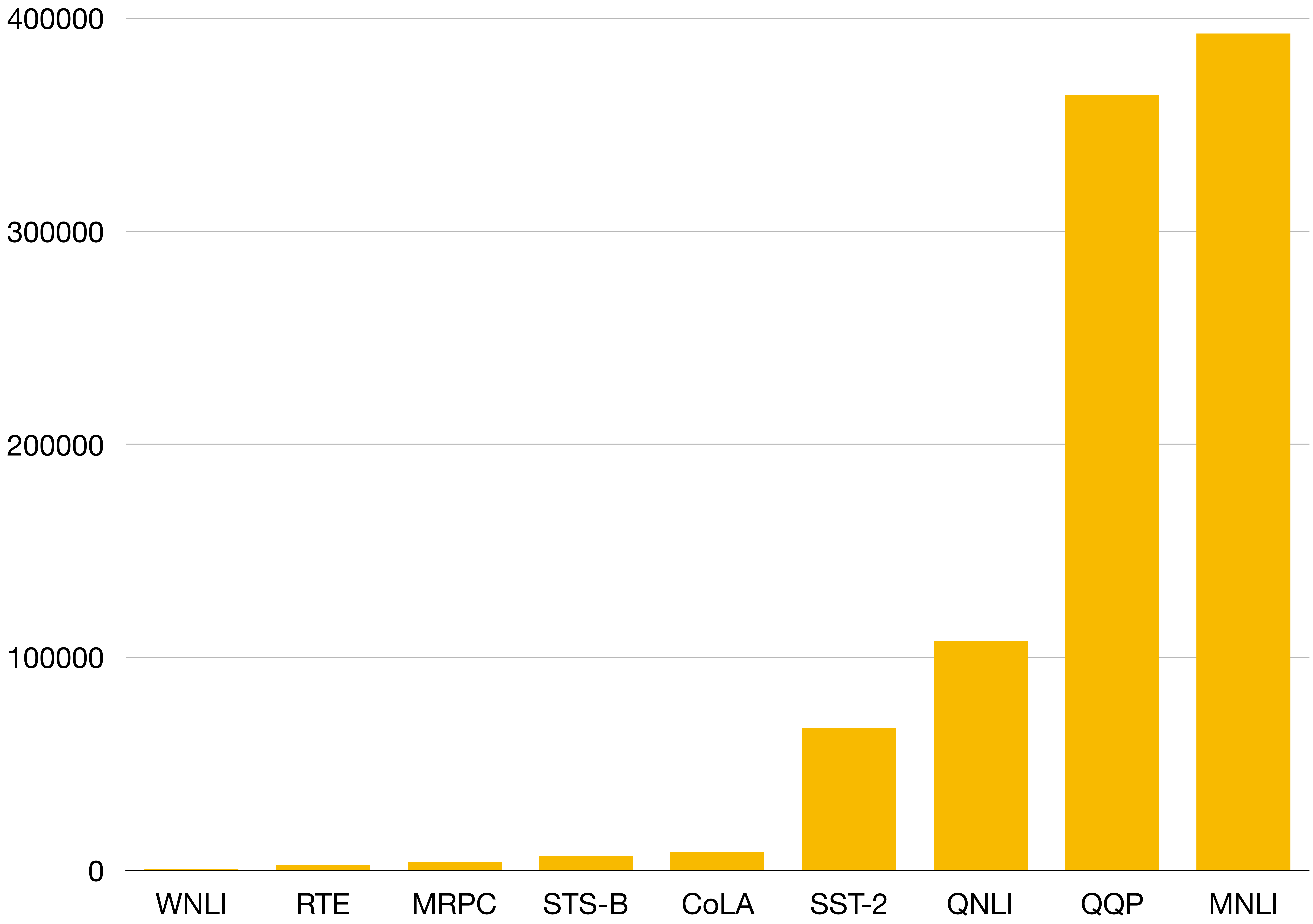
http://jiant.info/

Wang, Singh, Michael, Hill, Levy & Bowman ICLR '19

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 ICLR '19


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 ICLR '19

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 ICLR '19

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 ICLR '19

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 Paraph			

Wang, Singh, Michael, Hill, Levy & Bowman ICLR '19

The Recognizing Textual Entailment Challenge

Dagan et al. '06 et seq.

- **Binary classification over sentence pairs: Does the first sentence entail the second?**
- **Drawn from several of the RTE annual competitions.**

Text: *Dana Reeve, the widow of the actor Christopher Reeve, has died of lung cancer at age 44, according to the Christopher Reeve Foundation.*
Hypothesis: *Christopher Reeve had an accident.*
no-entailment

Corpus	|Train|	|Dev|	|Test|	Task	Metrics	Domain
				Single-Sentence Tasks		
CoLA						
SST-2						
MRPC						
STS-B						
QQP						
				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 ICLR '19

The Winograd Schema Challenge

NLI format, based on Levesque et al., 2011

- **Binary classification for expert-constructed pairs of sentences: What does the pronoun refer to?**
- **Manually constructed to foil superficial statistical cues.**
- **Private evaluation data used only in GLUE.**

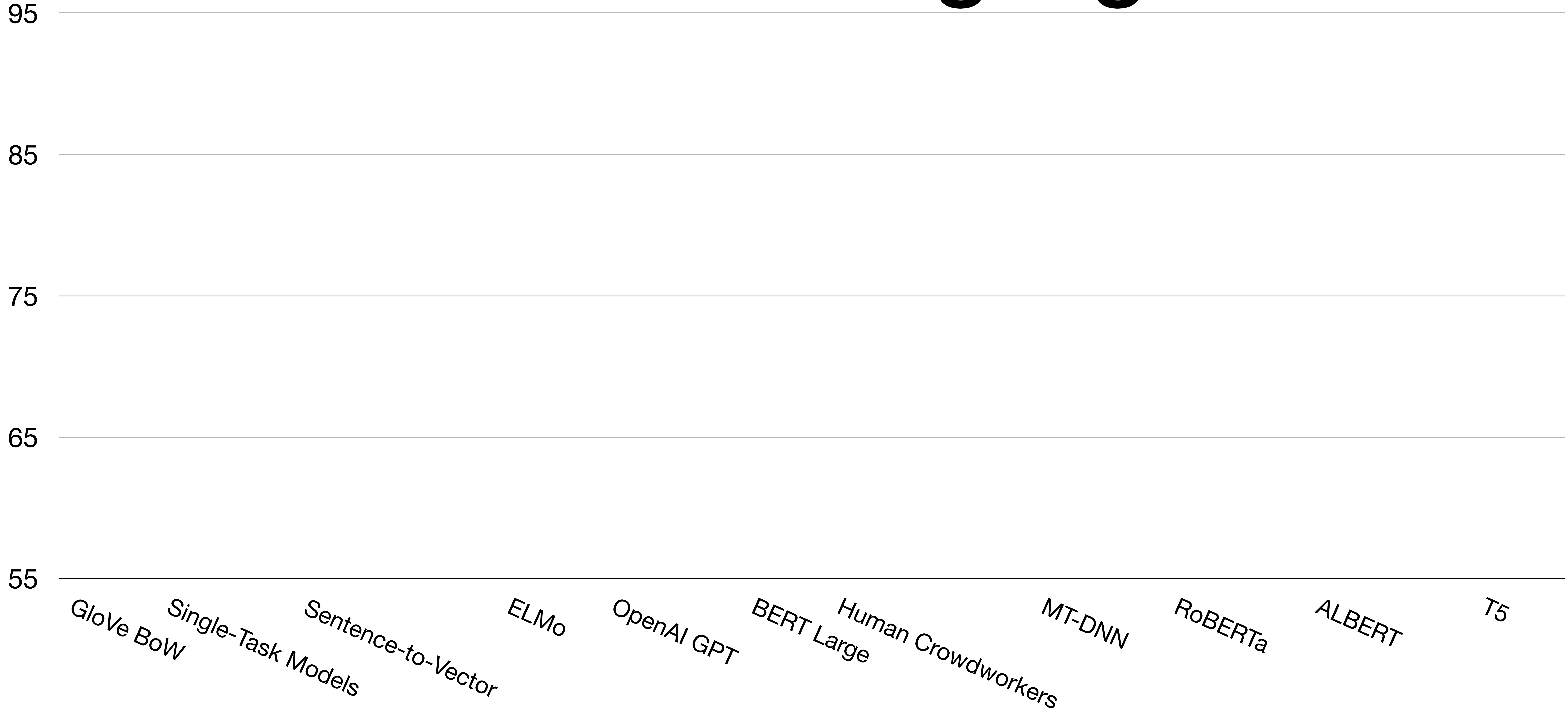
P: *Jane gave Joan candy because she was hungry.*
H: *Joan was hungry.*
entailment

Corpus	|Train|	|Dev|	|Test|	Task	Metrics	Domain
CoLA						
SST-2						
MRPC						
STS-B						
QQP						
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 ICLR '19

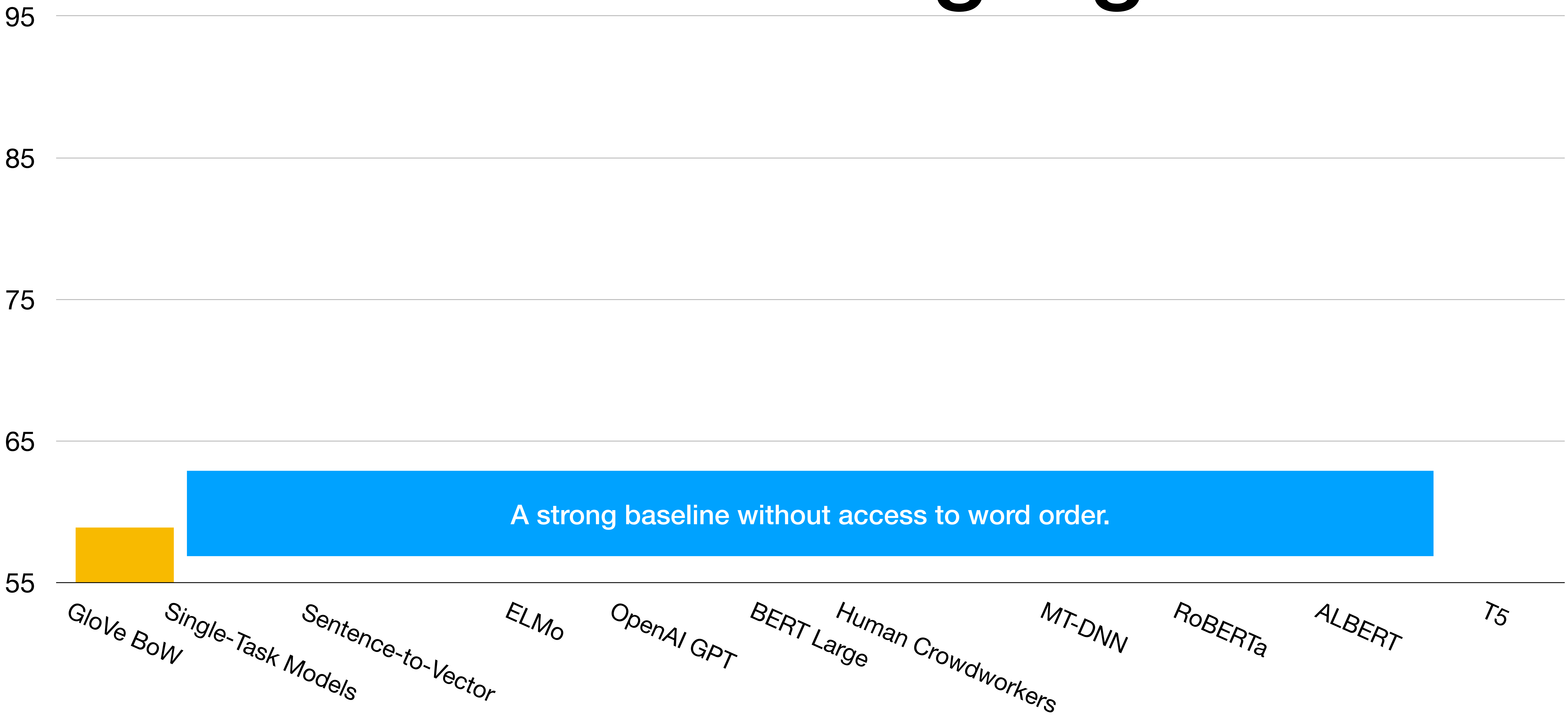
GLUE: What methods work?

