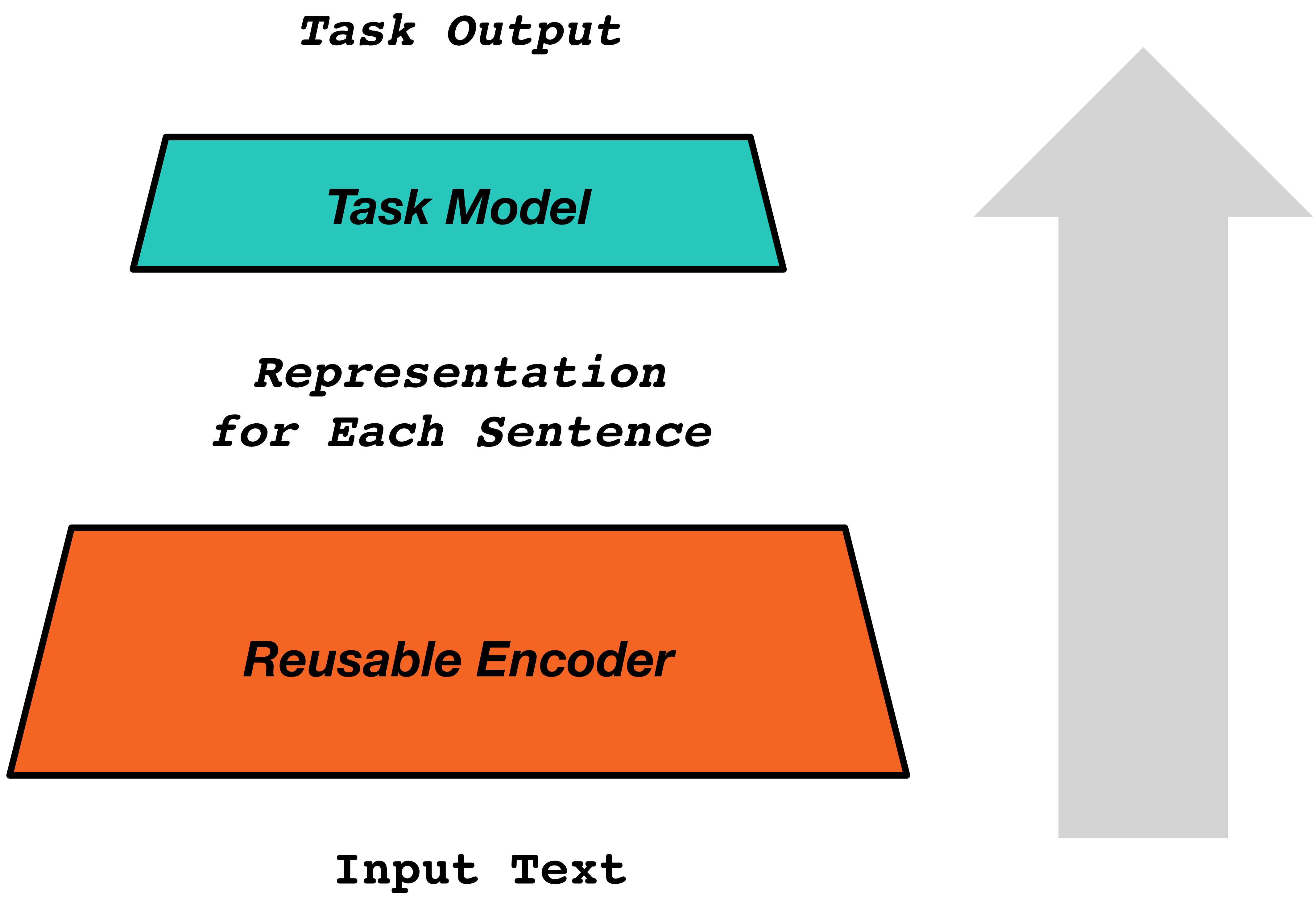
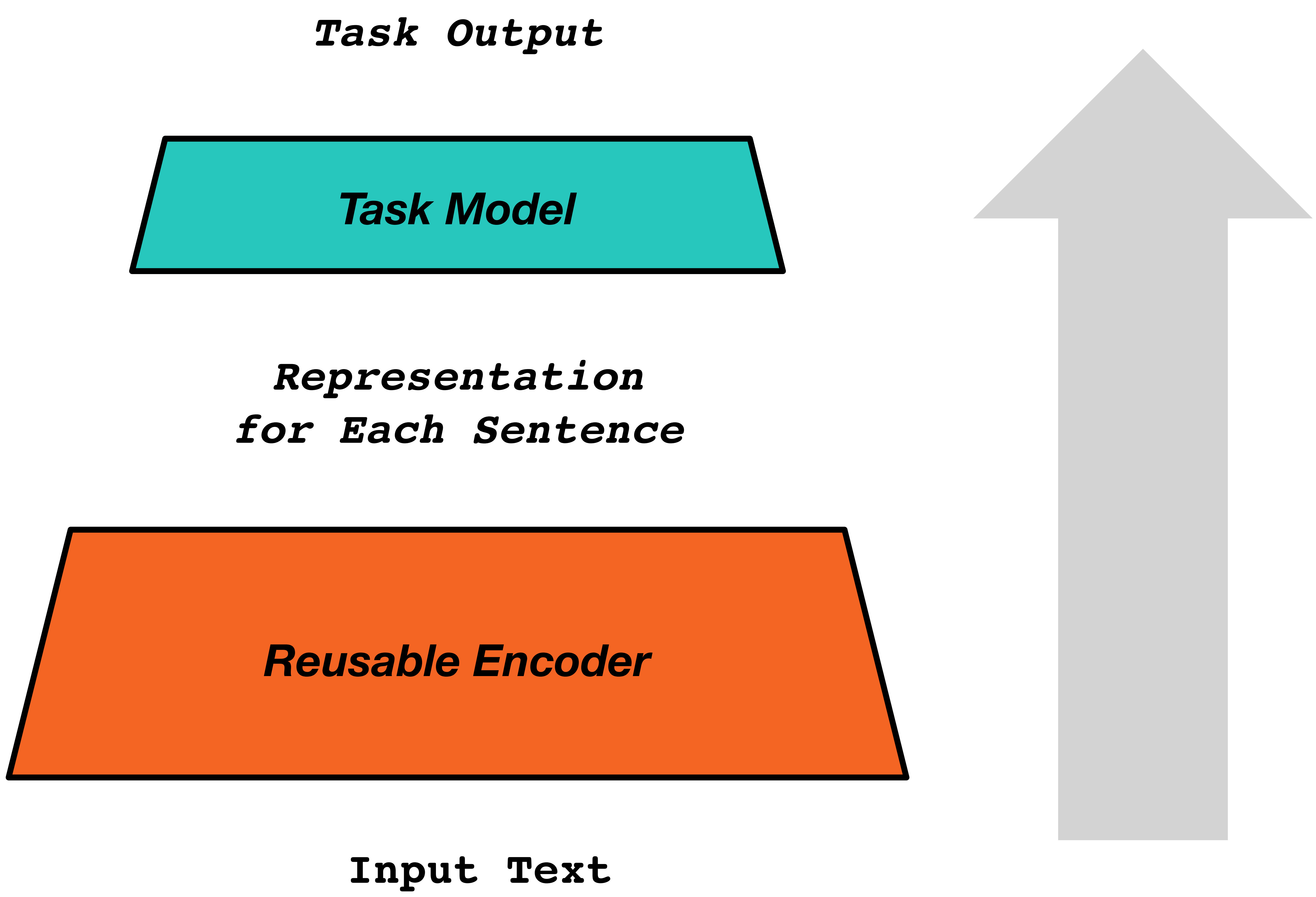
Task-Independent Sentence Understanding Models

Sam Bowman
@sleepinyourhat
NYU
Machine Learning
for Language



A general-purpose sentence encoder



The Goal

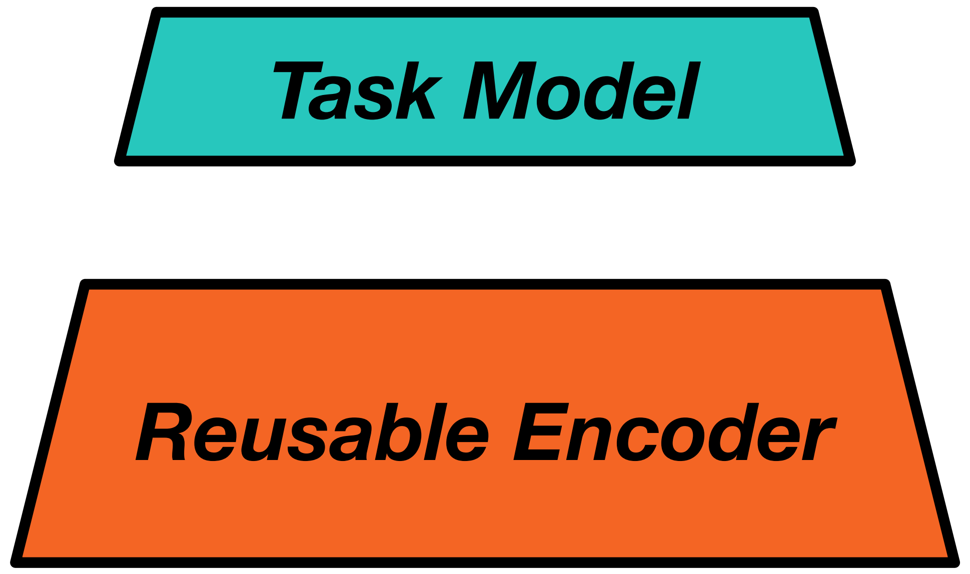
To develop a **general-purpose neural network sentence encoder** which makes it possible to solve any new **language understanding task** using only enough training data to **define the possible outputs**.

A general-purpose encoder

- Roughly, we might expect effective encoder representations to capture:
  - Word contents and word order.
  - (Rough) grammatical structure.
  - Cues to connotation and social meaning.
  - Unambiguous propositional information (of the kind expressed in a semantic parse).

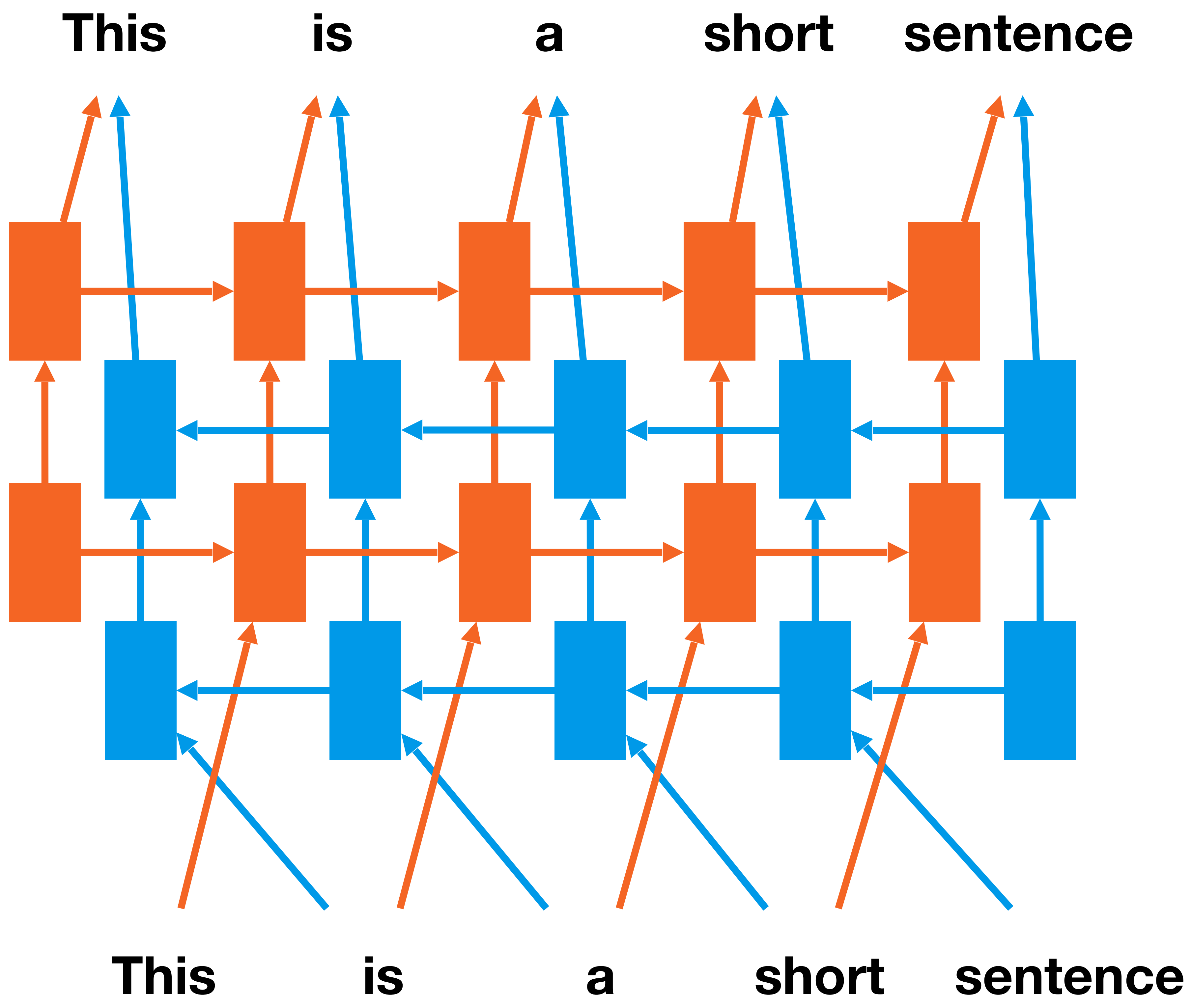
$$\forall x[\text{patient}'(x) \rightarrow \exists y[\text{doctor}'(y) \wedge \text{treat}'(y, x)]]$$

- *These are still neural networks, so all of this will be implicit.*



Case Study: ELMo

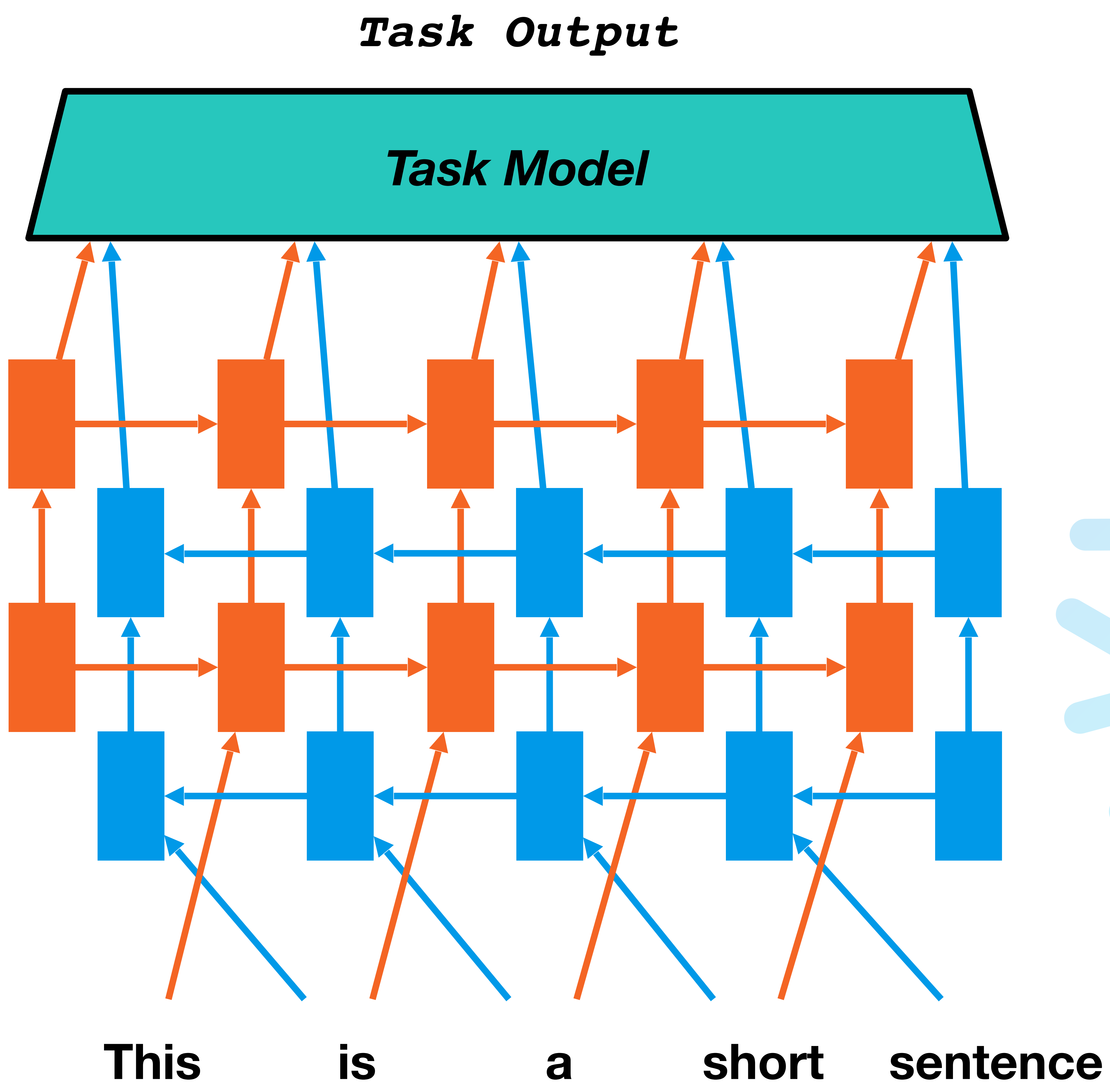
Train large forward ***and backward*** deep LSTM language models.



Peters et al. '18

Case Study: ELMo

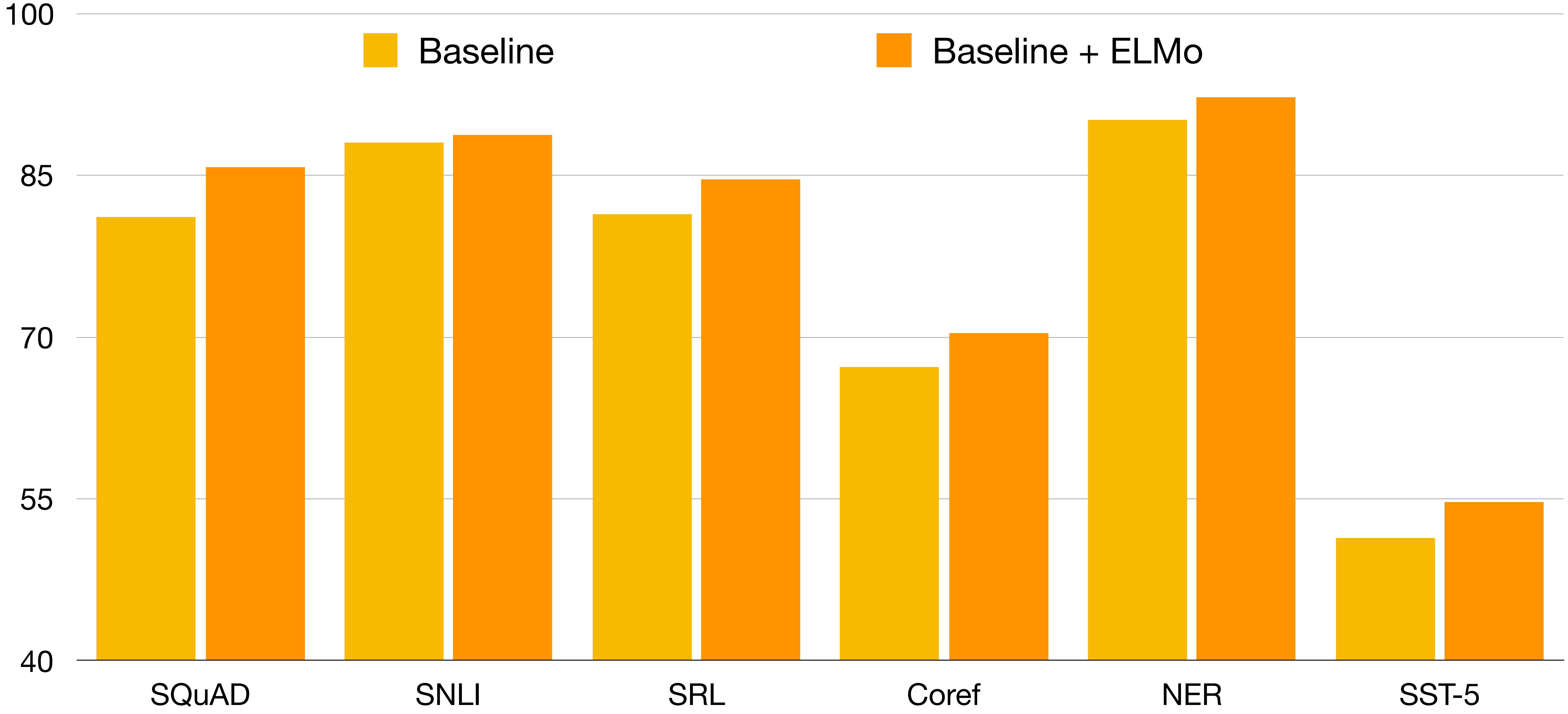
Train large (~100m-param) forward ***and backward*** deep LSTM language models.



Peters et al. '18

Case Study: ELMo

Best paper at NAACL 2018!



Peters et al. '18

The Rest of the Talk

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

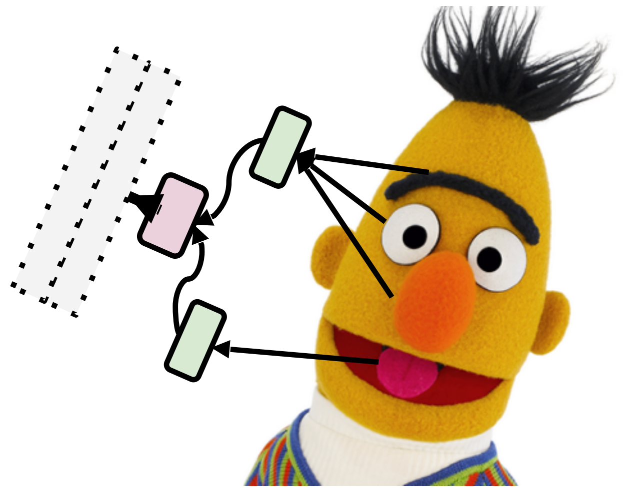
  

  - ...and successes with unsupervised pretraining and fine-tuning on GLUE
    Radford et al. '18 (OpenAI GPT), Devlin et al. '18 (BERT)

- The updated SuperGLUE benchmark
  Wang et al. '19a

- Easy transfer learning with STILTs
  Phang et al. '19

- A few more things we've learned about these models
  Wang et al. '19b, Tenney et al. '19



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 '19

GLUE, in short

- 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.
    - Easy to use!
- Simple leaderboard API: Upload predictions for a test set (Kaggle-style)
    - Usable with any kind of method/model!