GLUE Score: Highlights



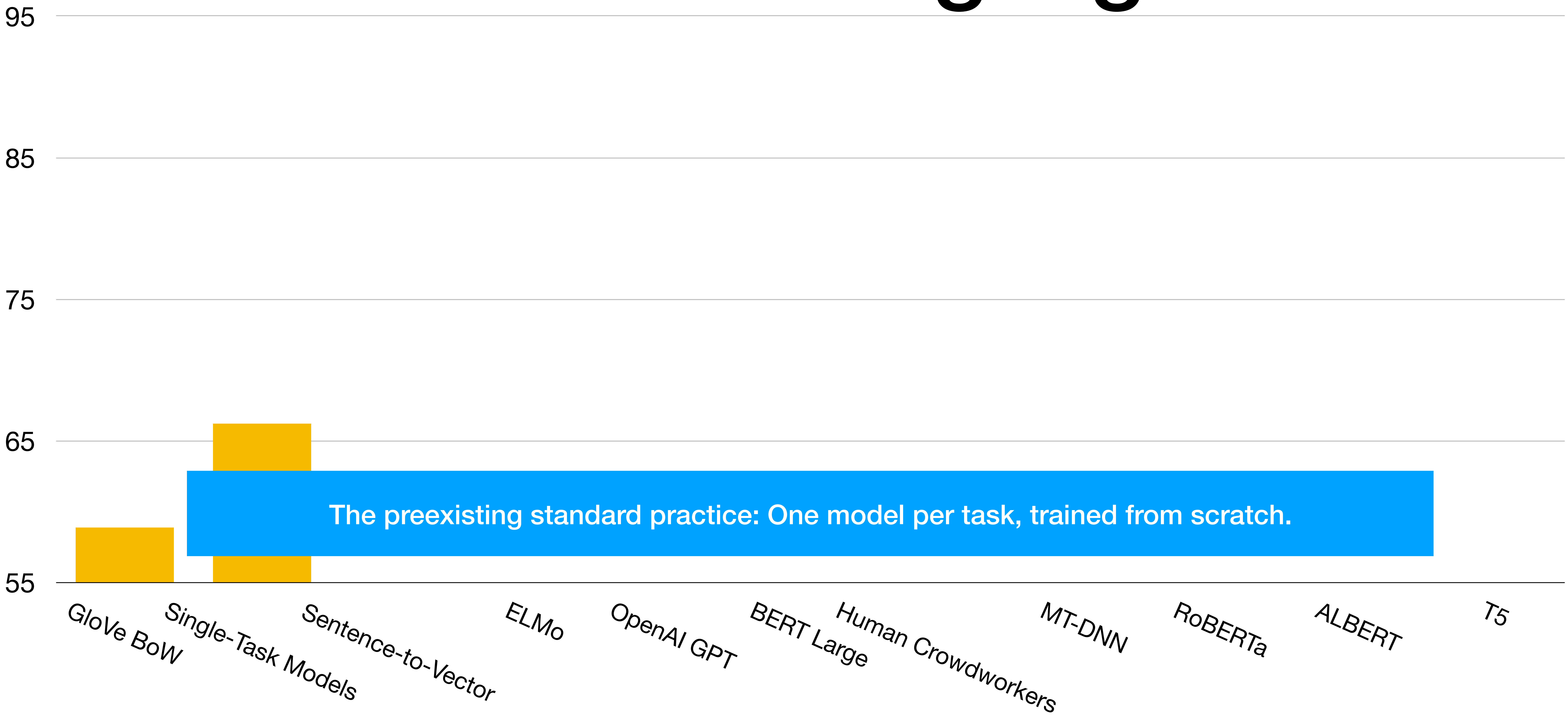
gluebenchmark.com

GLUE Score: Highlights



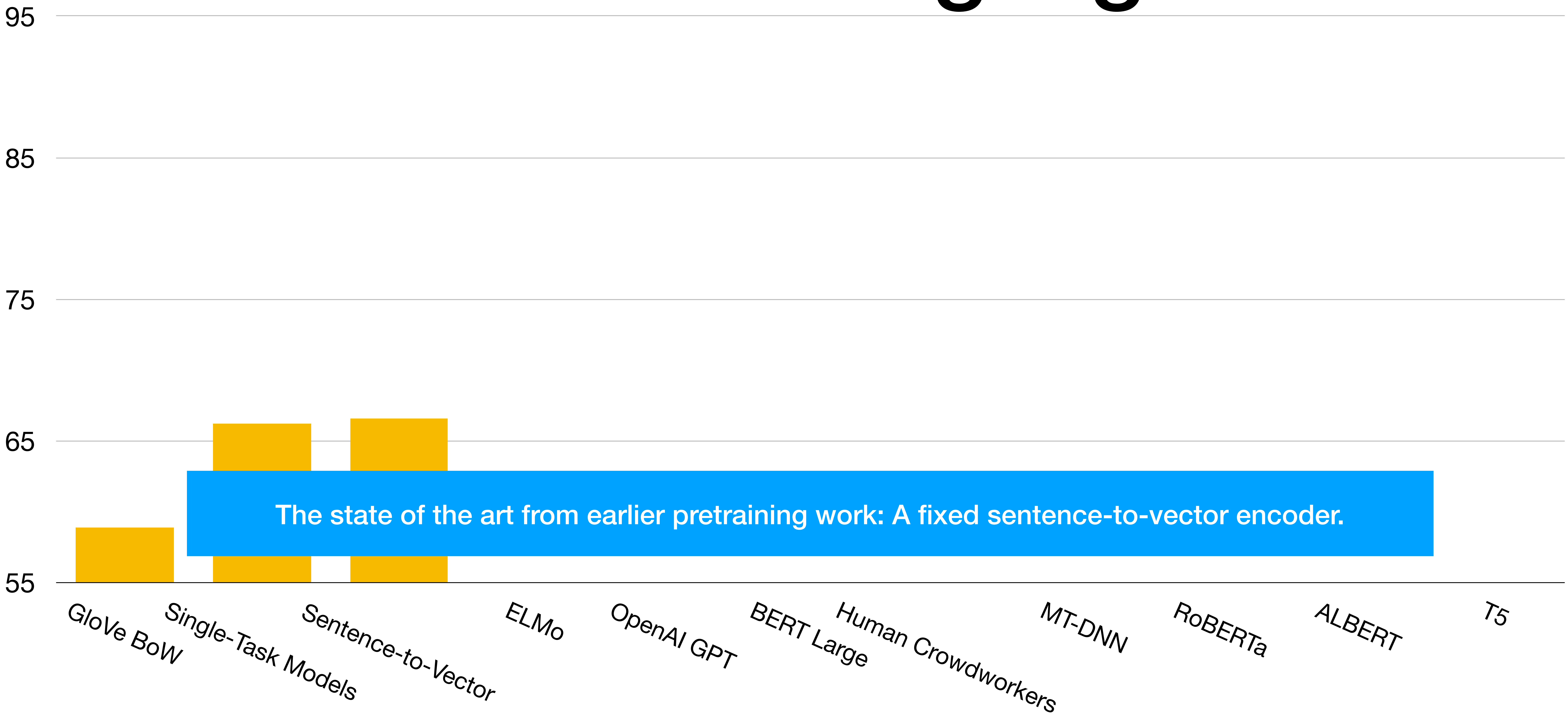
gluebenchmark.com

GLUE Score: Highlights



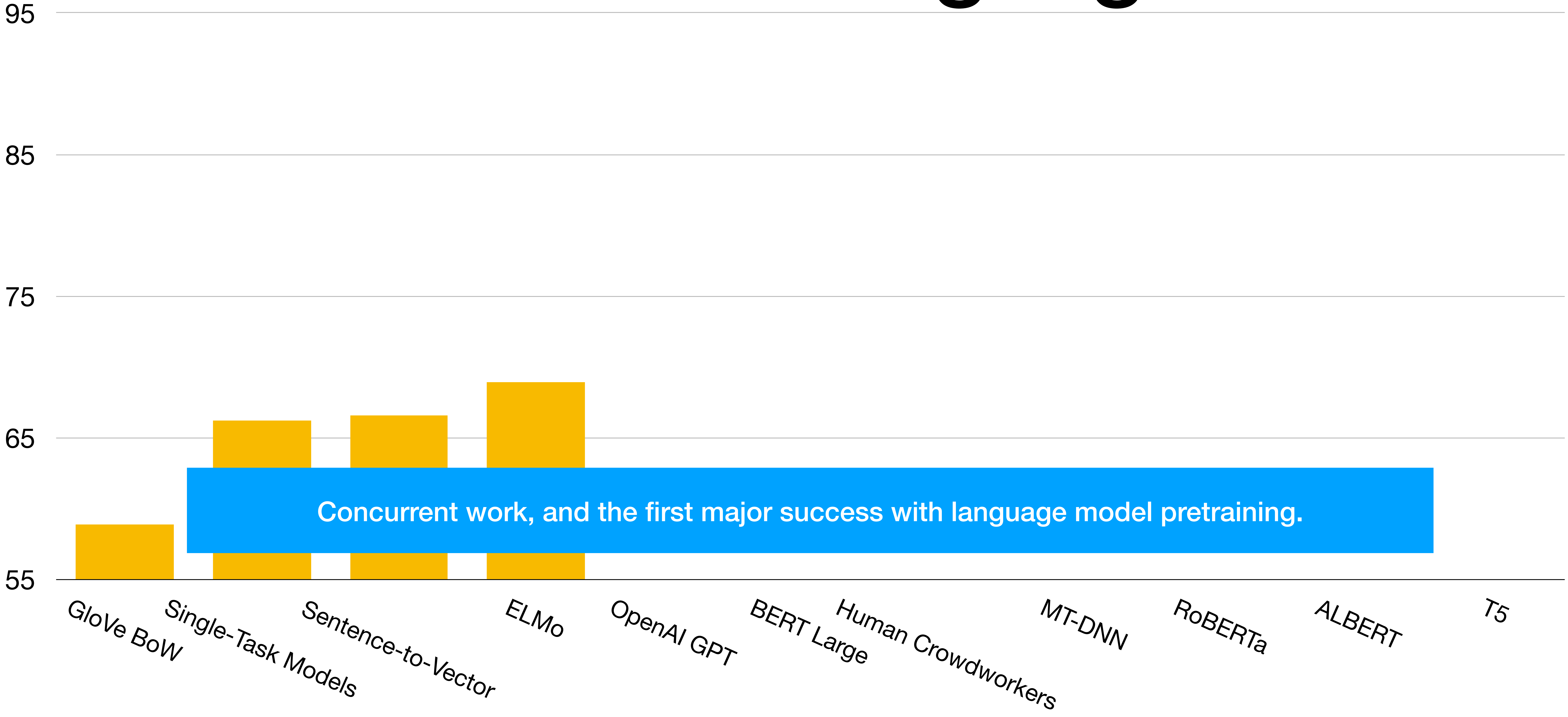
gluebenchmark.com

GLUE Score: Highlights



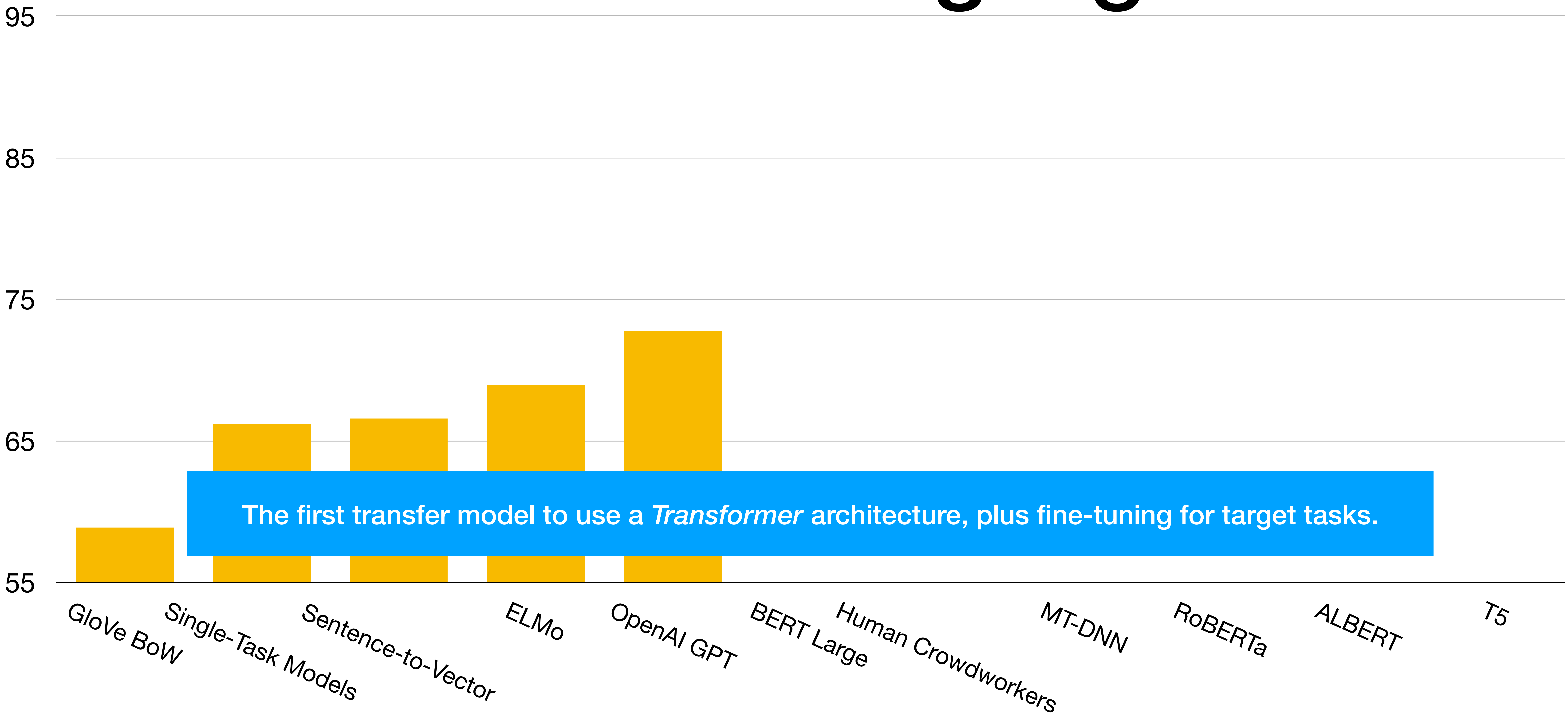
gluebenchmark.com

GLUE Score: Highlights

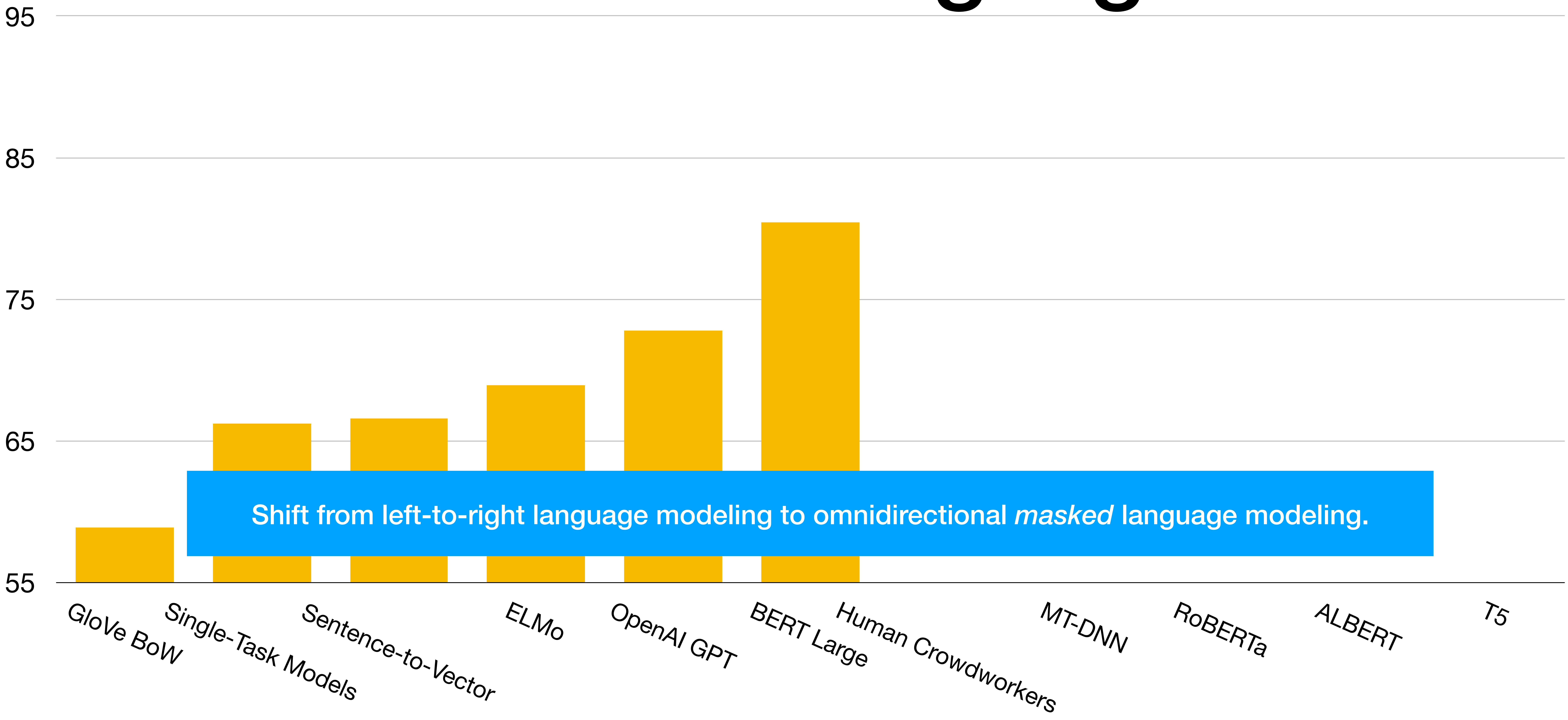


gluebenchmark.com

GLUE Score: Highlights



GLUE Score: Highlights



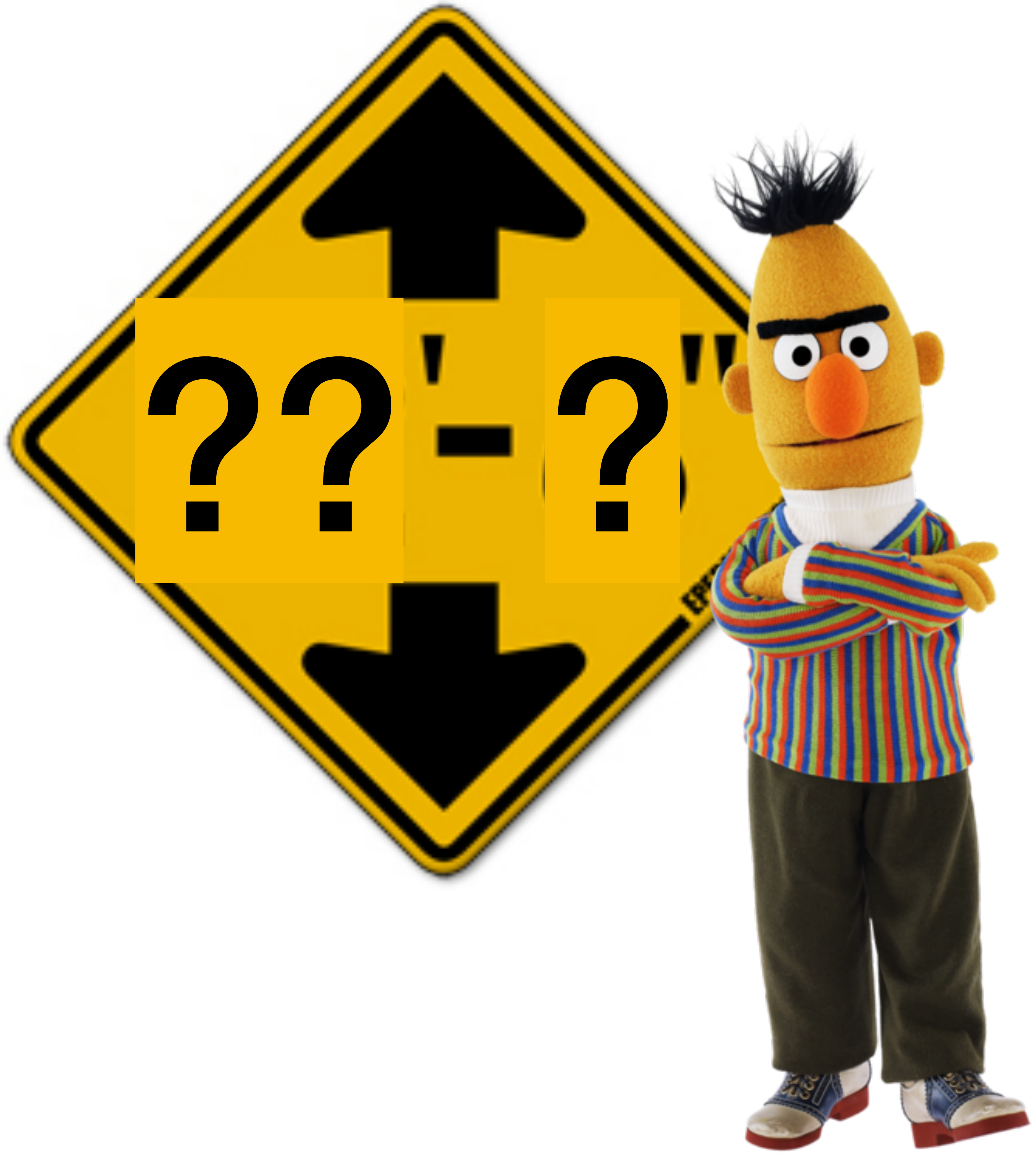