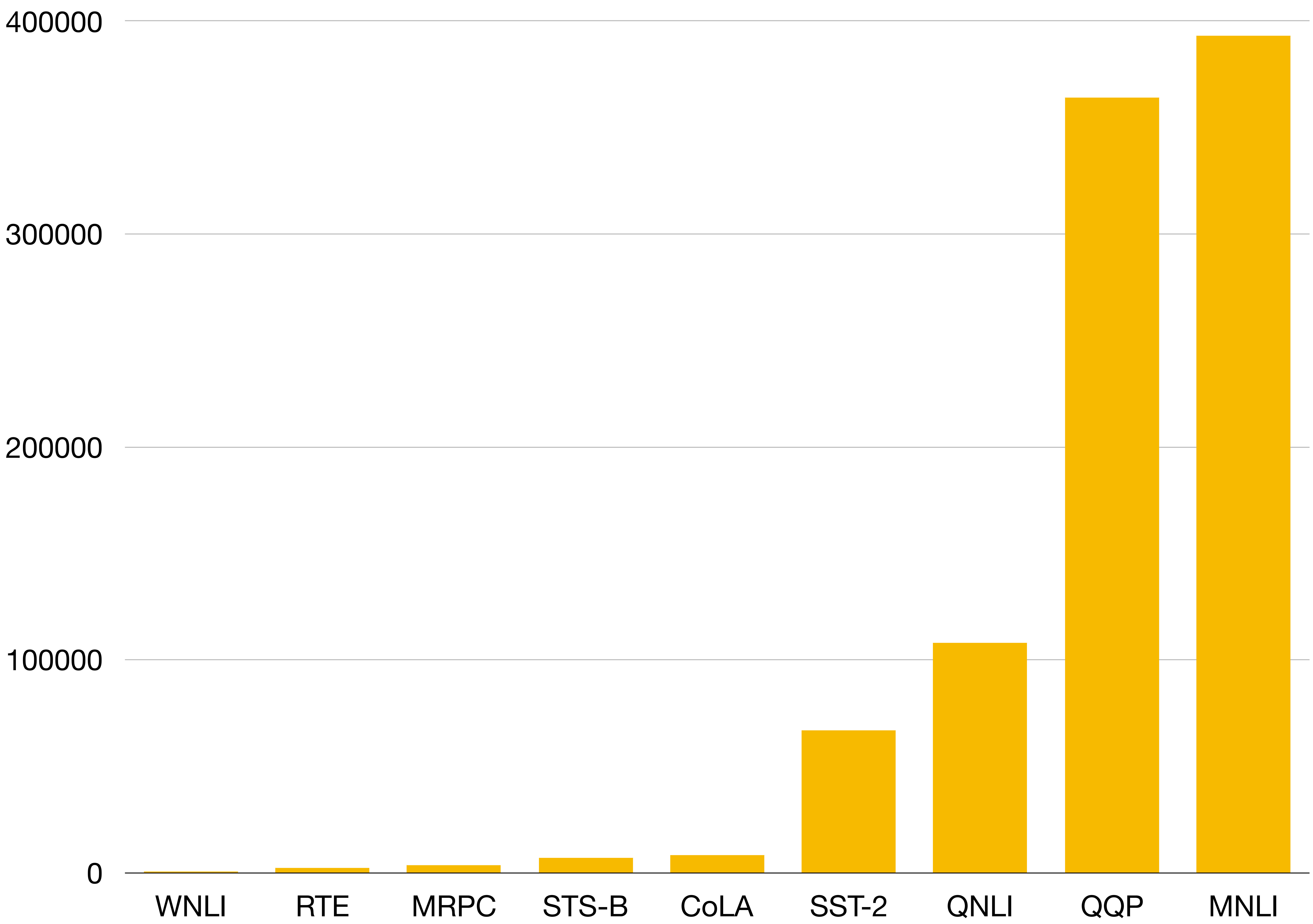
Wang, Singh, Michael, Hill, Levy & Bowman '19

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 '19

Corpus	|Train|
CoLA	8.5k
SST-2	67k
MRPC	3.7k
STS-B	7k
QQP	364k
MNLI	393k
QNLI	108k
RTE	2.5k
WNLI	634



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 '19

The Corpus of Linguistic Acceptability

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 aph			

Wang, Singh, Michael, Hill, Levy & Bowman '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.**

P: *Dana Reeve, the widow of the actor Christopher Reeve, has died of lung cancer at age 44, according to the Christopher Reeve Foundation.*
H: *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 '19

Multi-Genre Natural Language Inference

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 '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.**

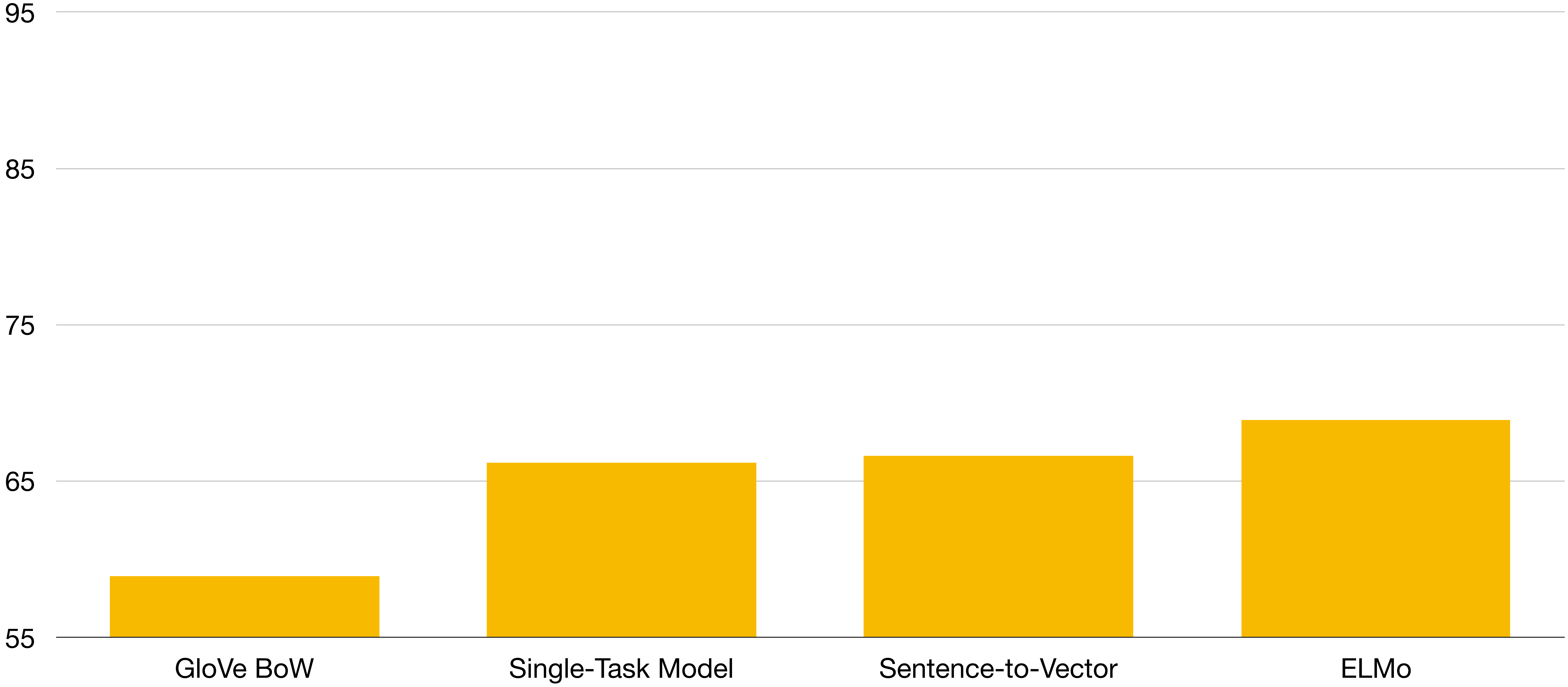
P: *Jane gave Joan candy because she was hungry.*
H: *Joan was hungry.*
entailment

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 '19

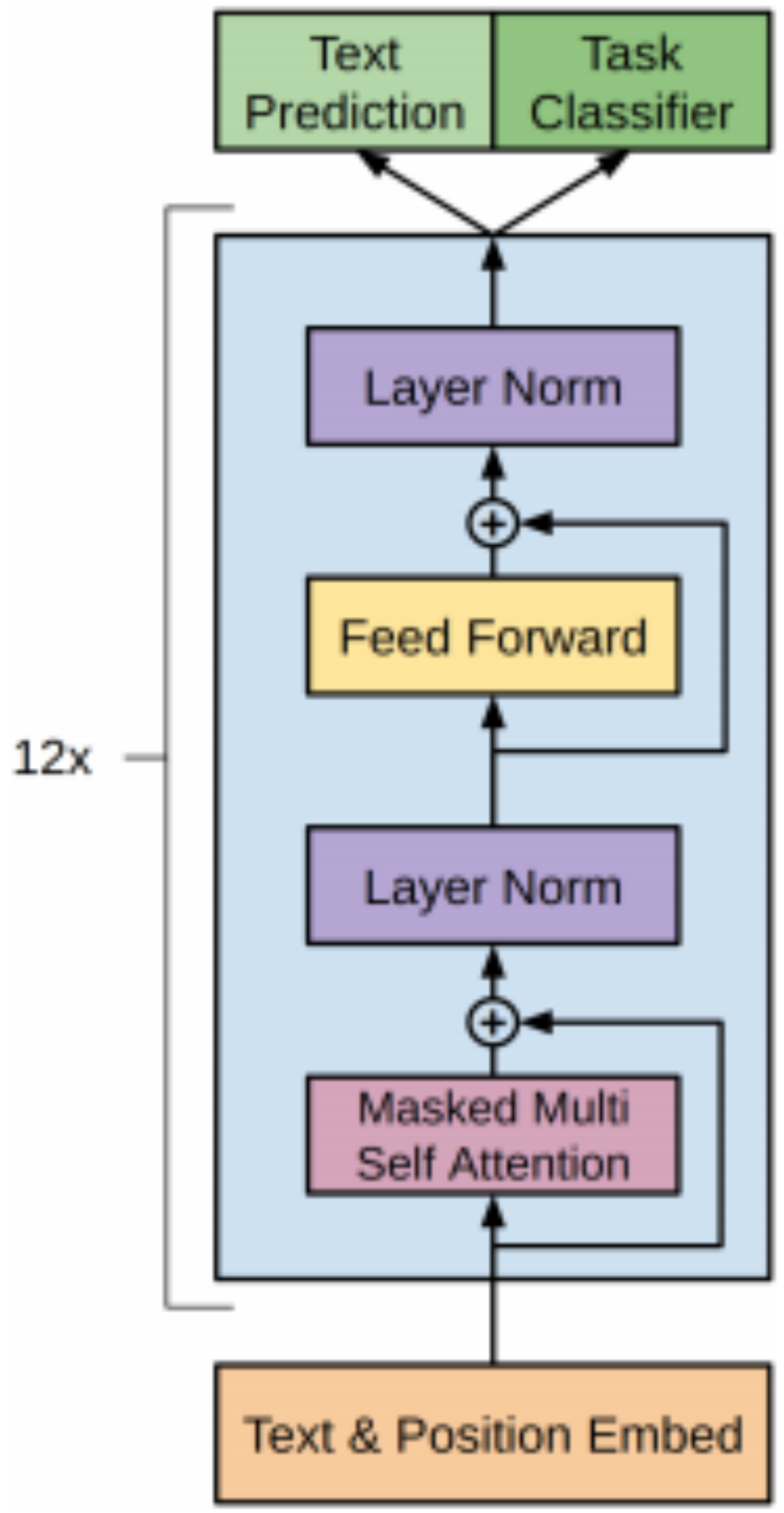
GLUE: What methods work?

Overall GLUE Score



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

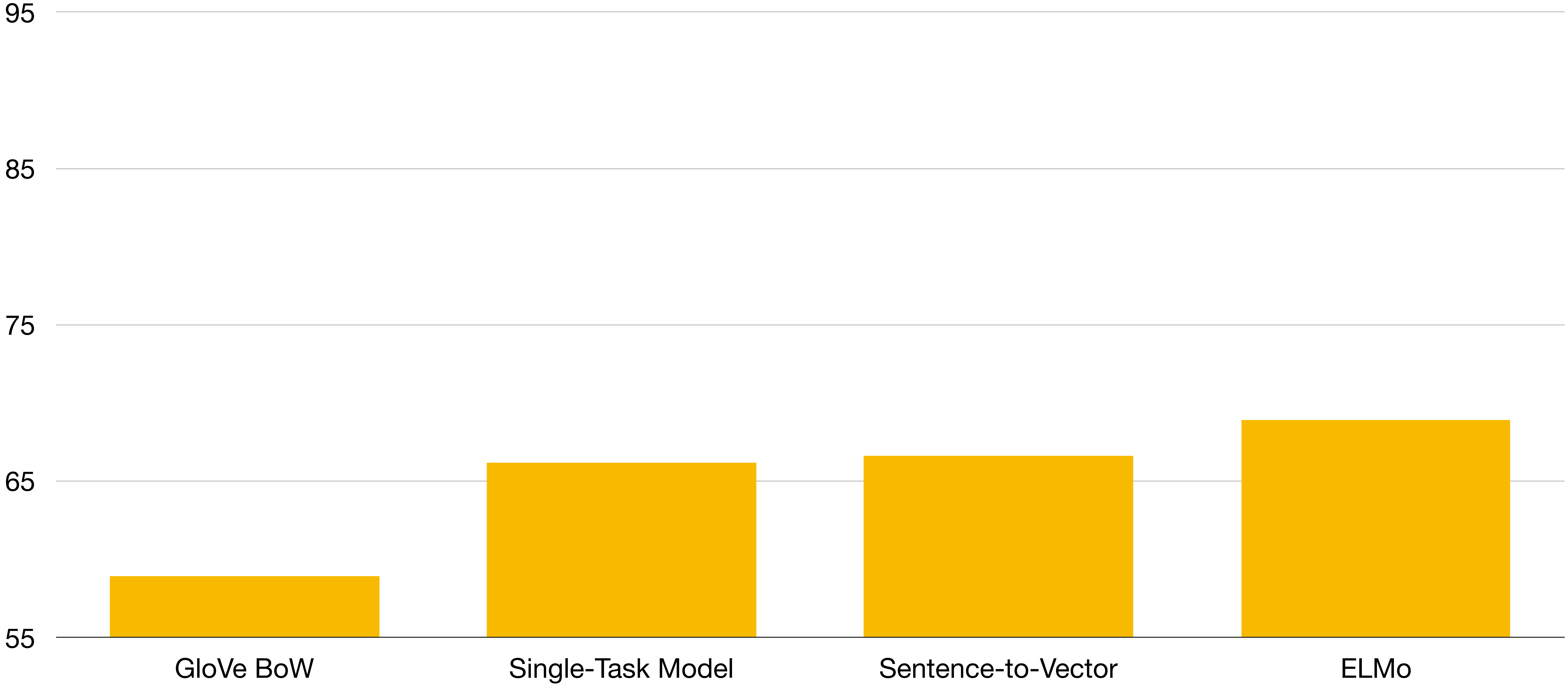
OpenAI's GPT Language Model



- Same basic idea as ELMo, but many changes (and many open questions!), 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.

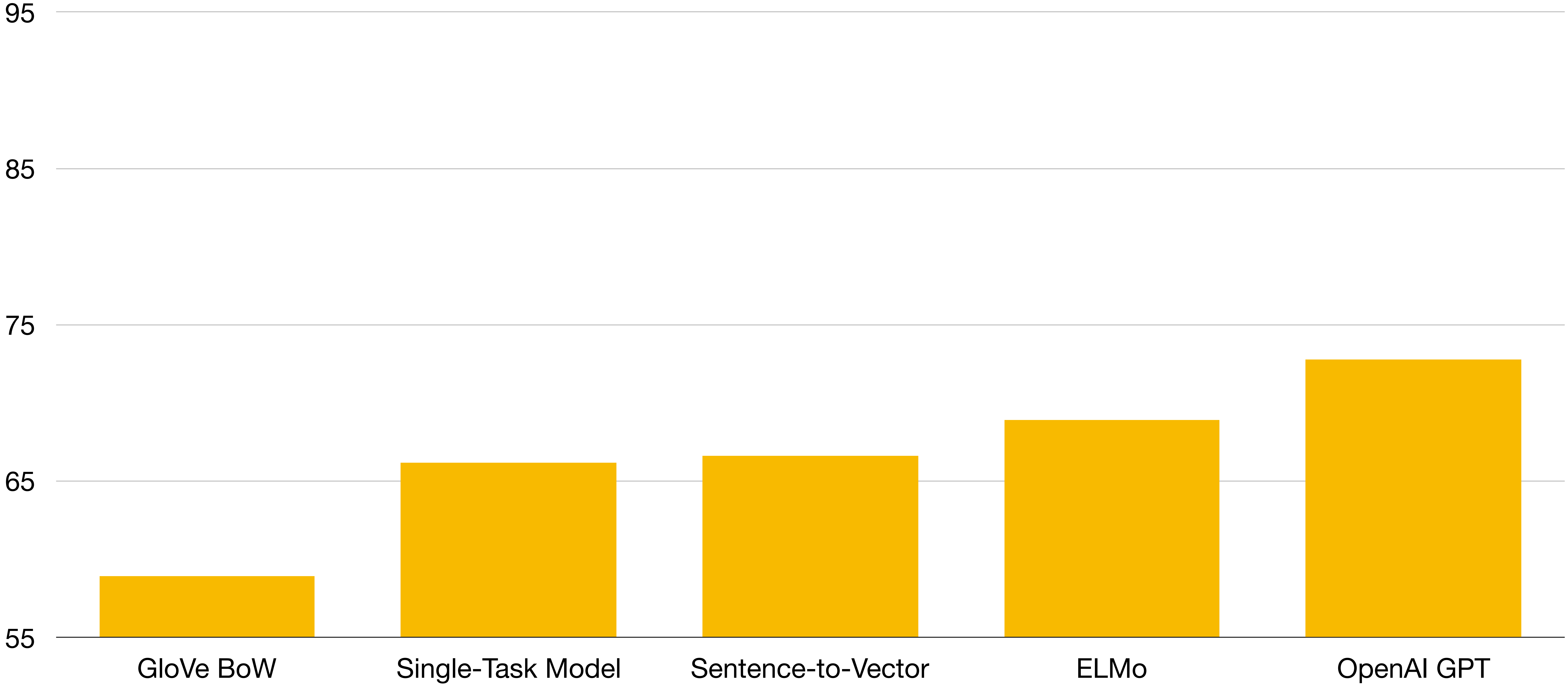
Radford et al. '18

GLUE Score



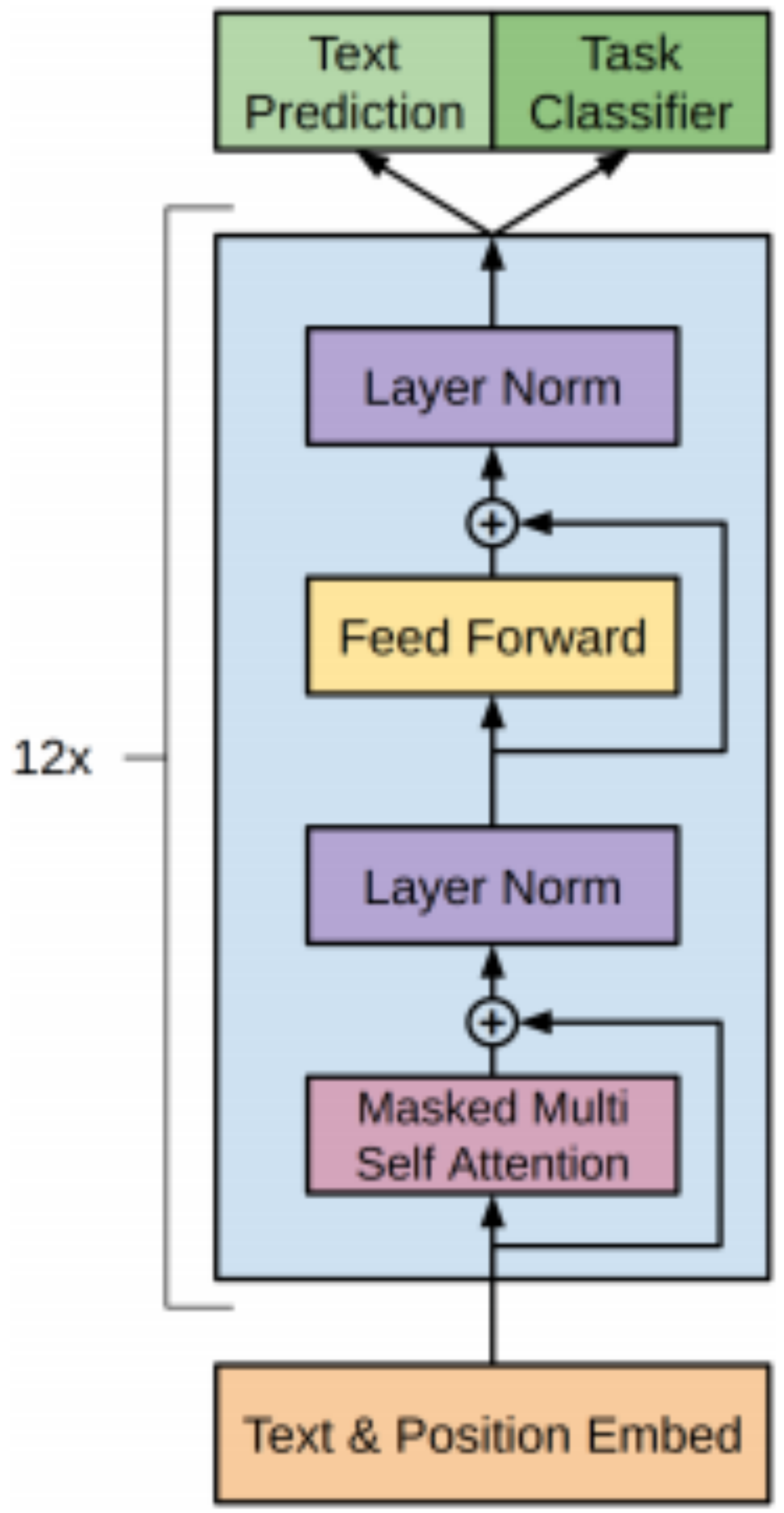
Radford et al. '18

GLUE Score



Radford et al. '18

OpenAI's Transformer Language Model



- Update this spring:
  - Announcement of 15x larger GPT-2 language model.
  - Impressive text generation results, but no transfer learning evaluations yet.
    - Systems issues yet to be solved internally, security concerns prevent sharing