Shift from left-to-right language modeling to omnidirectional *masked* language modeling.

gluebenchmark.com

Human Performance Estimate

How much headroom does GLUE have left?

- To compute a conservative estimate for each task:
    - Train crowdworkers.

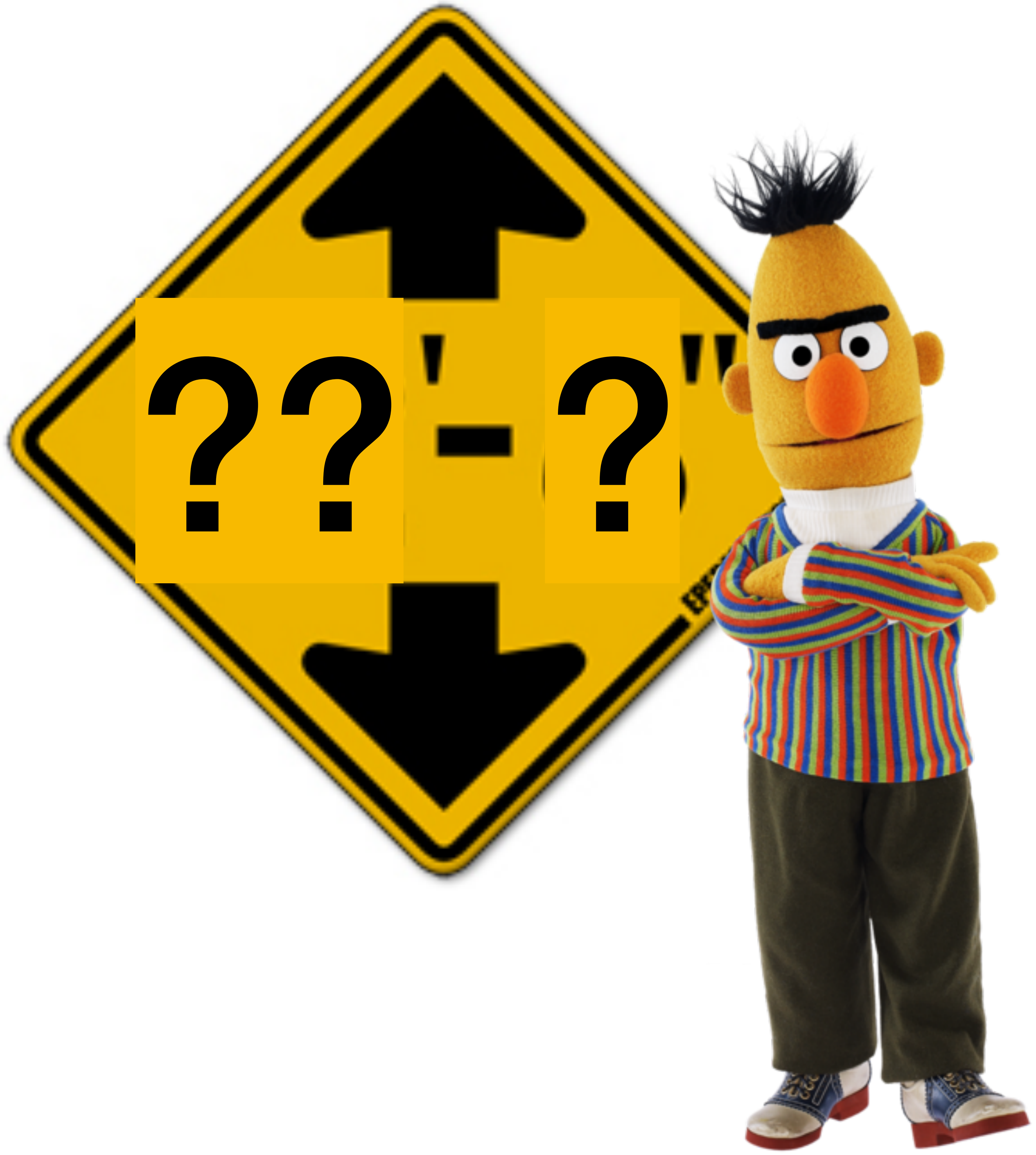


Nangia & Bowman '19

Human Performance Estimate

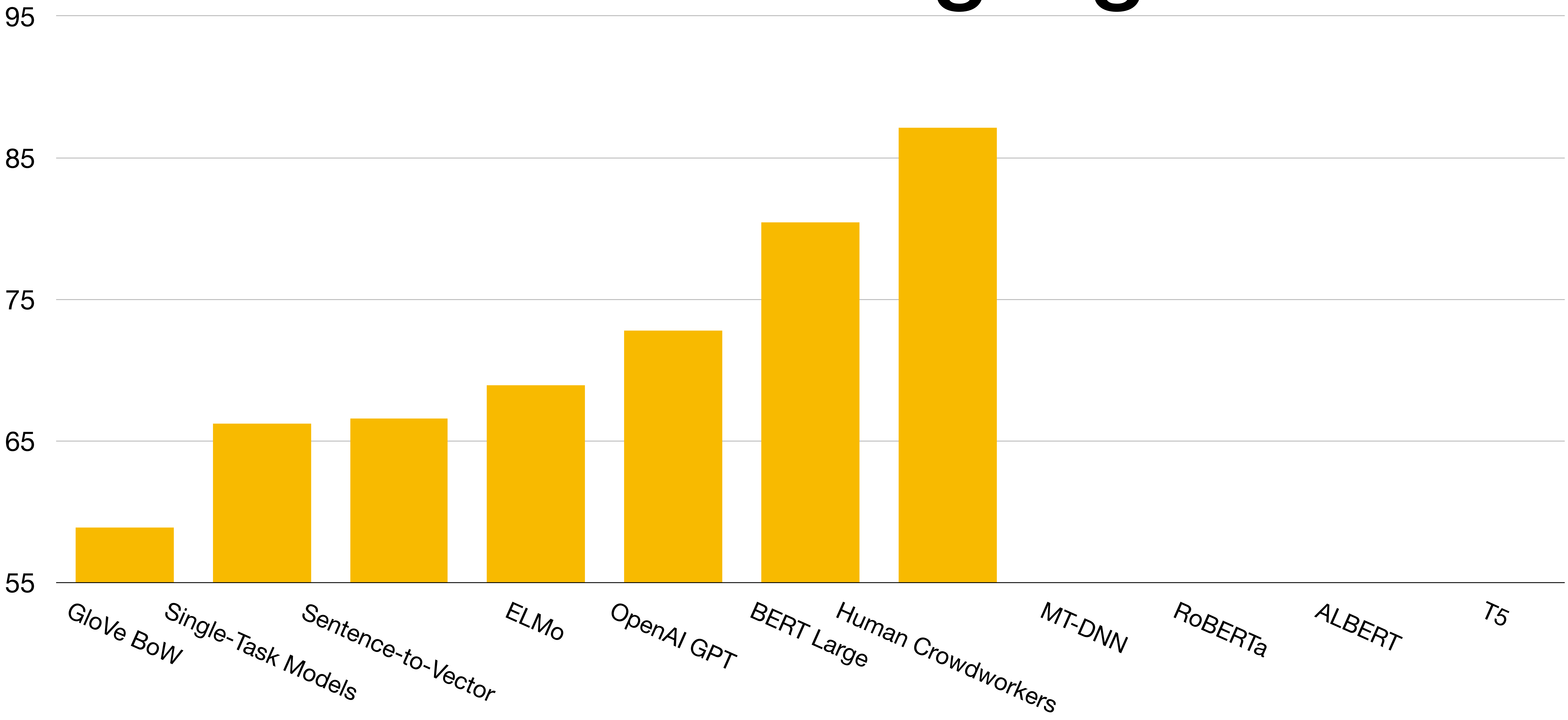
How much headroom does GLUE have left?

- To compute a conservative estimate for each task:
  - Train crowdworkers.
  - Get multiple crowdworker labels for each example, take a majority vote.