Radford et al. '18

Google's BERT

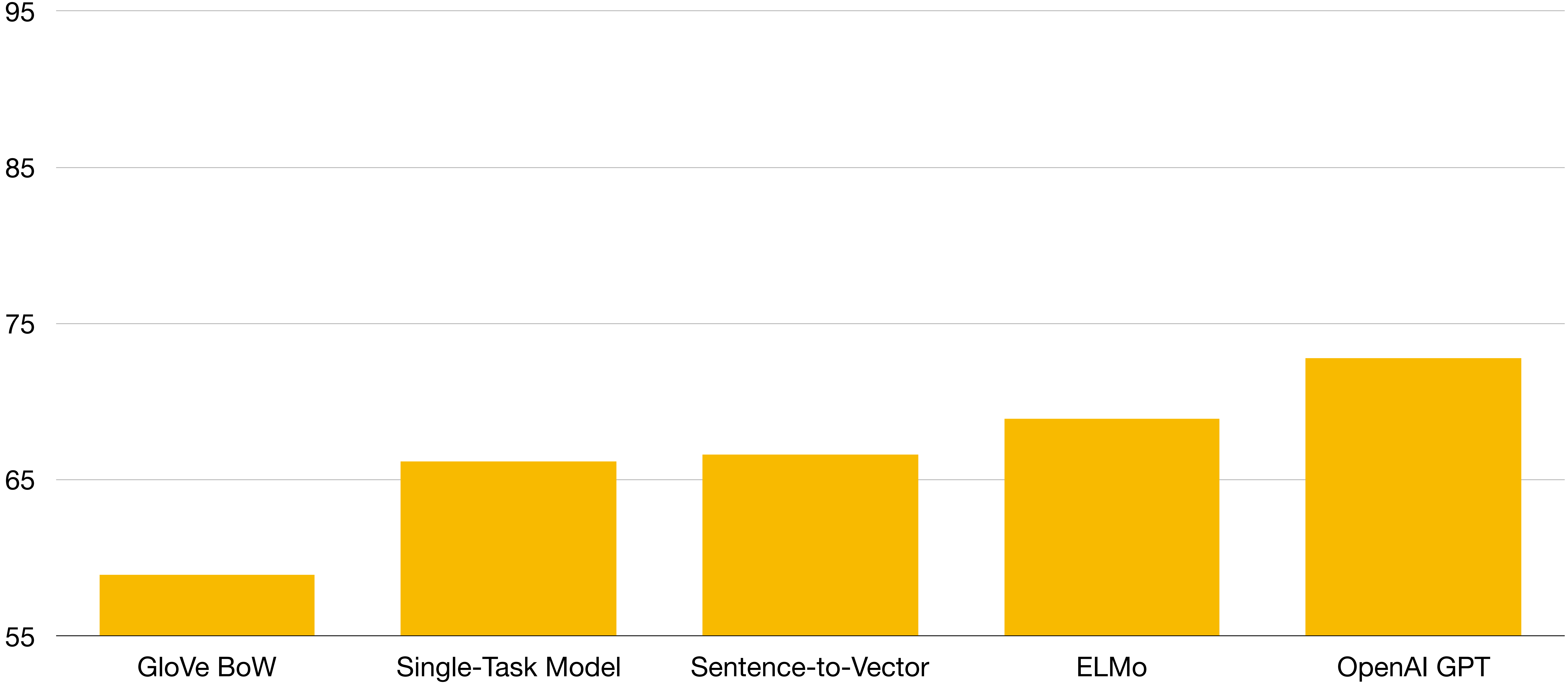
Devlin et al. '18
see Baevski et al. '19 for similar concurrent work

The BERT Model

- Same basic idea as OpenAI 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.
  - *Very* big (>300M params).

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

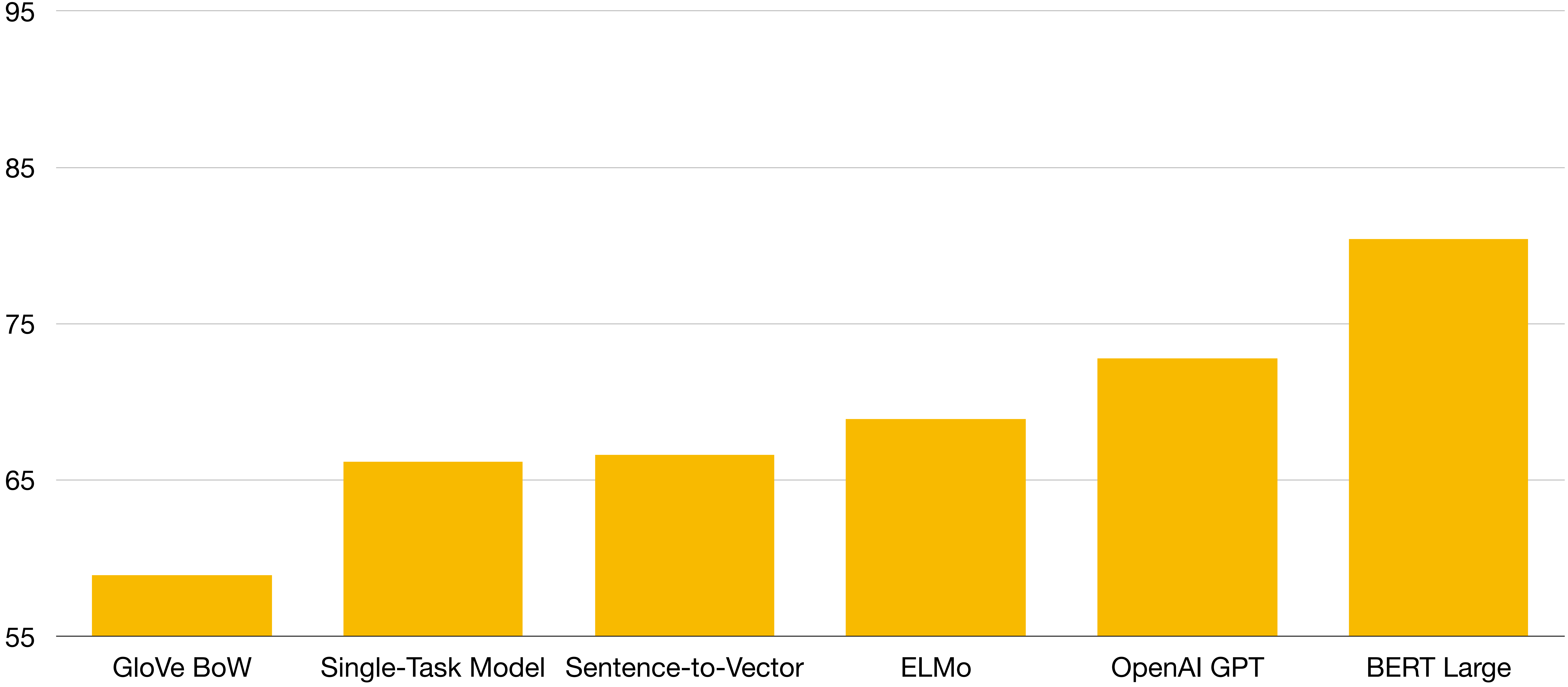
GLUE Score



Devlin et al. '18

see Baevski et al. '19 for similar concurrent work

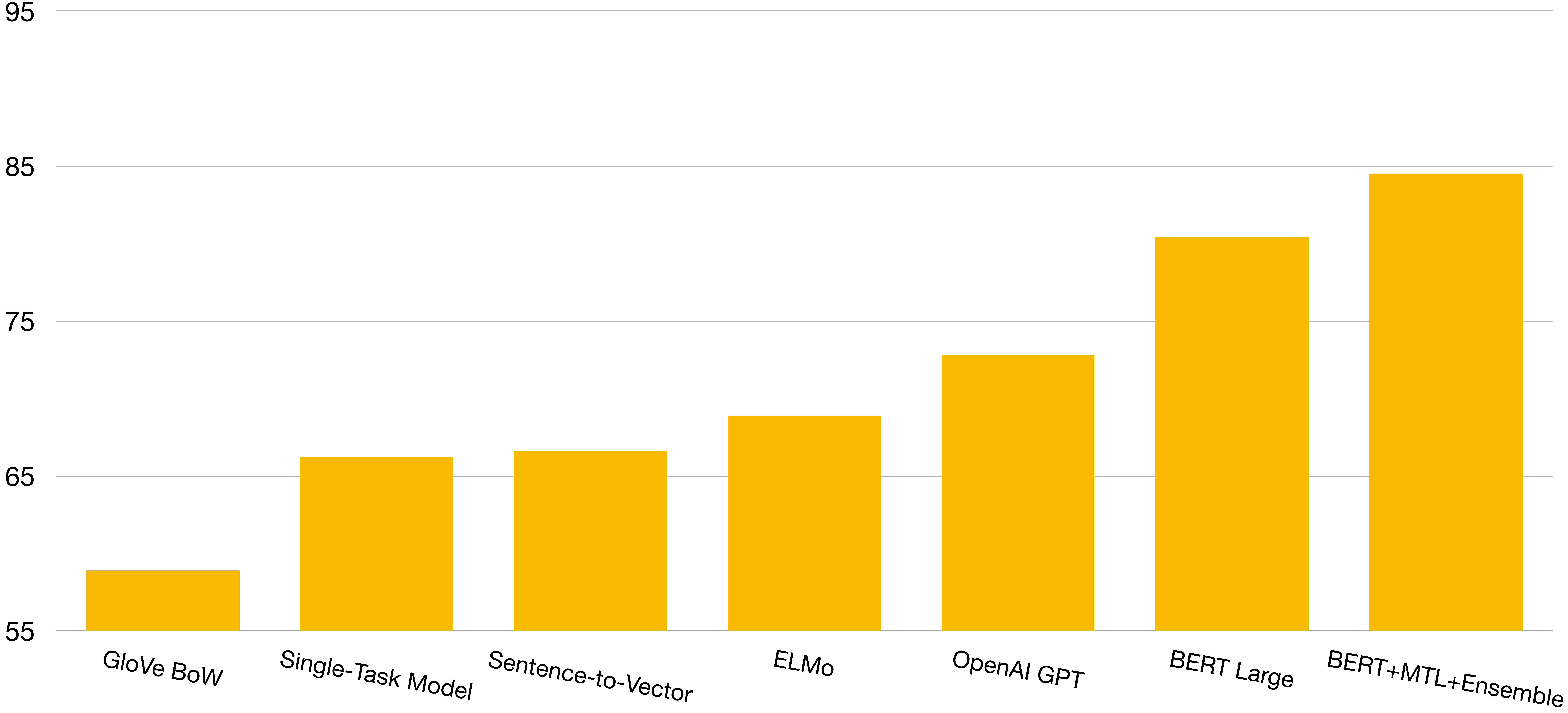
GLUE Score



Devlin et al. '18

see Baevski et al. '19 for similar concurrent work

GLUE Score



Liu et al. '19

Human Baseline

- **How much headroom does GLUE have left?**
- To compute a conservative estimate for each task:
    - Show crowdworkers a brief description of each task, plus twenty labeled *development set* examples in an interactive training mode.
    - Collect five crowdworker labels per example for 500 *test set* examples.
    - Take a majority vote and compare the result with the gold labels.

Nangia & Bowman '19

Human Baseline

	Avg	Single Sentence		Sentence Similarity			Natural Language Inference			
	Avg	**CoLA**	**SST-2**	**MRPC**	**STS-B**	**QQP**	**MNLI**	**QNLI**	**RTE**	**WNLI**
Training Size	-	*8.5k*	*67k*	*3.7k*	*7k*	*364k*	*393k*	*108k*	*2.5k*	*634*
Human	**87.1**	**66.4**	**97.8**	80.8/86.3	**92.7/92.6**	80.4/59.5	**92.0/92.8**	91.2	**93.6**	**95.9**
BERT	80.5	60.5	94.9	85.4/89.3	87.6/86.5	**89.3/72.1**	86.7/85.9	**92.7**	70.1	65.1
BigBird	83.9	65.4	95.6	**88.2/91.1**	89.5/89.0	**89.6/72.7**	87.9/87.4	95.8*	85.1	65.1
Δ_{bert} (-)	6.6	5.9	2.9	-4.6/-3.0	5.1/6.1	-8.9/-12.6	5.3/6.9	-1.5	23.5	30.8
Δ_{bird} (-)	3.2	1.0	2.2	-7.4/-4.8	3.2/3.6	-9.2/-13.2	4.1/5.4	-4.6*	8.5	30.8

Nangia & Bowman '19

Human Baseline

	Avg	Single Sentence CoLA	SST-2	Sentence Similarity MRPC	STS-B	QQP	Natural Language Inference MNLI	QNLI	RTE	WNLI
Training Size	-	*8.5k*	*67k*	*3.7k*	*7k*	*364k*	*393k*	*108k*	*2.5k*	*634*
Human	**87.1**	**66.4**	**97.8**	80.8/86.3	**92.7/92.6**	80.4/59.5	**92.0/92.8**	91.2	**93.6**	**95.9**
BERT	80.5	60.5	94.9	85.4/89.3	87.6/86.5	**89.3/72.1**	86.7/85.9	**92.7**	70.1	65.1
BigBird	83.9	65.4	95.6	**88.2/91.1**	89.5/89.0	**89.6/72.7**	87.9/87.4	95.8*	85.1	65.1
Δ_{bert} (-)	6.6	5.9	2.9	-4.6/-3.0	5.1/6.1	-8.9/-12.6	5.3/6.9	-1.5	23.5	30.8
Δ_{bird} (-)	3.2	1.0	2.2	-7.4/-4.8	3.2/3.6	-9.2/-13.2	4.1/5.4	-4.6*	8.5	30.8

Nangia & Bowman '19

Human Baseline

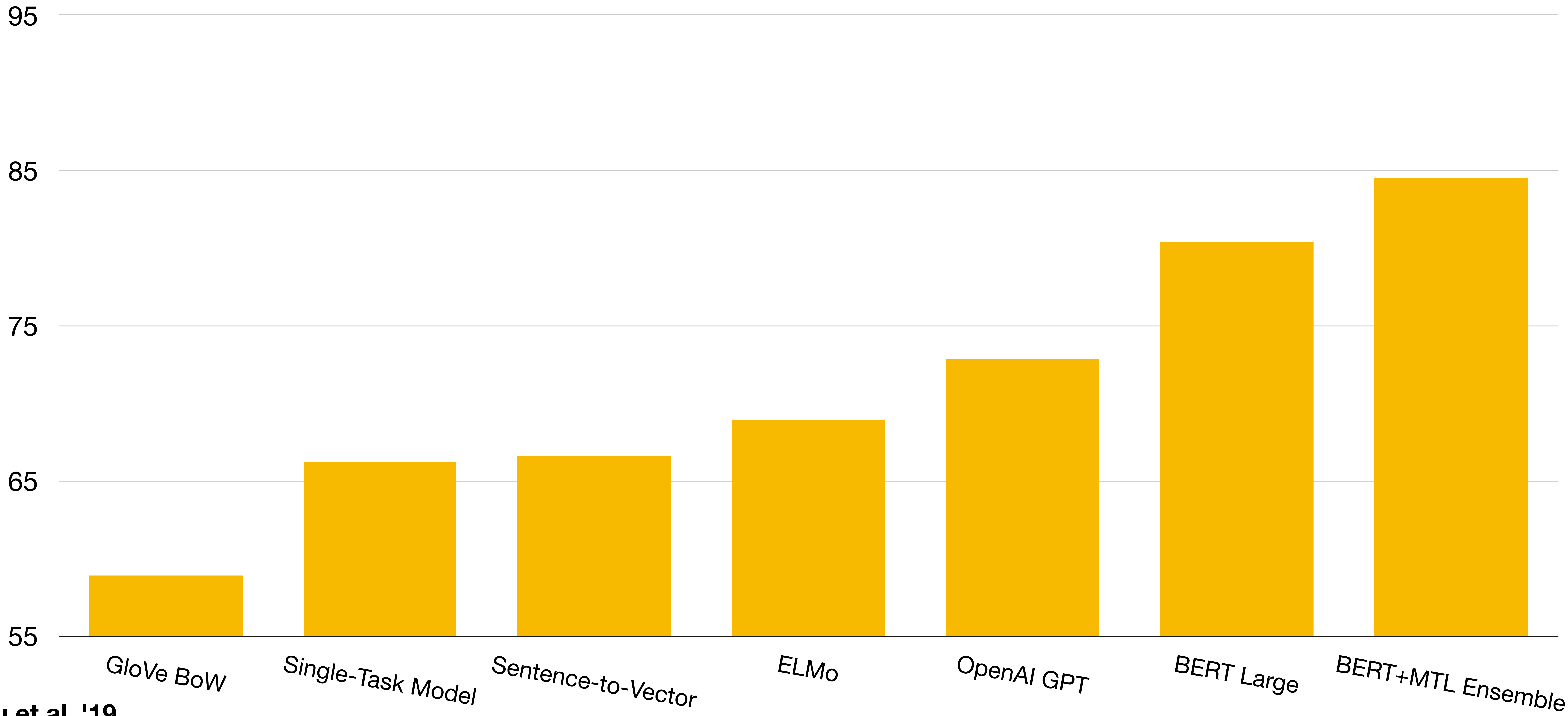
	Avg	Single Sentence CoLA	SST-2	Sentence Similarity MRPC	STS-B	QQP	Natural Language Inference MNLI	QNLI	RTE	WNLI
Training Size	-	*8.5k*	*67k*	*3.7k*	*7k*	*364k*	*393k*	*108k*	*2.5k*	*634*
Human	**87.1**	**66.4**	**97.8**	80.8/86.3	**92.7/92.6**	80.4/59.5	**92.0/92.8**	91.2	**93.6**	**95.9**
BERT	80.5	60.5	94.9	85.4/89.3	87.6/86.5	**89.3/72.1**	86.7/85.9	**92.7**	70.1	65.1
BigBird	83.9	65.4	95.6	**88.2/91.1**	89.5/89.0	**89.6/72.7**	87.9/87.4	95.8*	85.1	65.1
Δ_{bert} (-)	6.6	5.9	2.9	-4.6/-3.0	5.1/6.1	-8.9/-12.6	5.3/6.9	-1.5	23.5	30.8
Δ_{bird} (	[illegible]	1.0	2.2	-7.4/-4.8	3.2/3.6	[illegible]	[illegible]	[illegible]	[illegible]	30.8

Devlin et al. '18:

...scored by the GLUE evaluation server. The number below each task denotes the number of training examples. The “Average” column is slightly different than the official GLUE score, since we exclude the problematic WNLI set. OpenAI GPT = (L=12, H=768, A=12); BERT$_{BASE}$ = (L=12, H=768,