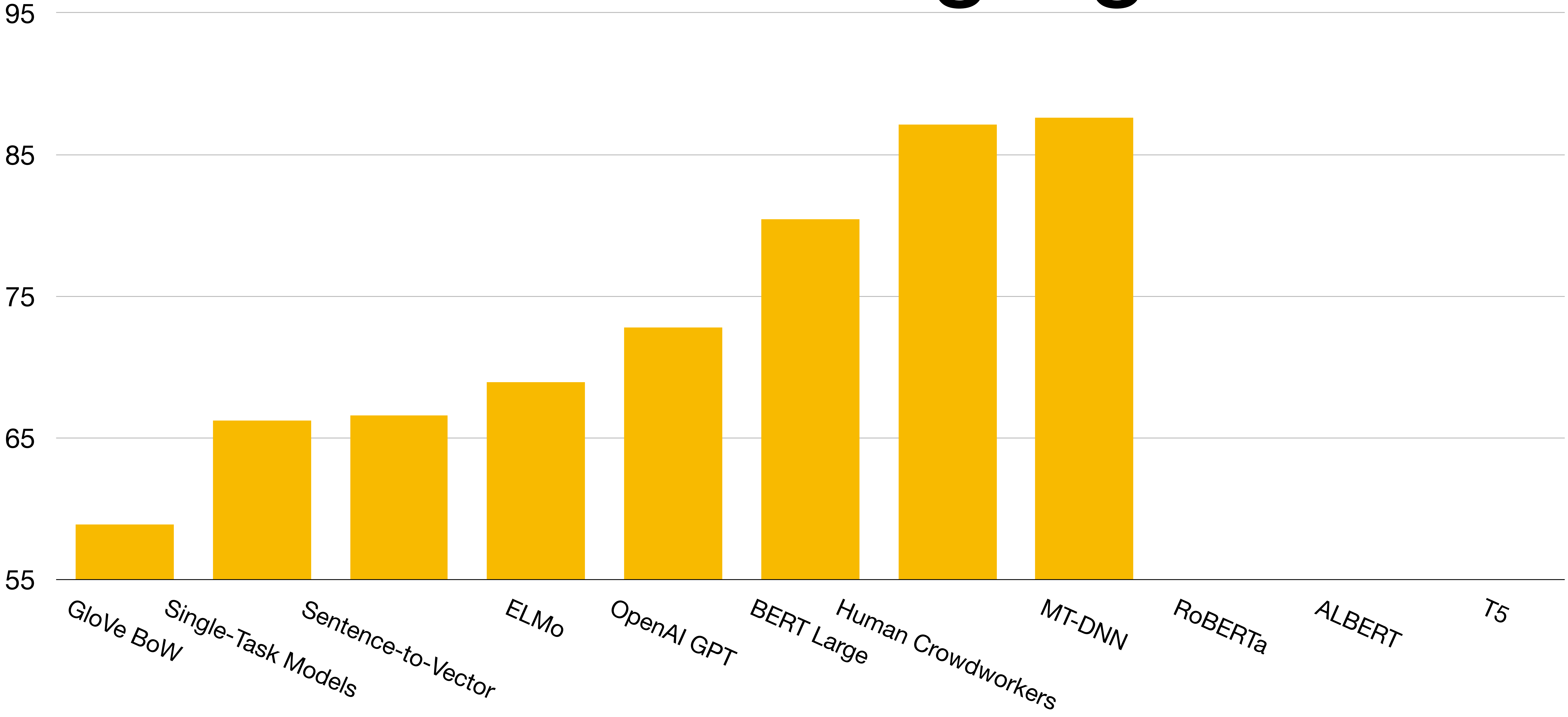
Nangia & Bowman '19

GLUE Score: Highlights



gluebenchmark.com

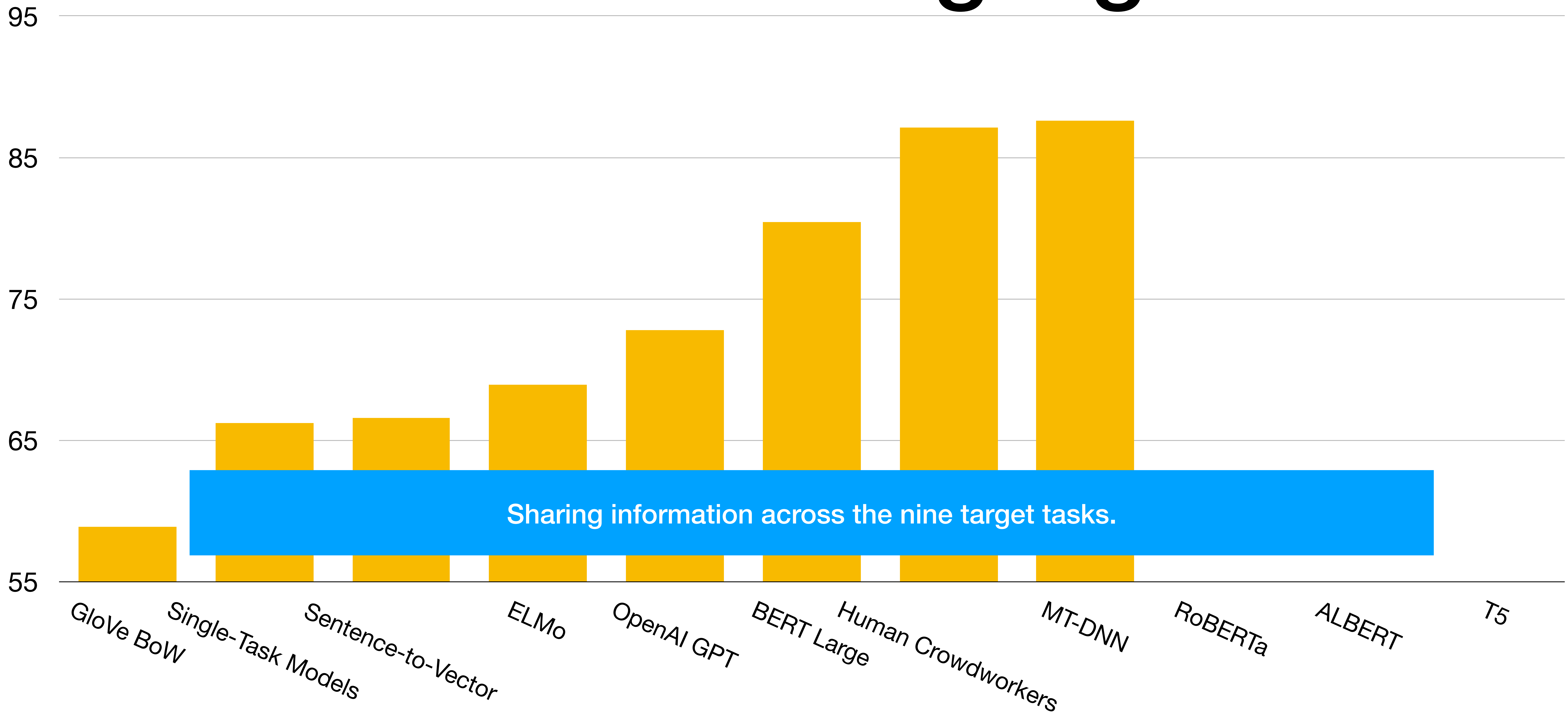
GLUE Score: Highlights



gluebenchmark.com

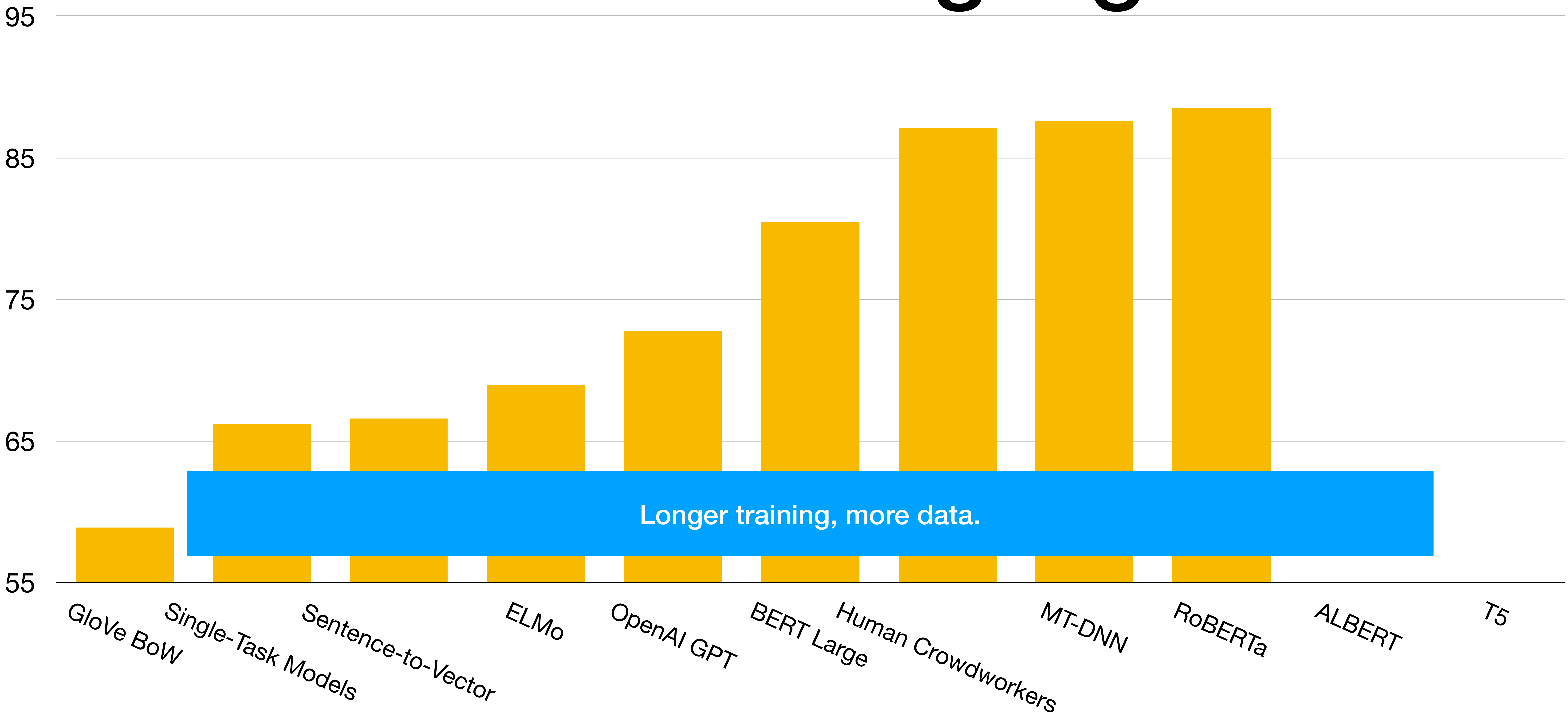

OVERHEIGHT
MUST TURN
11'-8"
11foot8.com

GLUE Score: Highlights



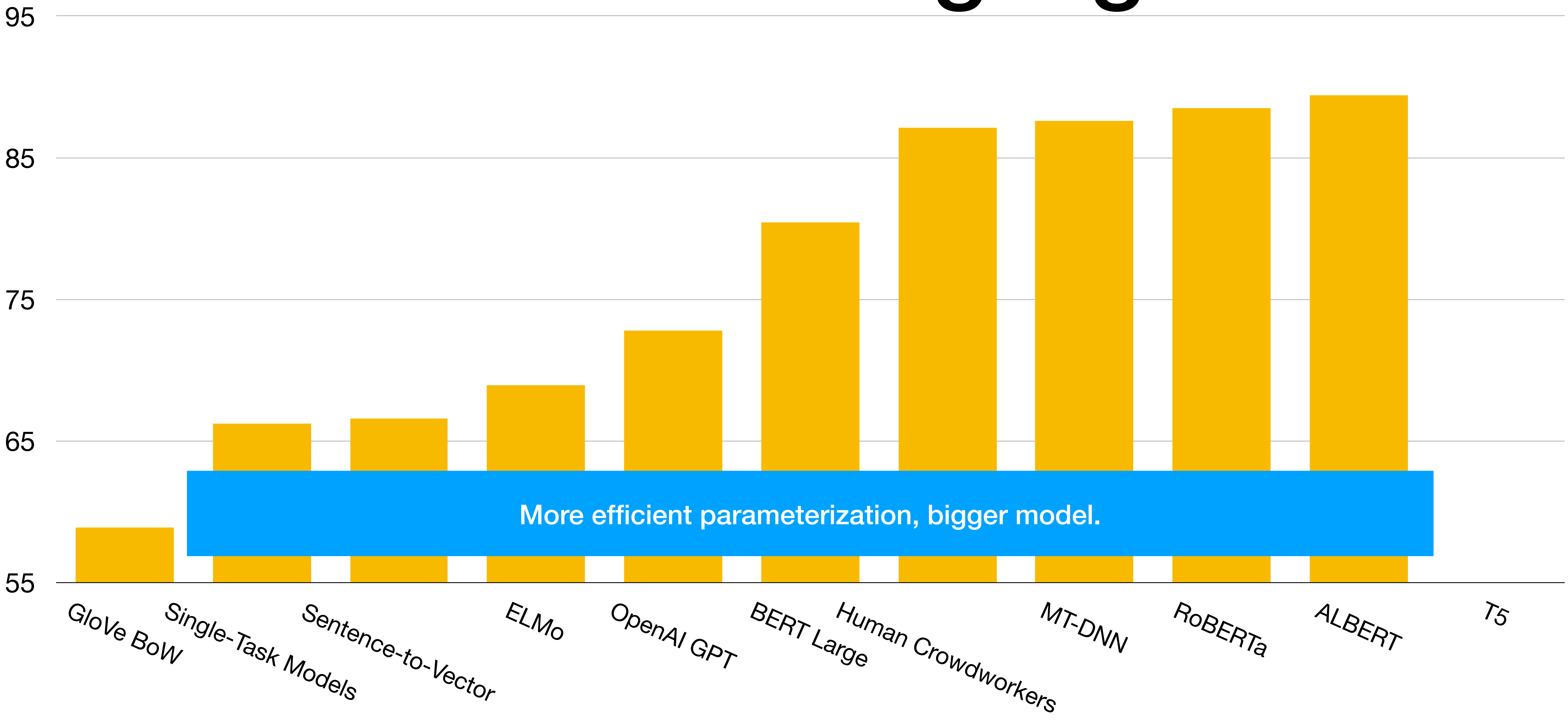
gluebenchmark.com

GLUE Score: Highlights



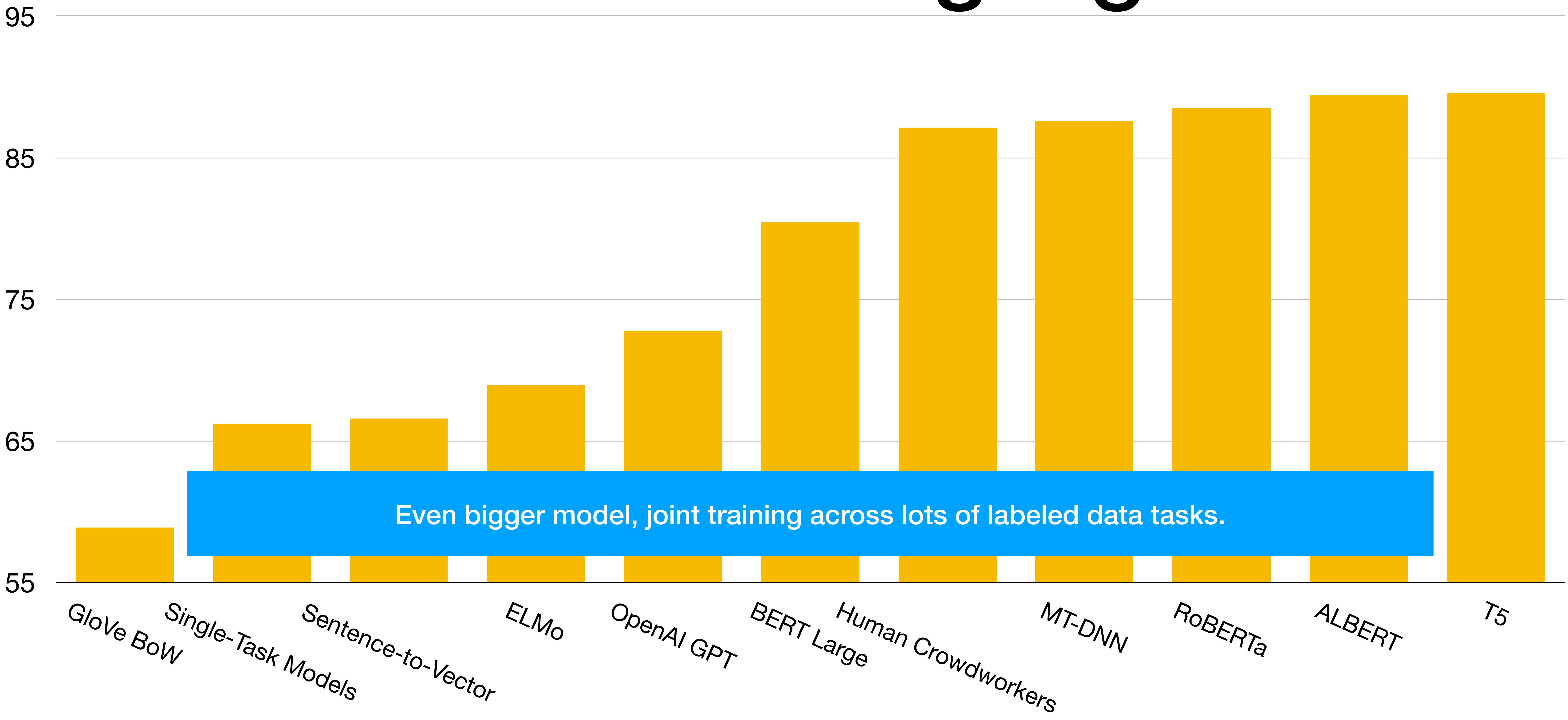
gluebenchmark.com

GLUE Score: Highlights



gluebenchmark.com

GLUE Score: Highlights



gluebenchmark.com

SuperGLUE

We rebuilt GLUE from scratch...

- ...starting with an open call for dataset proposals
- …yielding 30–40 candidates
- ...which we filtered using human evaluation and BERT-base baselines
- …and a final set of eight tasks
- ...following a slightly expanded set of task APIs.