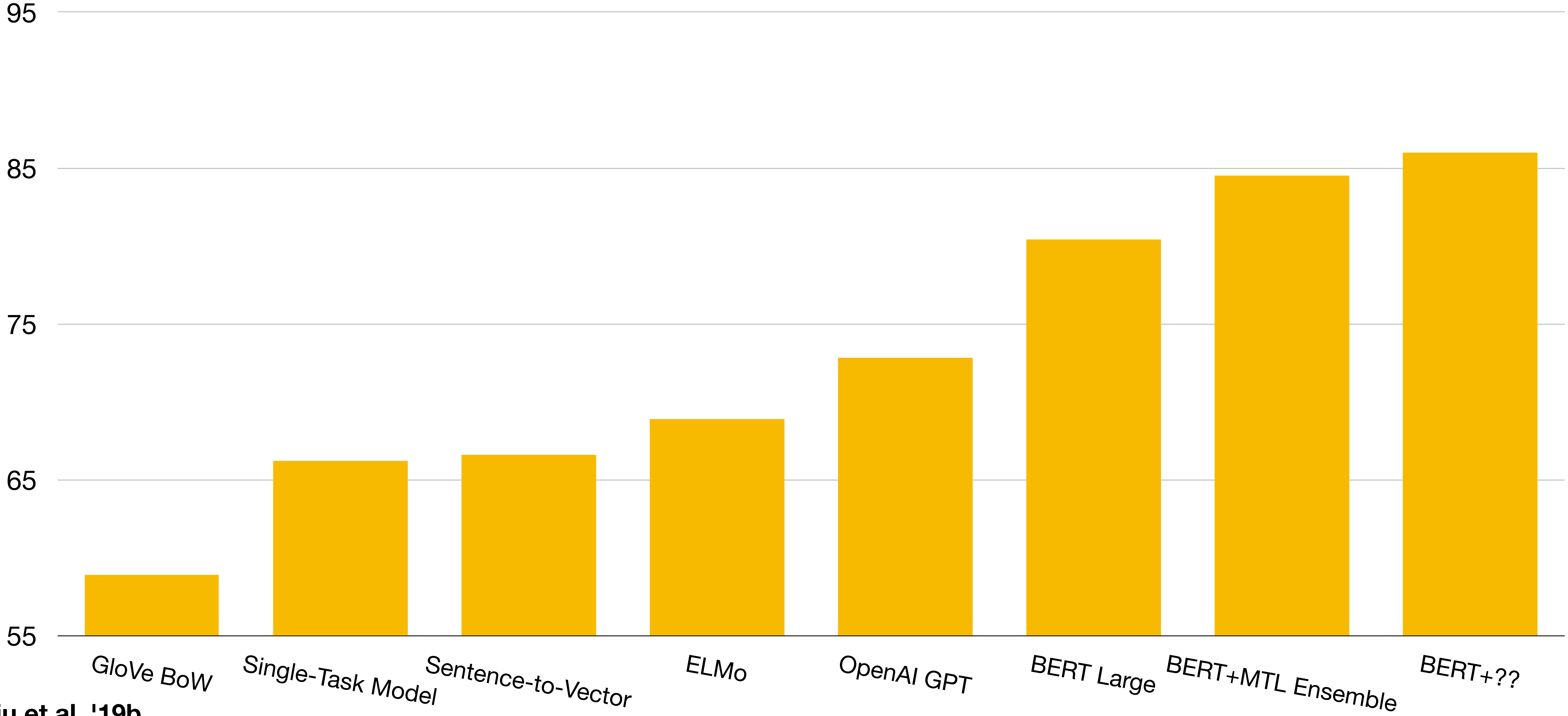
Nangia & Bowman '19

GLUE Score



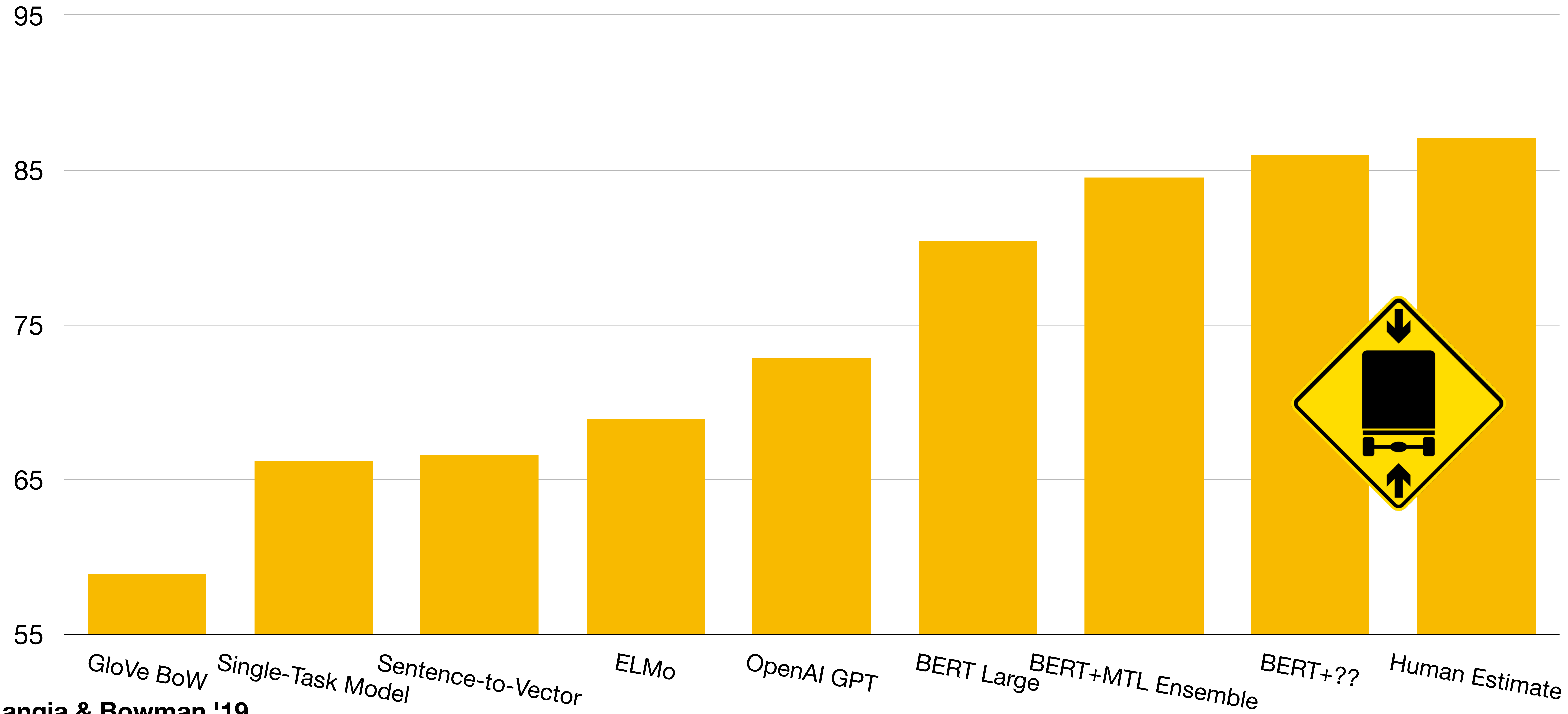
Liu et al. '19

GLUE Score



Liu et al. '19b

GLUE Score



Nangia & Bowman '19

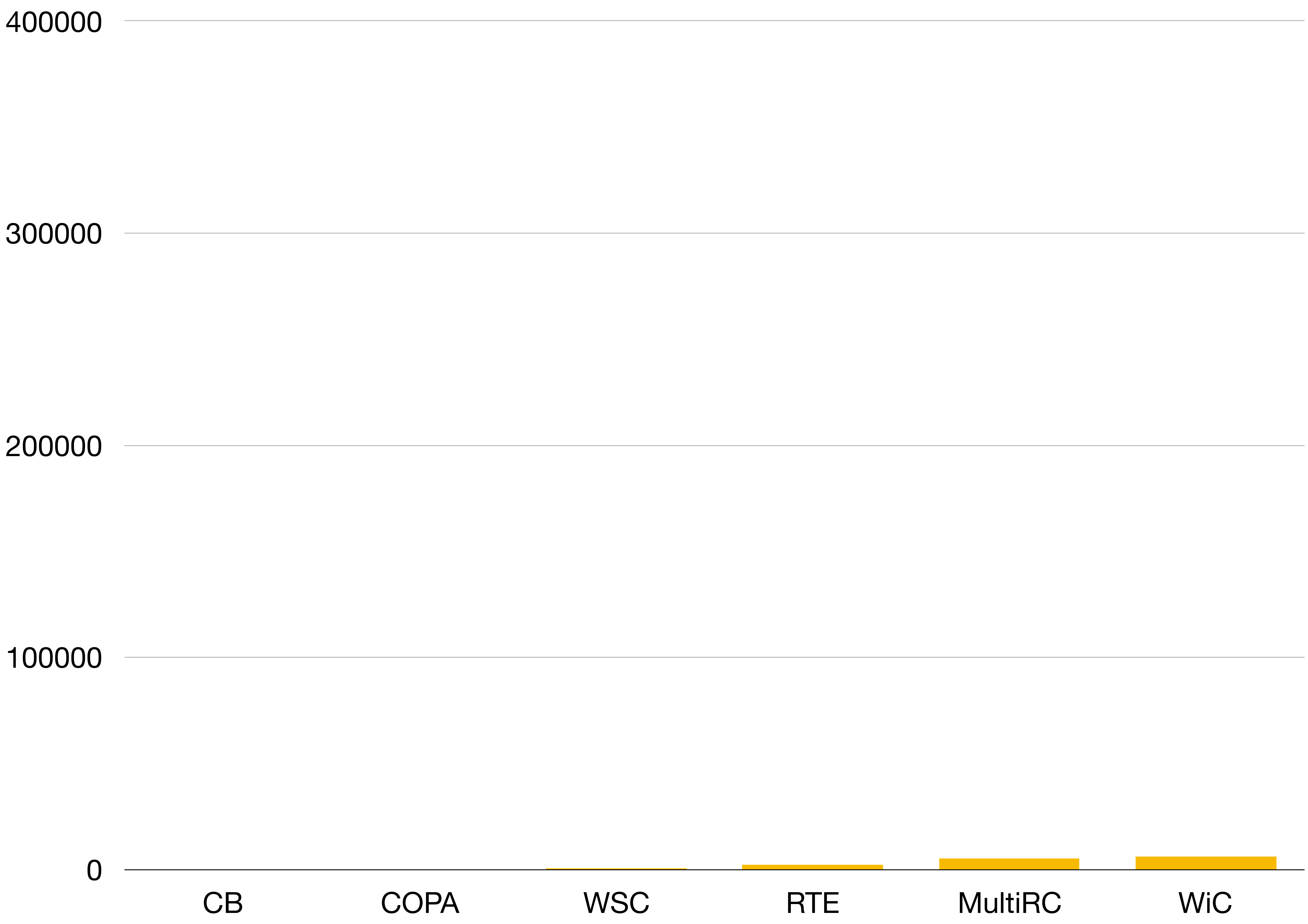
This Spring: SuperGLUE

A revised version of GLUE with:

- A new set of six 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)
- An extensible software toolkit (`jiant`) with built-in support for state-of-the-art methods on the tasks.

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

Corpus	|Train|
CB	250
COPA	400
MultiRC	5100
RTE	2500
WiC	6000
WSC	554



SuperGLUE: The Main Tasks

Corpus	|Train|	|Dev|	|Test|	Task	Metrics	Text Sources
CB	250	57	250	NLI	acc./F1	various
COPA	400	100	500	SC	acc.	online blogs, photography encyclopedia
MultiRC	5100	953	1800	QA	$F1_m/F1_a$	various
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: The Main Tasks

Corpus	|Train|	|Dev|	|Test|	Task	Metrics	Text Sources
CB	250	57	250	NLI	acc./F1	various
COPA	400	100	500	SC	acc.	online blogs, photography encyclopedia
MultiRC	5100	953	1800	QA	$F1_m/F1_a$	various
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

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
CB	250	57	250	NLI	acc./F1	various
COPA	400	100	500	SC	acc.	online blogs, photography encyclopedia
MultiRC	5100	953	1800	QA	$F1_m$/F1	[illegible]
RTE	2500	278	300	NLI	acc.	[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	$F1_m/F1_a$	various
RTE	2500	278	300	NLI	acc.	

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

Winograd Schema Challenge

Pilehvar and Camacho-Collados et al. '19

- **Same data and task as WNLI, but using a standard Boolean coreference format, without recasting.**

Text: *Mark told Pete many lies about himself, which Pete included in his book. He should have been more truthful.* **Coreference:** `False`

COPA	400	100	500	SC	acc.	online blogs, photography encyclopedia
MultiRC	5100	953	1800	QA	$F1_m/F1_a$	various
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: The Main Tasks

Corpus	|Train|	|Dev|	|Test|	Task	Metrics	Text Sources
CB	250	57	250	NLI	acc./F1	various
COPA	400	100	500	SC	acc.	online blogs, photography encyclopedia
MultiRC	5100	953	1800	QA	$\text{F1}_m/\text{F1}_a$	various
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

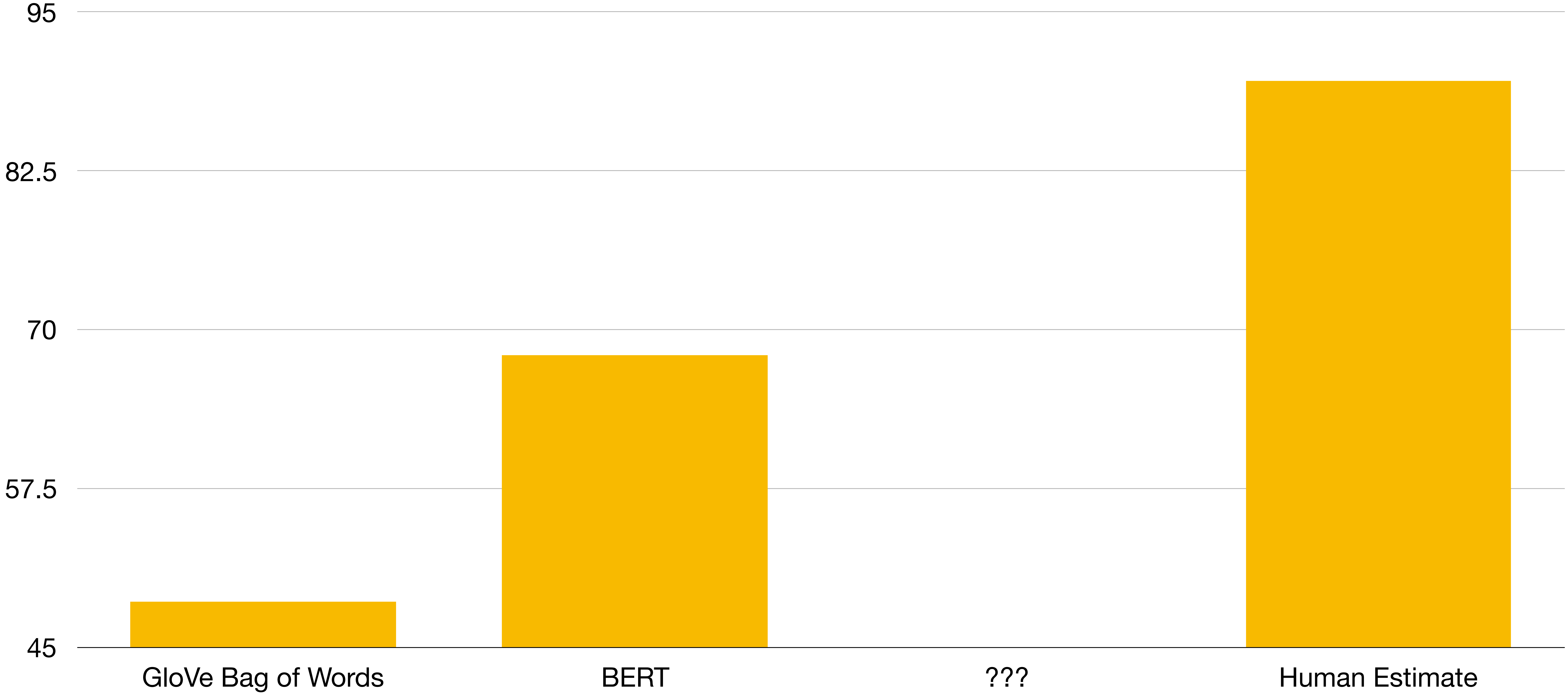
{Wang, Pruksachatkun, Nangia}, Singh, Michael, Hill, Levy & Bowman '19

SuperGLUE

- Preliminary public release out now:
    - `super.gluebenchmark.com`
- Final release coming in mid-summer.
- Expect additional tasks!