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

SuperGLUE: The Main Tasks

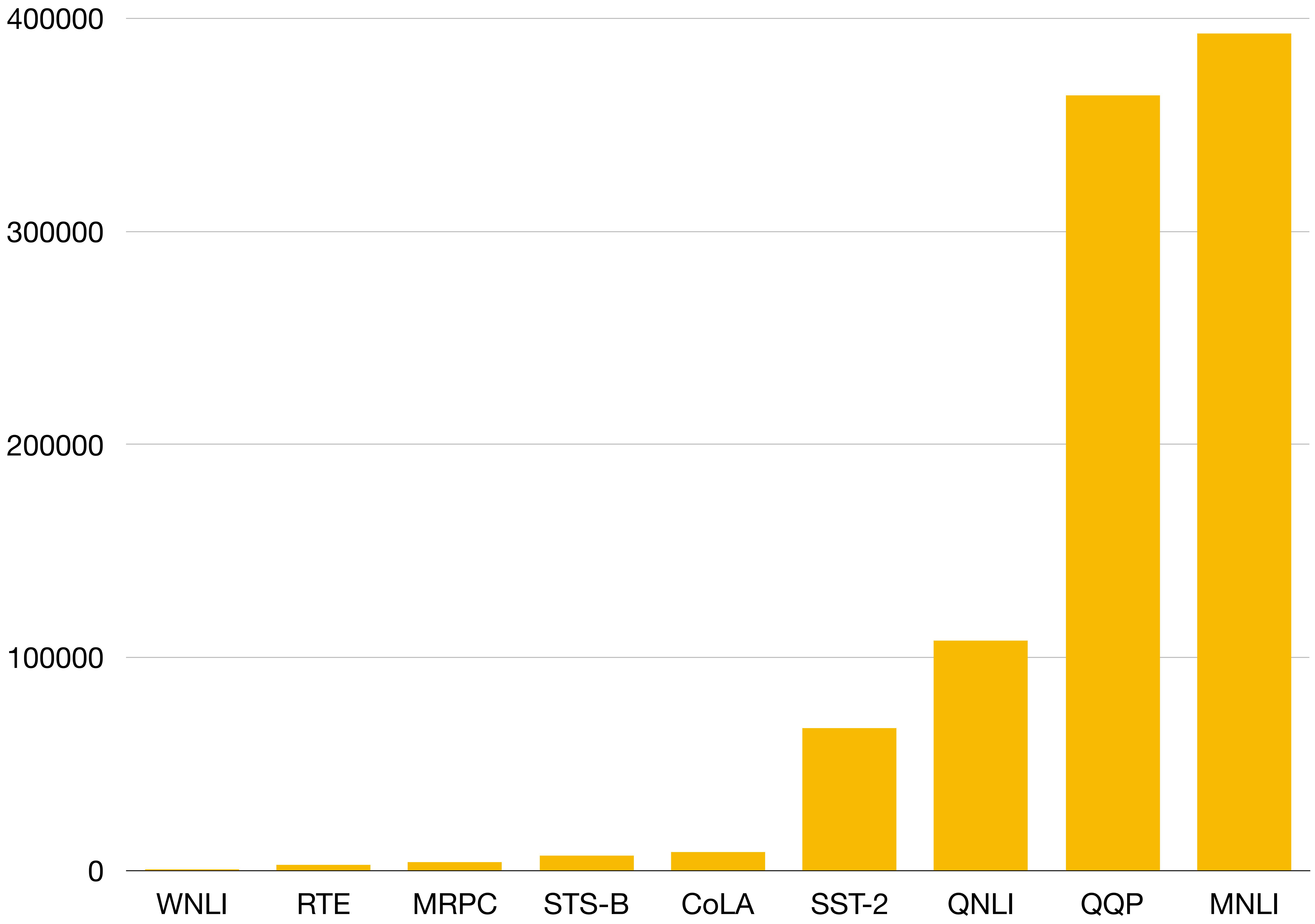
Corpus	|Train|	|Dev|	|Test|	Task	Metrics	Text Sources
BoolQ	9427	3270	3245	QA	acc.	Google queries, Wikipedia
CB	250	57	250	NLI	acc./F1	various
COPA	400	100	500	QA	acc.	blogs, photography encyclopedia
MultiRC	5100	953	1800	QA	$F1_a$/EM	various
ReCoRD	101k	10k	10k	QA	F1/EM	news (CNN, Daily Mail)
RTE	2500	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 NeurIPS '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 NeurIPS '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

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 NeurIPS '19

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

  no-entailment

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	va
ReCoRD	101k	10k	10k	QA	F1/EM	ne

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

MultiRC

Khashabi et al. '18

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

Paragraph: *Susan wanted to have a birthday party. She called all of her friends. She has five friends. Her mom said that Susan can invite them all to the party. Her first friend could not go to the party because she was sick. Her second friend was going out of town. Her third friend was not so sure if her parents would let her. The fourth friend said maybe. The fifth friend could go to the party for sure. Susan was a little sad. On the day of the party, all five friends showed up. Each friend had a present for Susan. Susan was happy and sent each friend a thank you card the next week.*
Question: *Did Susan's sick friend recover?*
Answers: *Yes, she recovered (T), No (F), Yes (T), No, she didn't recover (F), Yes, she was at Susan's party (T)*

COPA	400	100	500	QA	acc.	blogs, photography encyclopedia
MultiRC	5100	953	1800	QA	$F1_a$/EM	va
ReCoRD	101k	10k	10k	QA	F1/EM	ne
RTE	2500	278	300	NLI	acc.	ne

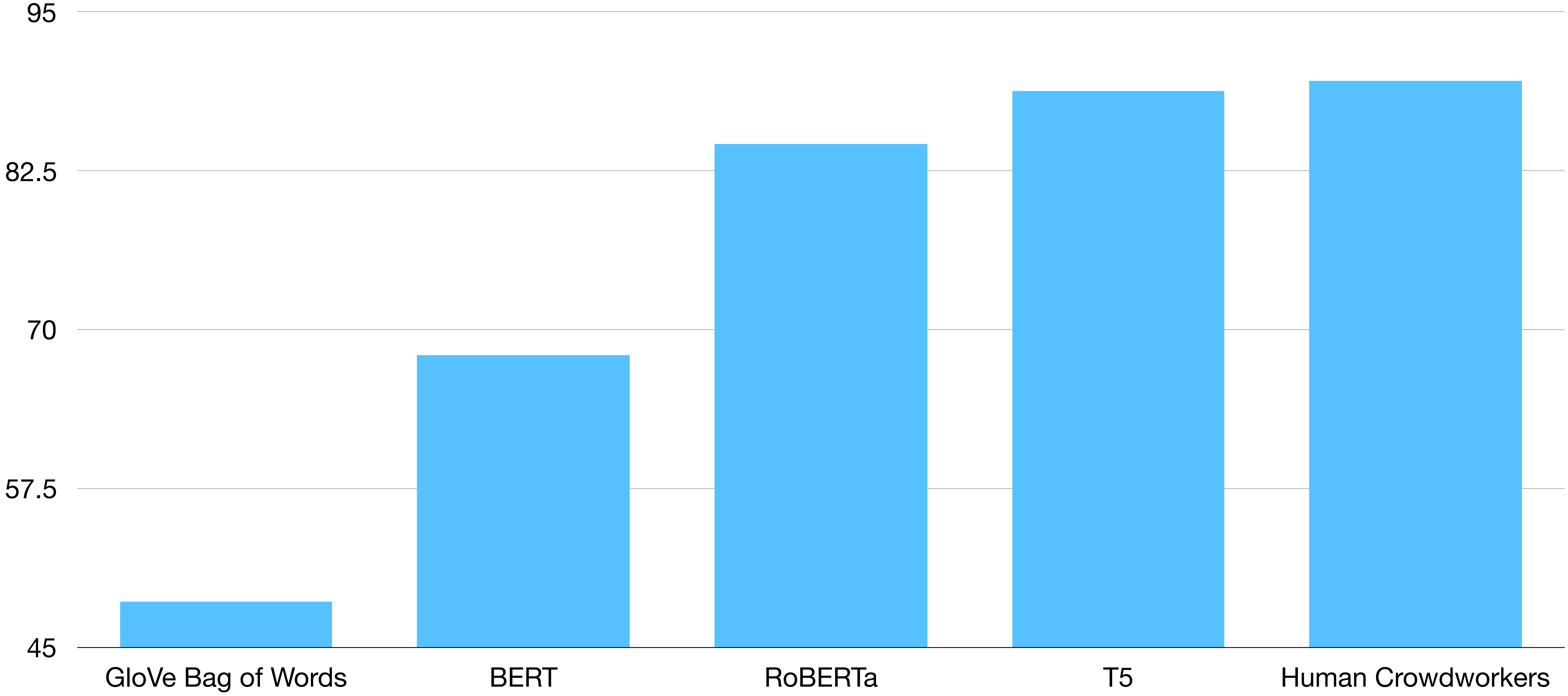
{Wang, Pruksachatkun, Nangia, Singh}, Michael, Hill, Levy & Bowman NeurIPS '19

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 NeurIPS '19

SuperGLUE Score: Highlights



super.gluebenchmark.com

⚠ 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: Limitations

GLUE and SuperGLUE use lots of naturally occurring or 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), T5 is 10x more like than humans to be confused by irrelevant gender cues.
- Mitigating these biases is a major open problem.

GLUE and SuperGLUE: Non-Limitations

GLUE and SuperGLUE don't test generation or structured prediction.

- These are hand and important problems, but mostly orthogonal to language understanding.

GLUE and SuperGLUE: Open Issues

GOOD LUCK

10-point gap between humans and T5!

We clearly haven't solved NLU.

SuperGLUE includes a broad-coverage NLI diagnostic:

Prepositional phrases section

I ate pizza with olives.
I ate olives.
entailment

I ate pizza with some friends.
I ate some friends.
neutral

GLUE and SuperGLUE: Open Issues

We can be pretty sure we haven't solved NLU *even for IID evaluations*.

- 6-point gap between T5 and humans on Winograd Schemas.
- *In-domain* evaluation for NLI, QA, etc., involves lots of phenomena that we know models aren't great at. Are these differences just drowned in the noise?

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

*Yes, BERT.

What's inside BERT?

In our work on *Edge Probing* (Tenney et al.), we observe that:

- ELMo and BERT both learn nearly perfect features for POS tagging.
- BERT learns better features than ELMo for parsing.
- ELMo and BERT Base do not learn coreference features, but BERT Large does.

What’s inside BERT?

What's inside BERT?

In further edge probing studies (Tenney, Das, and Pavlick):

- Lower layers of BERT express features for 'lower level' tasks.
- Higher layers express more abstract/ semantic knowledge.

What’s inside BERT?

Structural probes (Hewitt and Manning):

- The geometry of BERT's activation vectors encode some syntactic structure.

What's inside BERT?

Evaluations on *handbuilt test sets* (Yaghoobzadeh et al.):

- BERT relies on brittle non-syntactic heuristics for tasks like NLI; but BERT Large much less so than BERT Base.

How much can we trust these conclusions?

How much can we trust these conclusions?

- Probing studies (loosely defined) like these 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 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 EMNLP '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.

- Well-characterized in the linguistics literature.
- Depends on long-distance dependencies and complex structures, rather than local co-occurrence.

Does BERT know where NPIs are licensed?

I see kids who are not [eating ***any*** *cookies].*

I see* ***any* *kids who are not [eating cookies].*

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

Case Study: NPI Licensing

NPI words like *any* or *ever* can only occur in the scope of specific linguistic *licensing environments* like negation or conditionals.

- Well-characterized in the linguistics literature.

- Depends on long-distance dependencies and complex structures, rather than local co-occurrence.

Does BERT know where NPIs are licensed?

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

*I see kids who are not [eating **any** cookies].*

I see **any kids who are not [eating cookies].*

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

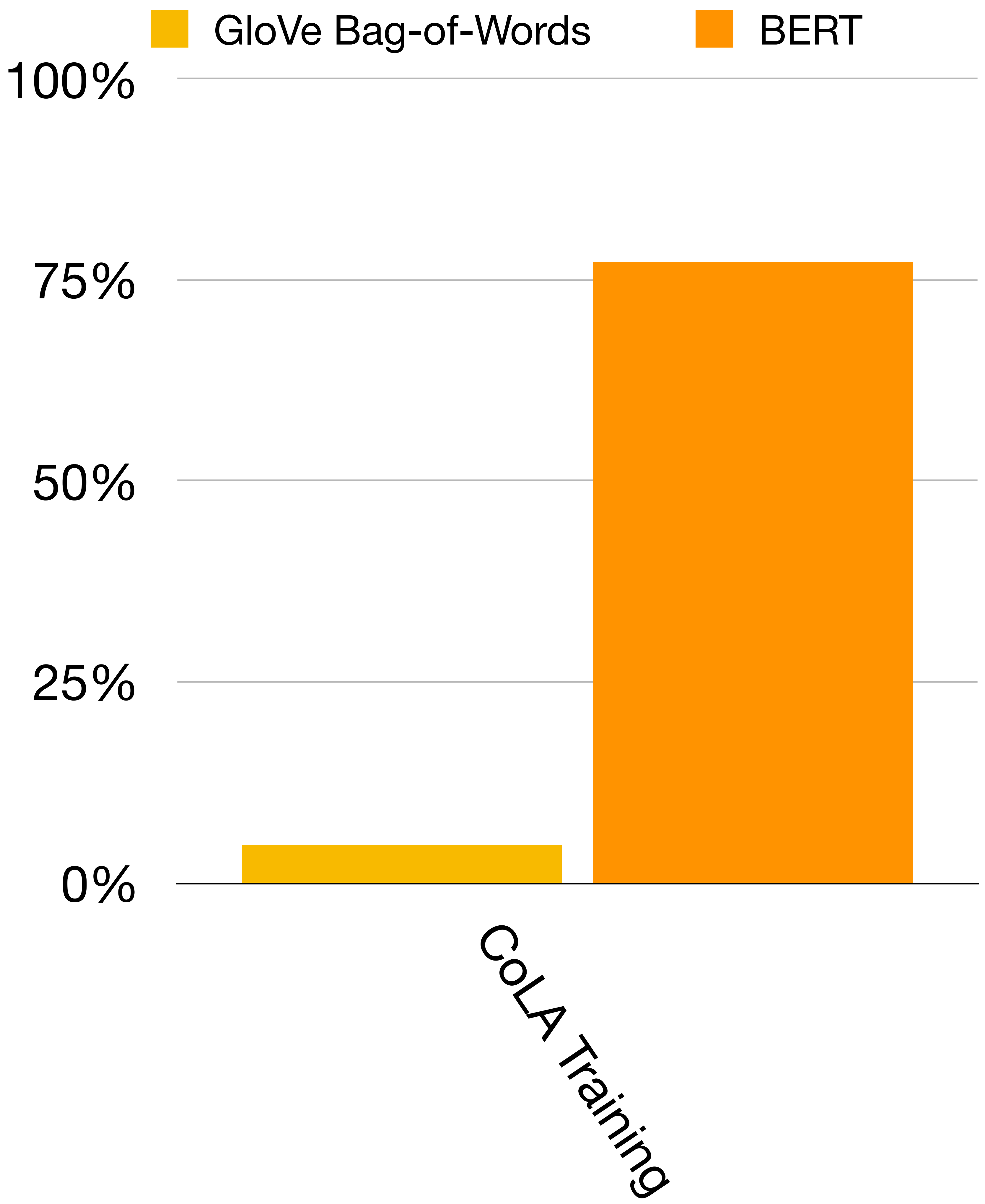
Let's teach the model to judge acceptability.

*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.

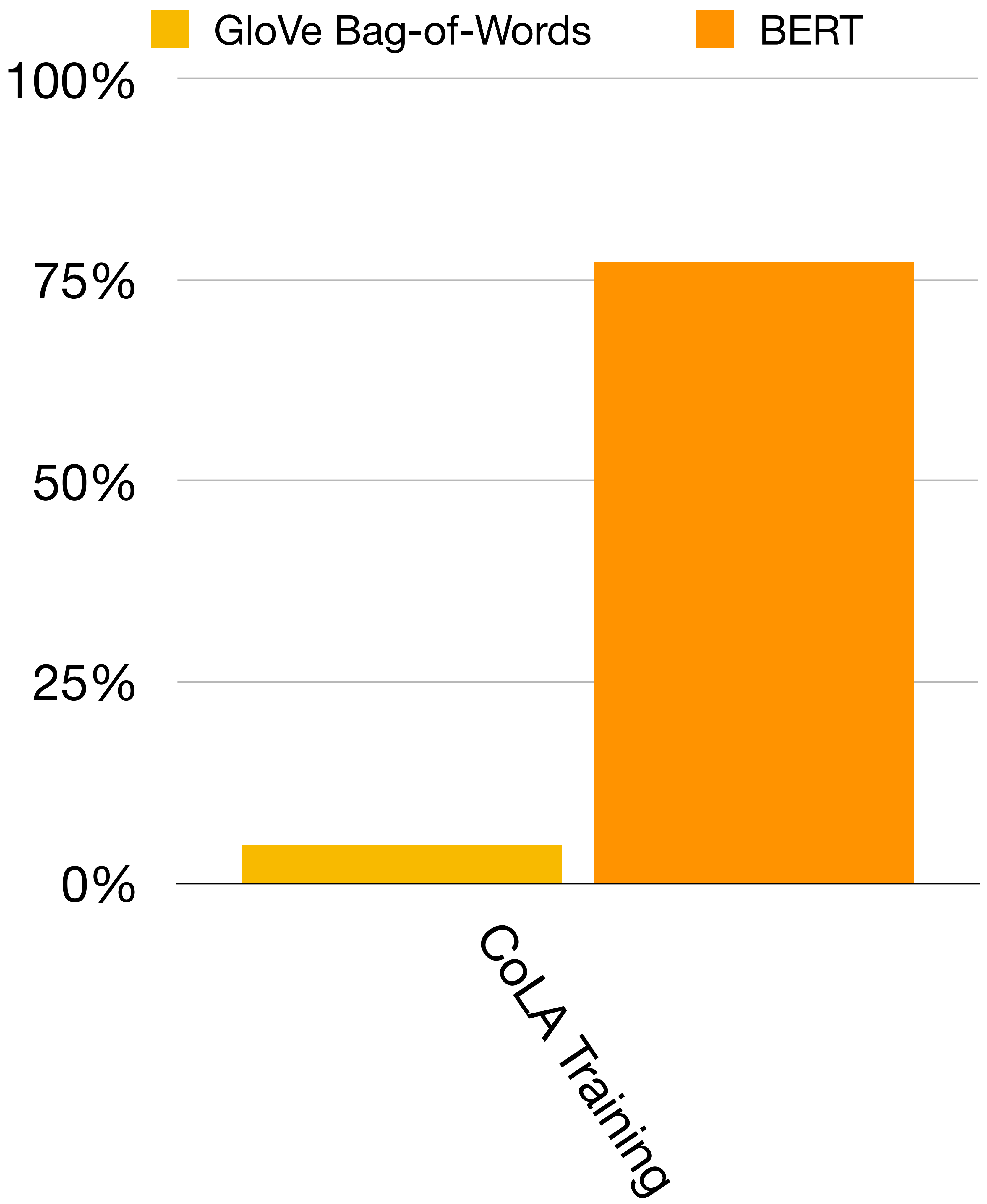
BERT knows a bit about NPIs,
but its not perfect.

*Whe
*Usually, any lion is majest
The gardener planted ro the garden.
I wrote Blair a letter 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



What if we train on NPI data directly?

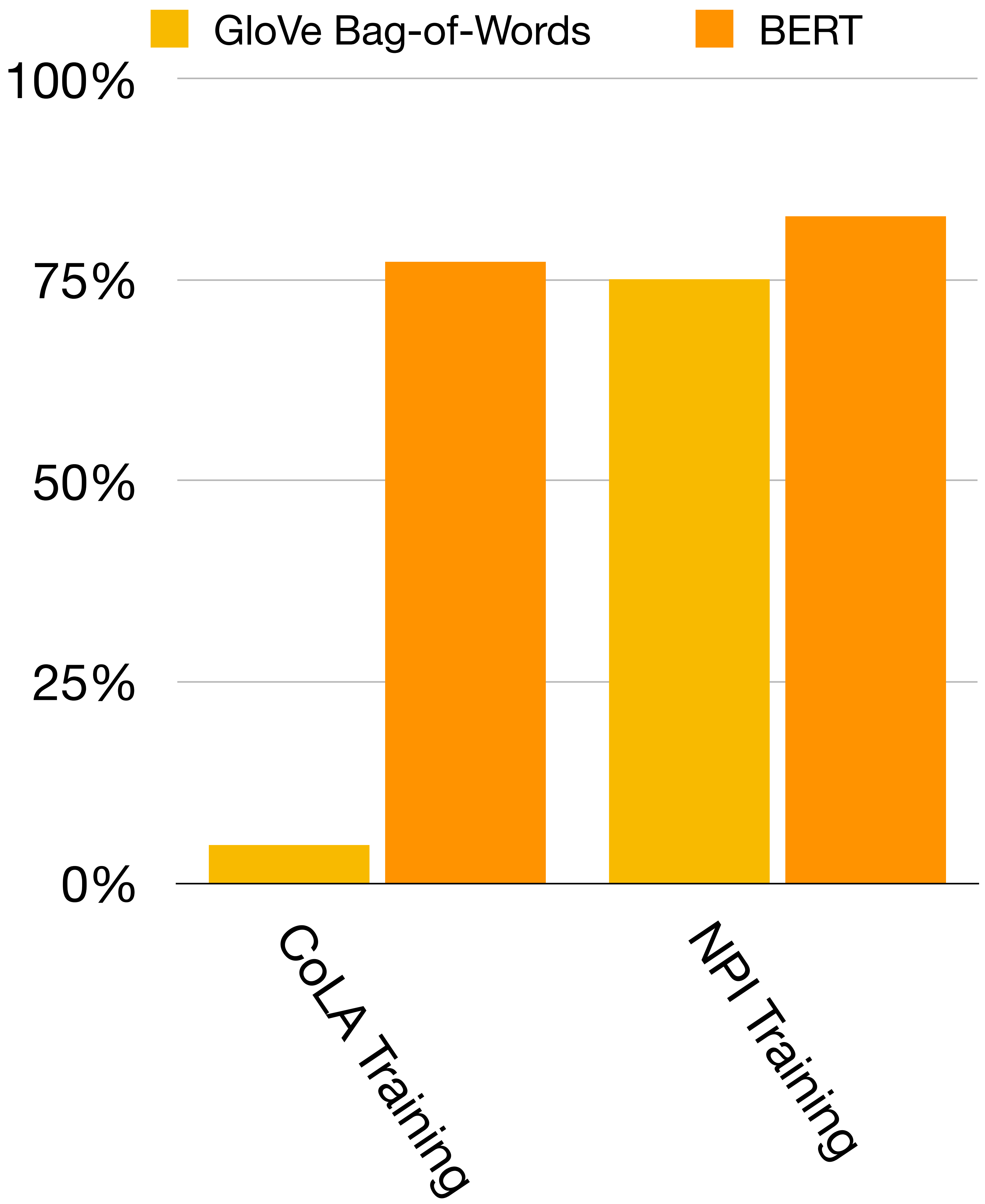
*Those boys say **that** [the doctors *ever* went to an art gallery.]
*Those boys *ever* say **that** [the doctors went to an art gallery.]
Those boys say **that** [the doctors *often* went to an art gallery.]
Those boys *often* say **that** [the doctors went to an art gallery.]

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

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.]

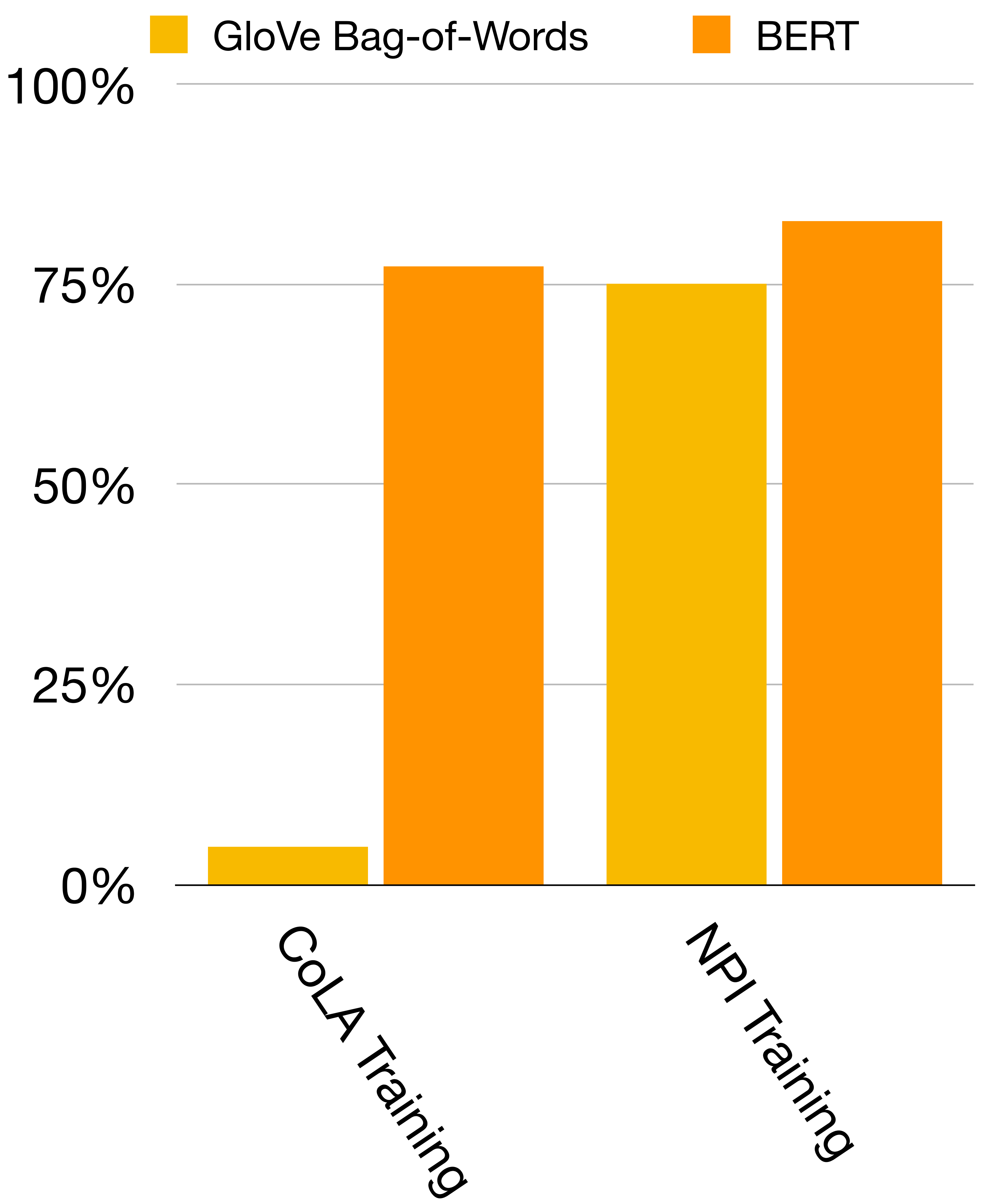
*Whe...
*Usually, any lion is majest...
The gardener planted r... the garden.
I wrote Blair a letter ... tore it up before I sent it.

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

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

Test:
NPI environment test sets

Metric:
Matthews Correlation (MCC) for acceptability



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

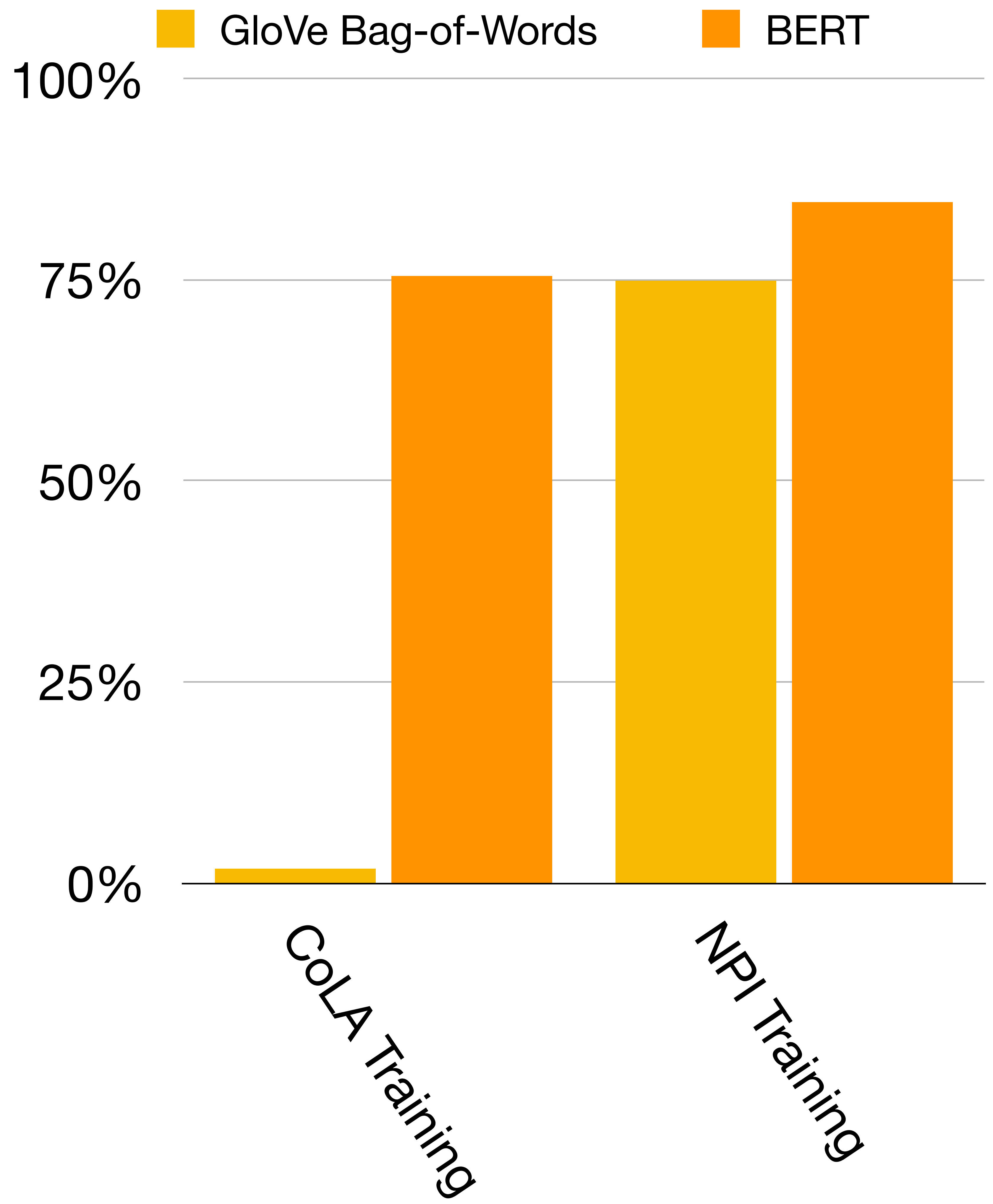
(1) Mary has**n't** eaten *any* cookies.

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

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

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...

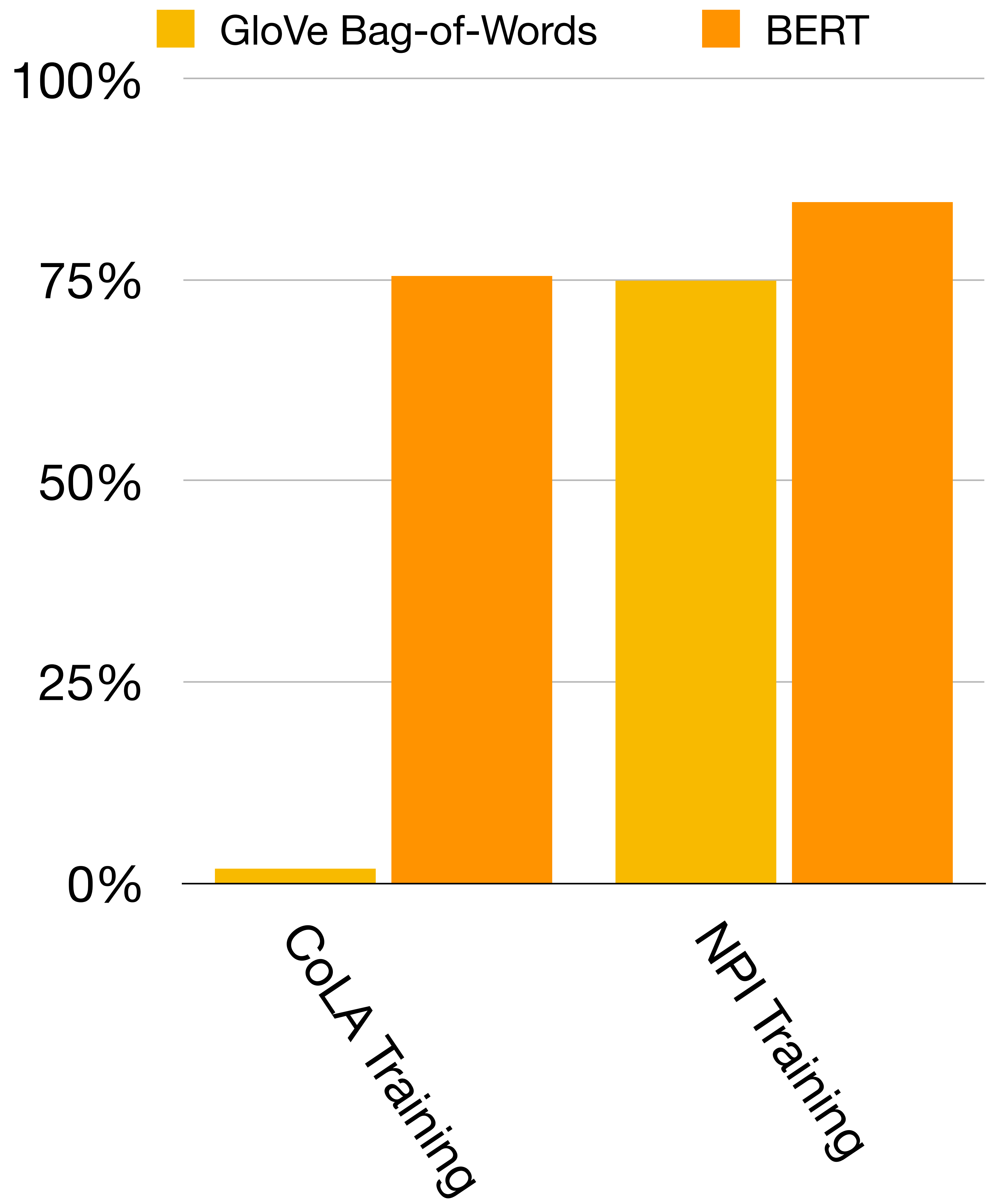
BERT knows something about NPIs,
but not all that much.

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

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

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...

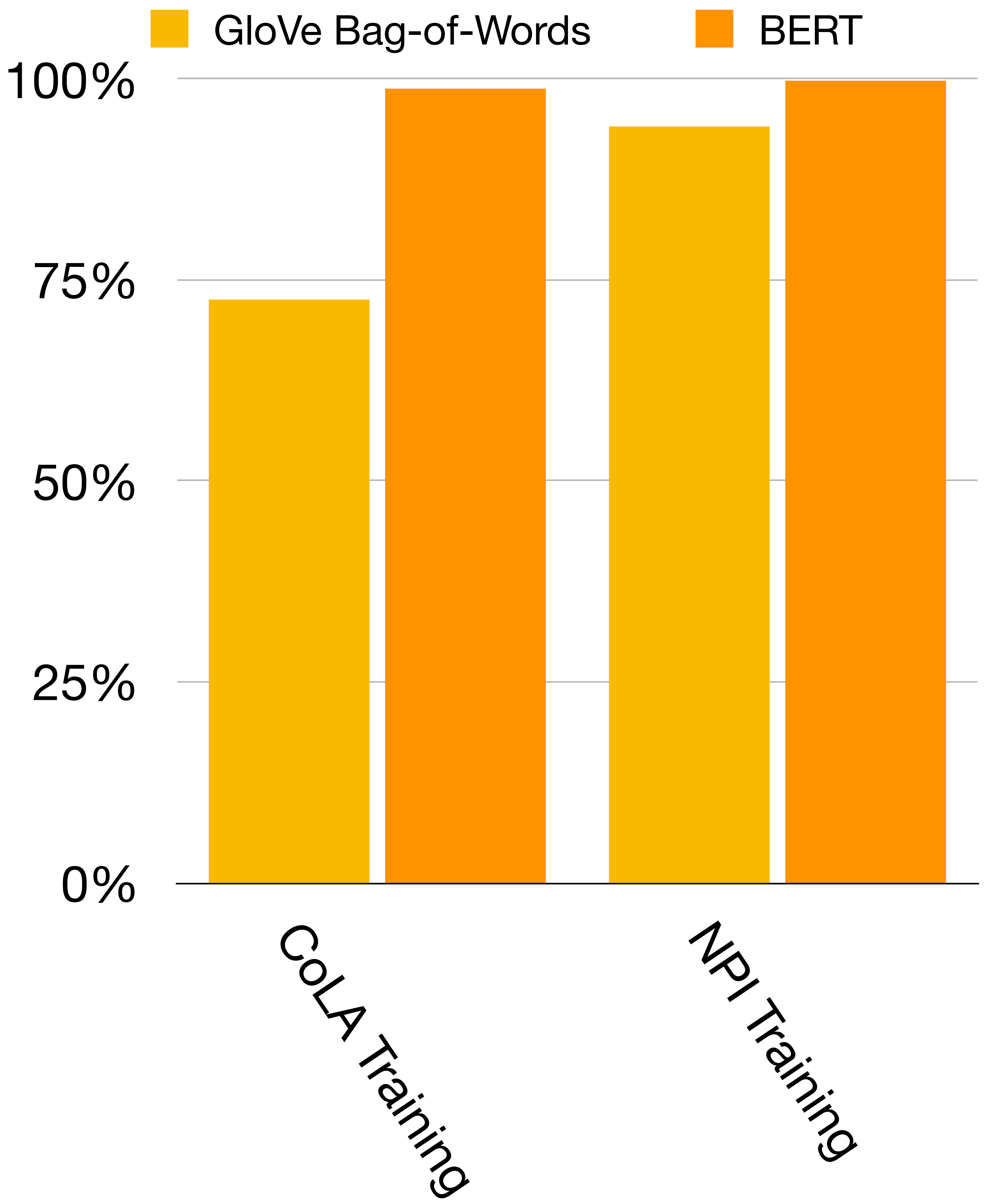
(1) Mary hasn**'t** eaten *any* cookies.

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

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

Test:
NPI environment test sets

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



MIGHTY JAXX SESAME STREET TM and © 2019 Sesame Workshop

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

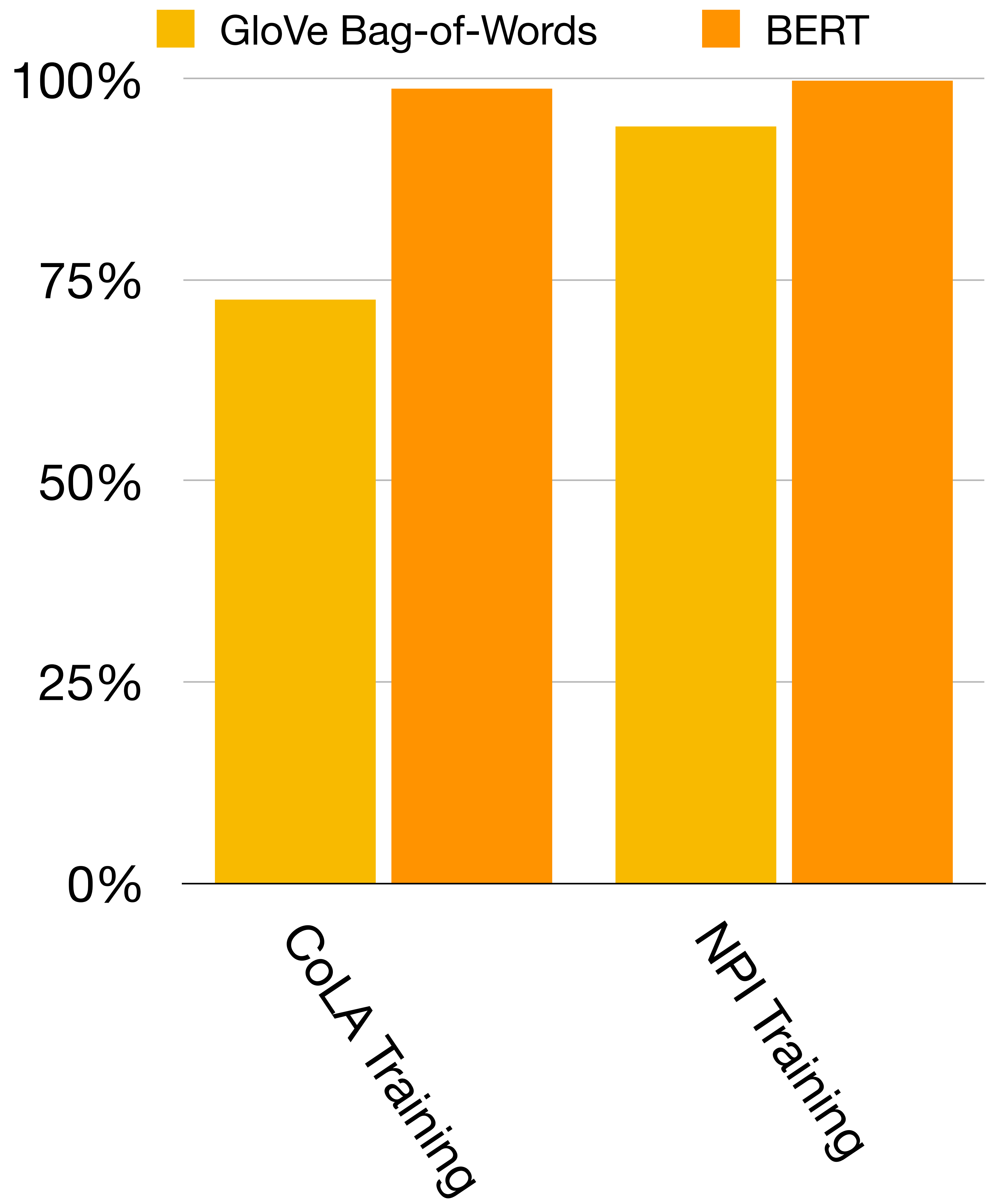
BERT has complete and perfect knowledge of NPI licensing.

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

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

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?

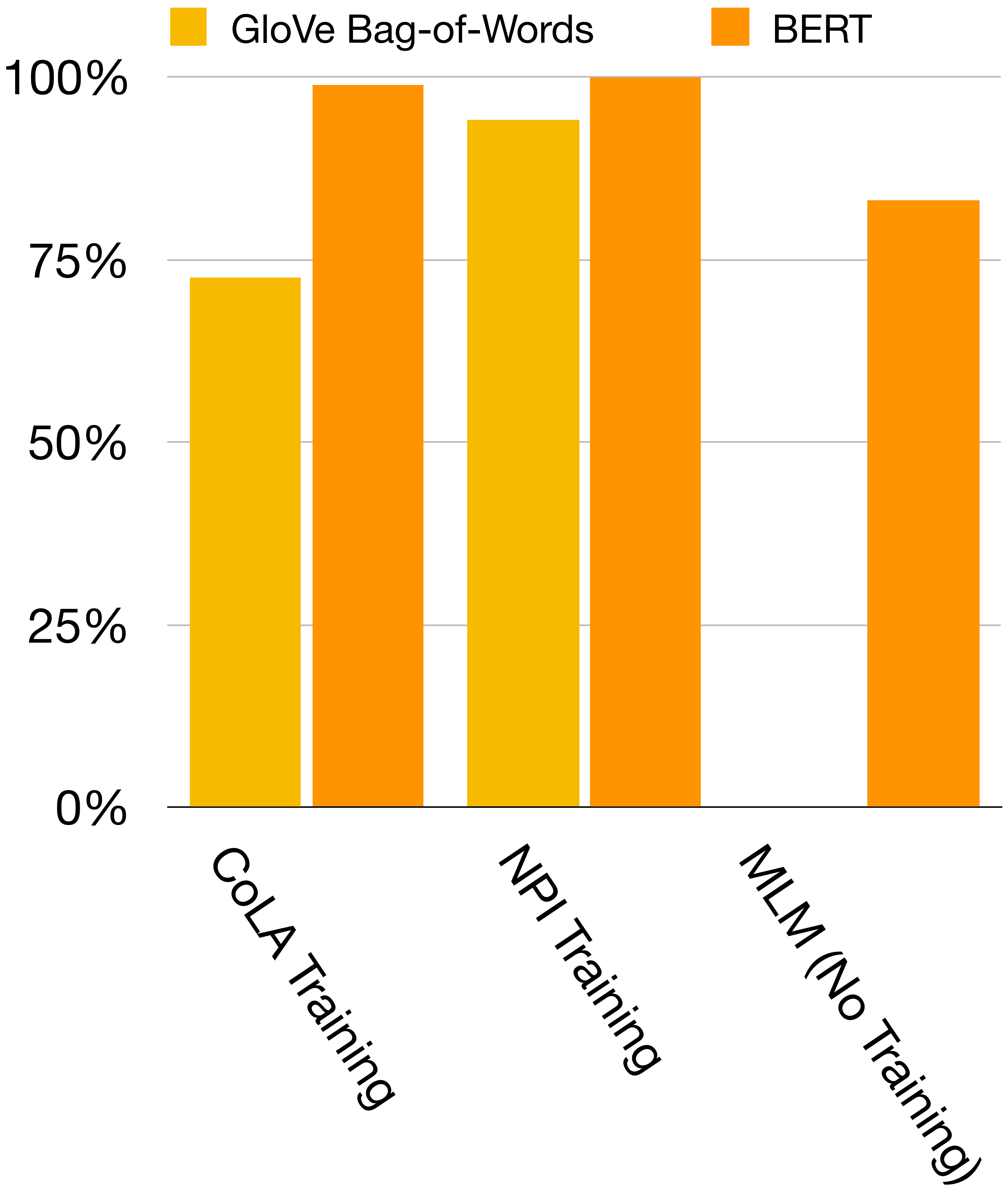
(1) Mary hasn**'t** eaten *any* cookies.

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

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

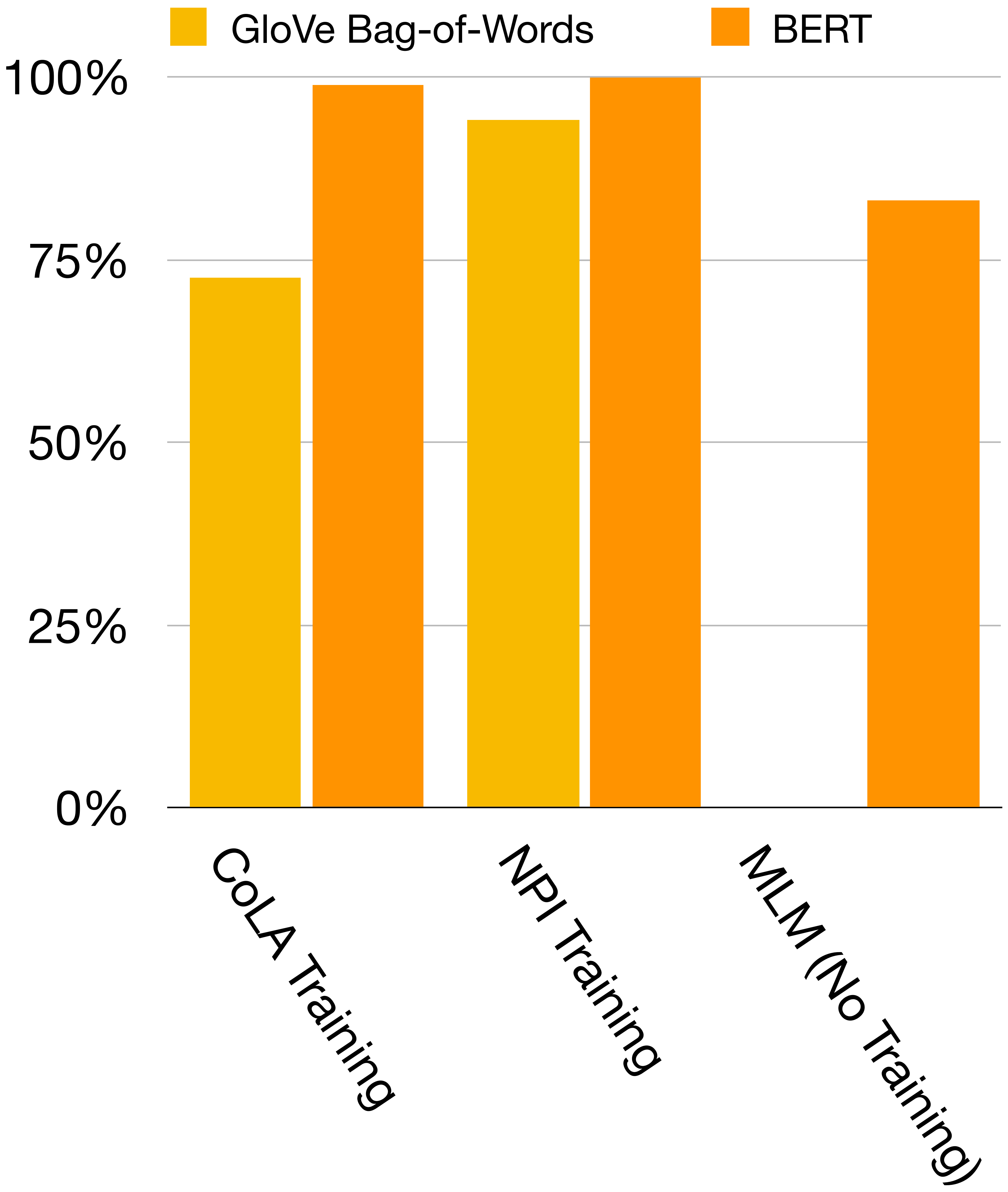
BERT does better than chance (50%), but not especially well.

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

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



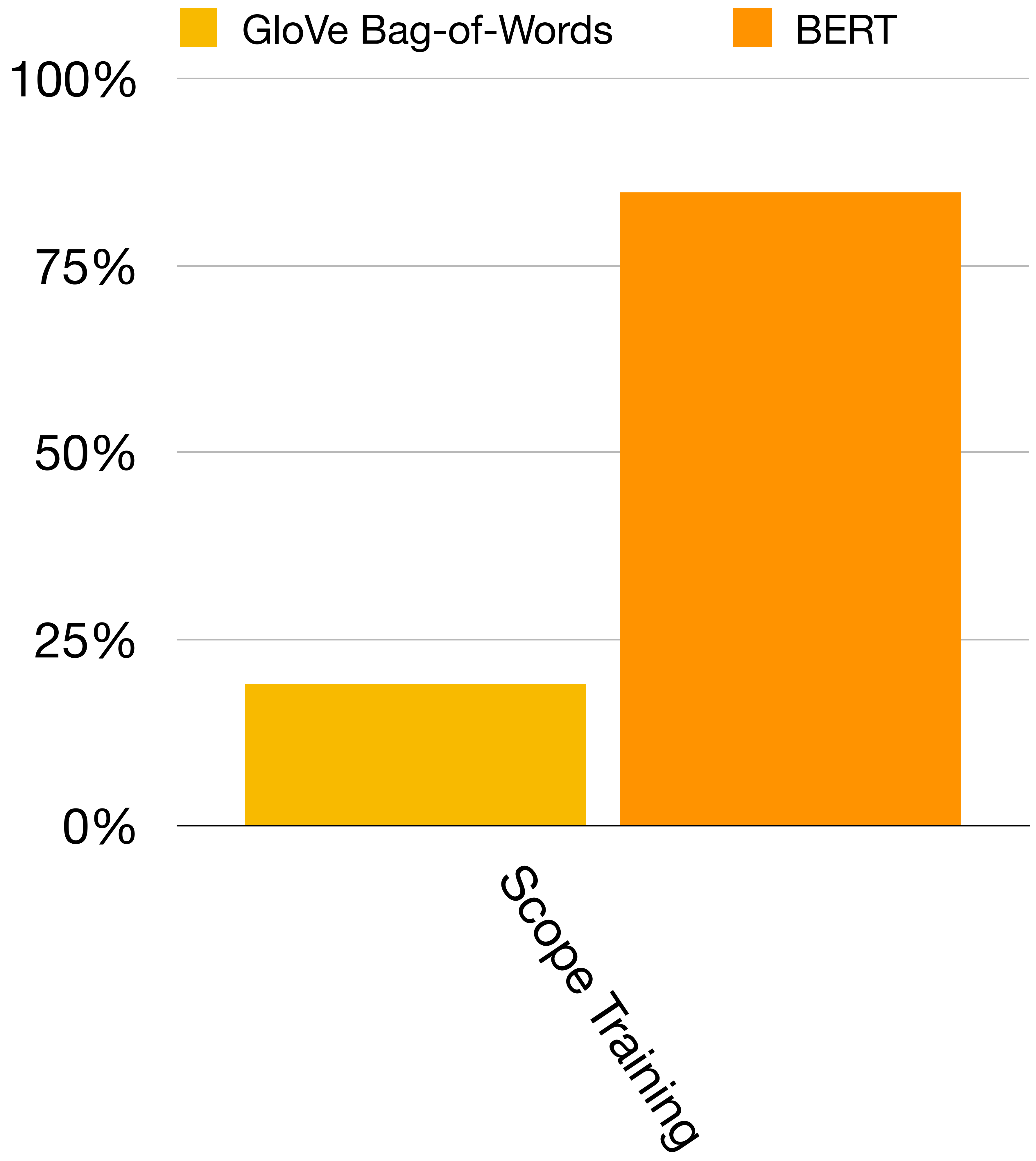
What if we use probing classifiers?

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?

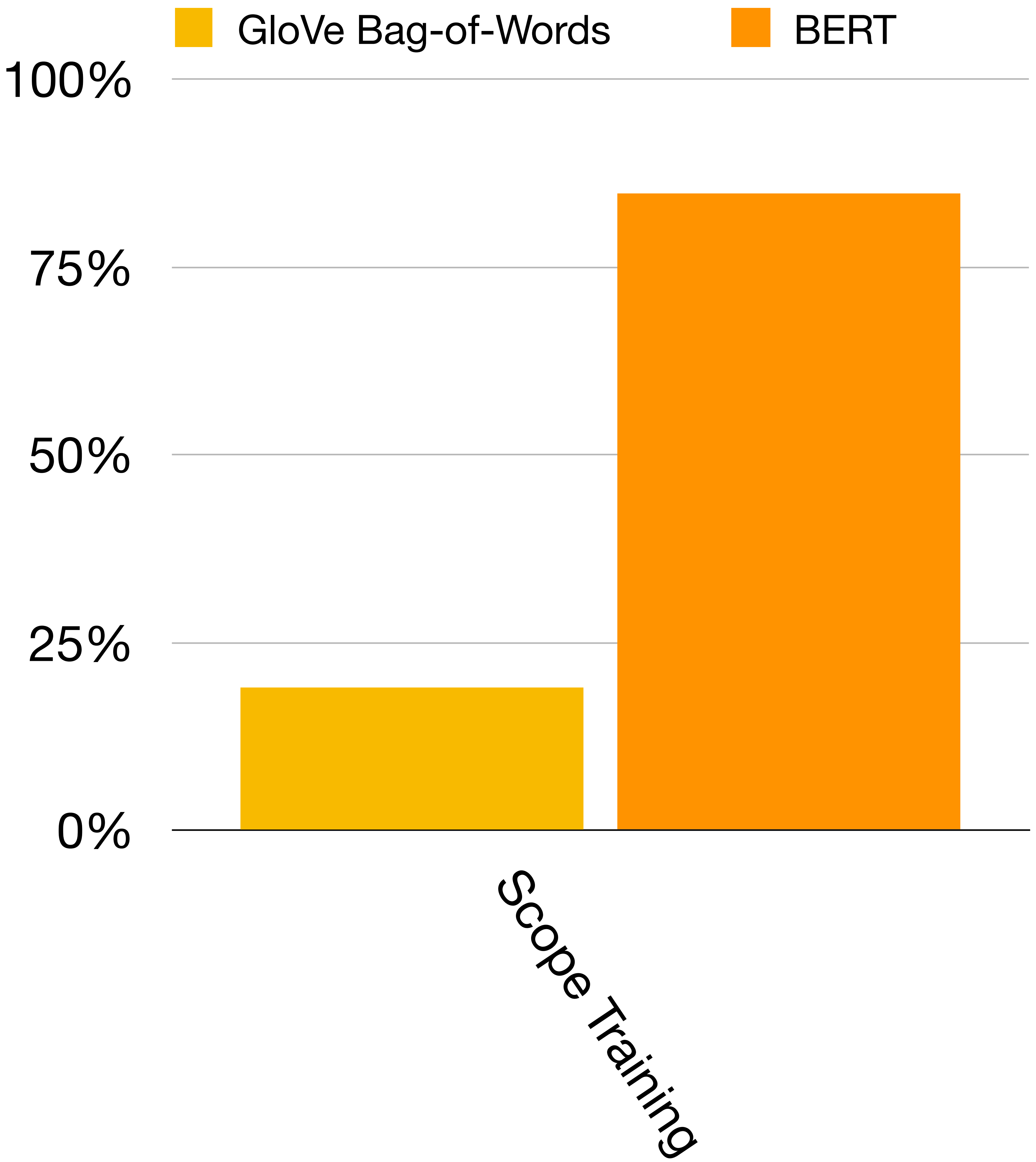
1 ery.]
0 ry.]
1 y.]
0 y.]
1
0 *Those boys *ever* … [the doctors went to an art gallery.]
1 Those boys say … [the doctors *often* went to an art gallery.]
0 Those boys *of*… say **that** [the doctors went to an art gallery.]

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 pro

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

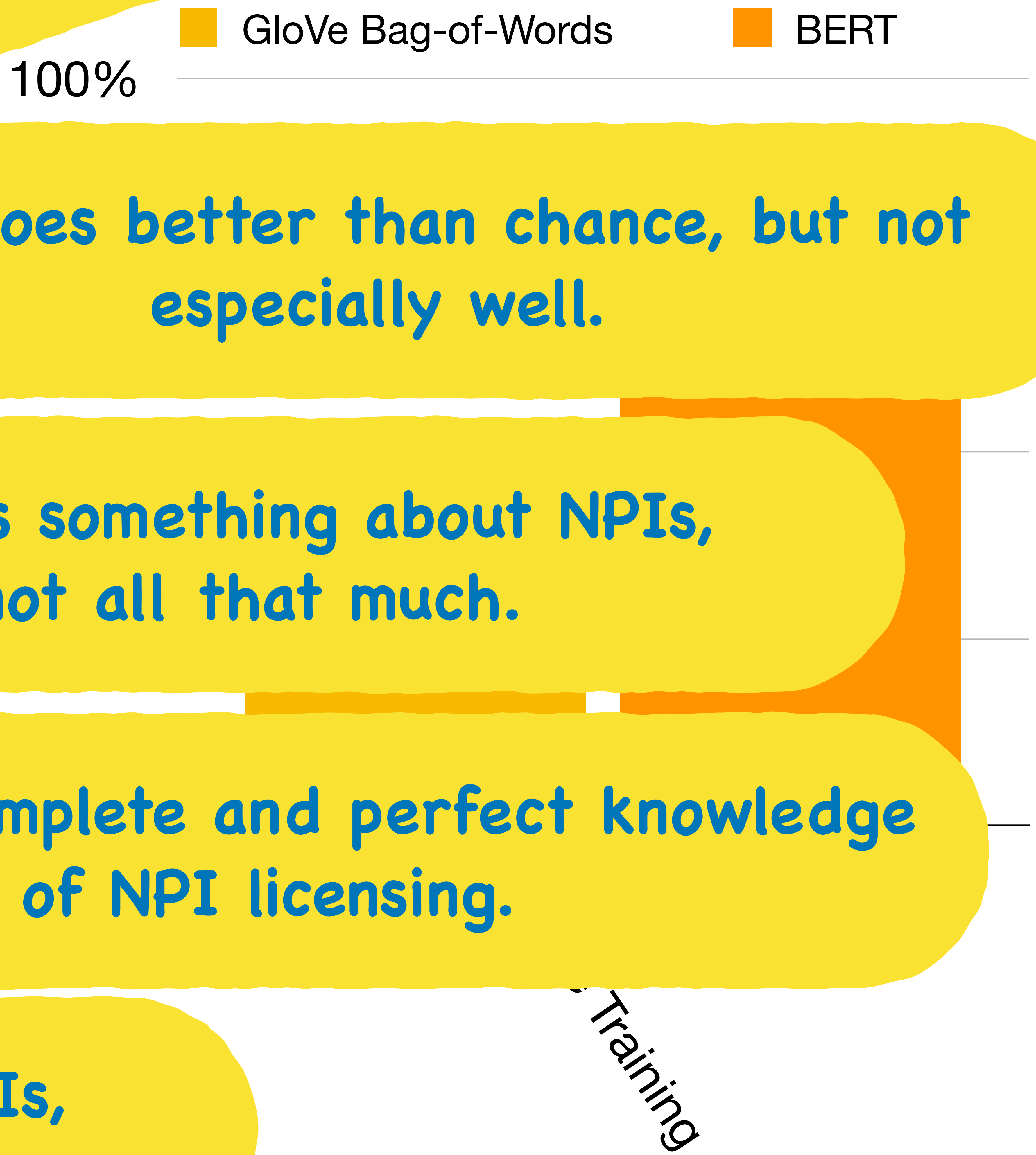
1 ...ery.]
0 ...ry.]
1
0
1
0 *Those boys *ever* ... [the doctors went ... art gallery
1 Those boys say ... [the doctors *ofte*... to an art gallery.]
0 Those boys *of*... say **that** [the ...ors went to an art gallery]

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

Train:
Scop... ...diction task, training o... ...(hold-one ...ithout fine-tu... en...nments)

Test:
Scope predic... ...sk

Met...



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.

Back to evaluation...

Evaluation: What’s Next?

There are plenty of big open problems in NLU, but doesn’t seem possible to build another GLUE-style benchmark again soon.

- Is our ability to build models improving faster than our ability to build hard evaluation sets?

Evaluation: What’s Next?

Give up and work on something else?

- I guess?
- or...

Evaluation: What’s Next?

Use *adversarial filtering* to semi-automatically create datasets that are hard for SotA models?

- Good source of data for training...
- Okay source of data for local hill-climbing evaluation...

Evaluation: What's Next?

Use *adversarial filtering* to semi-automatically create datasets that are hard for SotA models?

- Good source of data for training...
- Okay source of data for local hill-climbing evaluation...
- ...but using these datasets as benchmarks risks encouraging models that are *different but not better.*
- Mitigated by fast iteration times, but logistics get complicated.

Evaluation: What's Next?

Build *growing* benchmarks like Build-it-Break-it or ORB, where experts can add test data to target weaknesses.

- Similar risks, though to a lesser degree.
- Some risk that we lose sight of the task we're trying to solve.

Evaluation: What’s Next?

Restrict the task training sets, or focus on *zero-shot* or *few-shot* adaptation to new tasks.

- Likely to encourage good representations…
- …but may not reflect the setting that we’re interested in.

Evaluation: What’s Next?

Build big, high-quality datasets?

- Aim for *hard* examples with human performance >99%.
- Doable! But slow, expensive, risky work.

One More Open Question

Is it possible to build benchmarks *for bias* that are robust and realistic enough that it’s worthwhile to hill-climb on them?

Evaluation: What’s Next?

Thanks!

SCHMIDT **FUTURES**

SAMSUNG

NSF

NVIDIA

NYU

ML²

Machine Learning
for Language

Sam Bowman

@sleepinyourhat