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

SuperGLUE Score

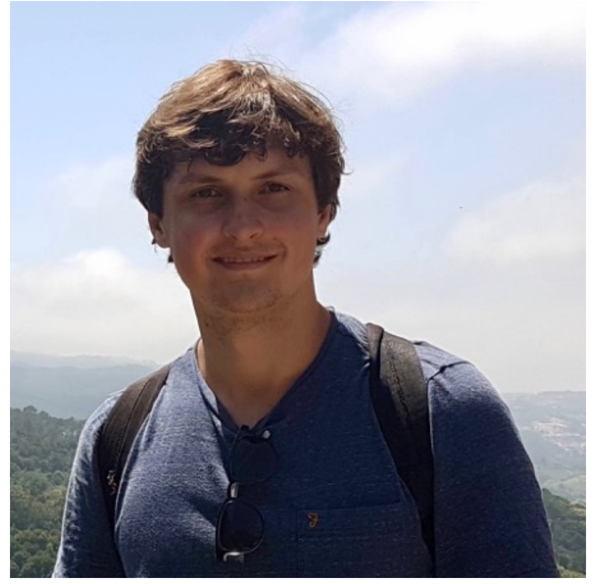


{Wang, Pruksachatkun, Nangia}, Singh, Michael, Hill, Levy & Bowman '19

GLUE and SuperGLUE: Limitations

- GLUE and SuperGLUE are built only on English data.
  - Sentence representation learning may look quite different in lower-resource languages!
- GLUE and SuperGLUE don't evaluate text ***generation***, and use only small amounts of context.
  - Isolates the problem of extracting sentence meaning, but avoids other hard parts of NLP.
- 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***.

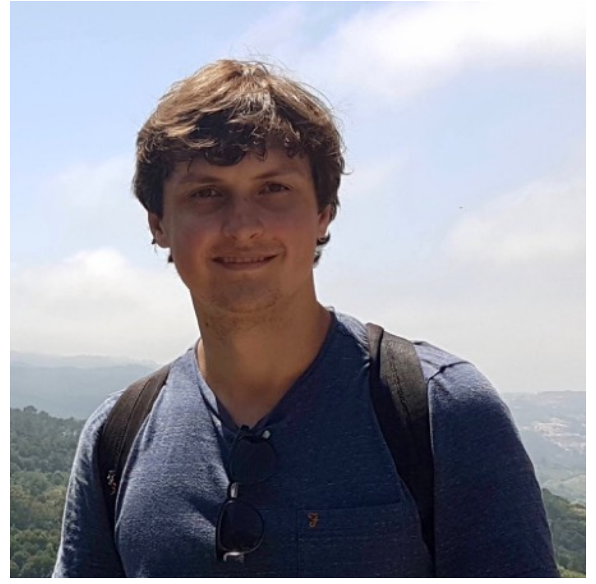
Muppets on STILTs?



- 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

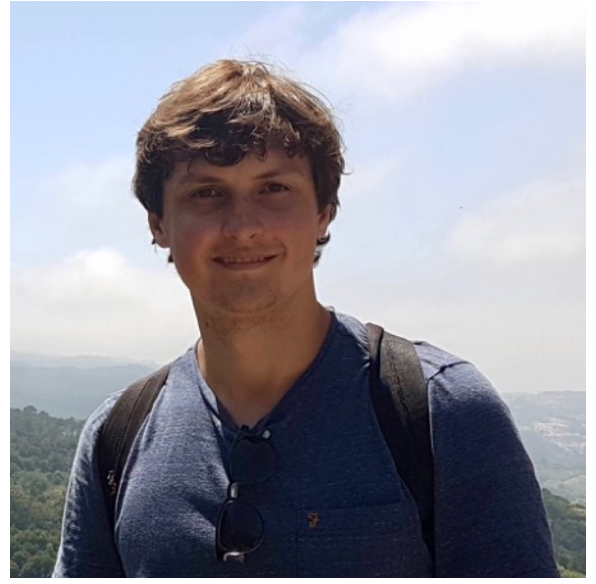
BERT on STILTs



- +1.5 on GLUE w/ MNLI and QQP
- +3.8 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

Phang, Févry & Bowman '18

BERT on STILTs

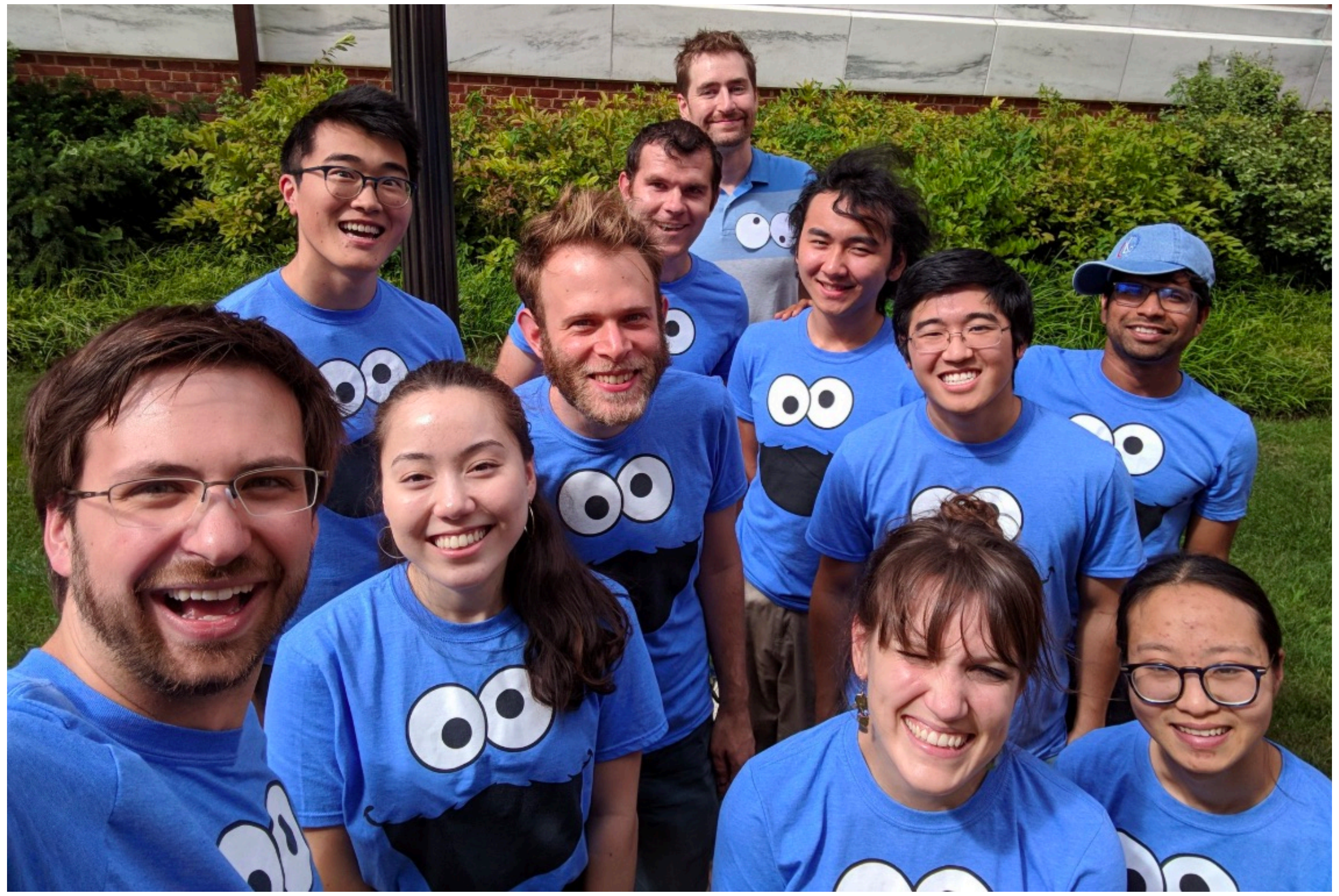


- +1.5 on GLUE w/ MNLI and QQP
- +3.8 on SuperGLUE w/ MNLI
- *Clark et al.* '19: +[illegible] on BoolQ w/ MNLI
- *Sap et al.* '18: +4 to +8 on commonsense tasks w/ SocialIQA

Tuning Not Required!

Phang, Févry & Bowman '18

A Few Big Empirical Studies

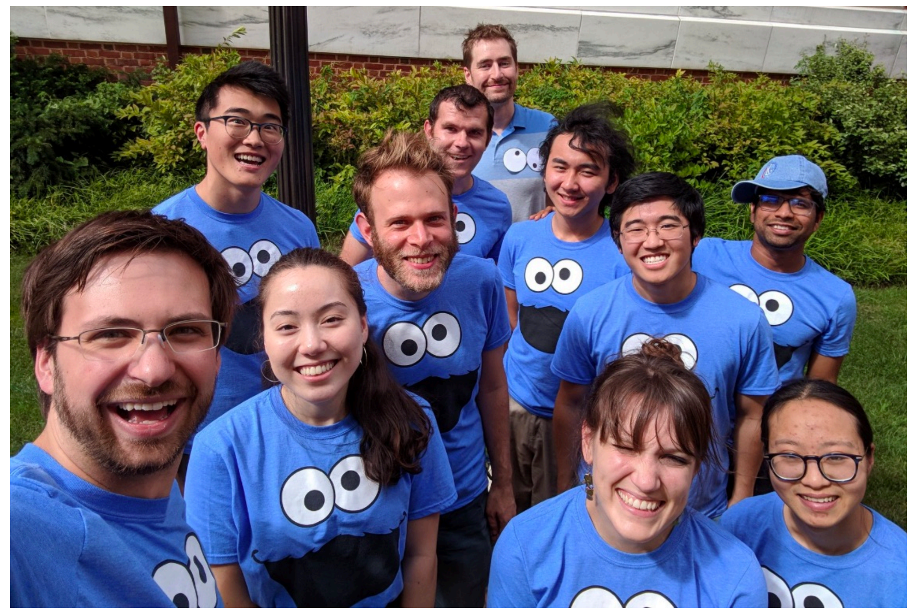


JSALT 2018 at Johns Hopkins U.:

- ~Twenty people, six weeks.
- Question:
  - What pretraining tasks are suitable for what target tasks and why?
- Today: Papers emerging from follow-up work
  - NYU: What pretraining tasks work?
  - Google: What do big language models know?
  - JHU/Brown: What do pretrained *NLI* models know?
    (****SEM best paper, tomorrow at noon***)

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

ELMo and BERT Base on STILTs

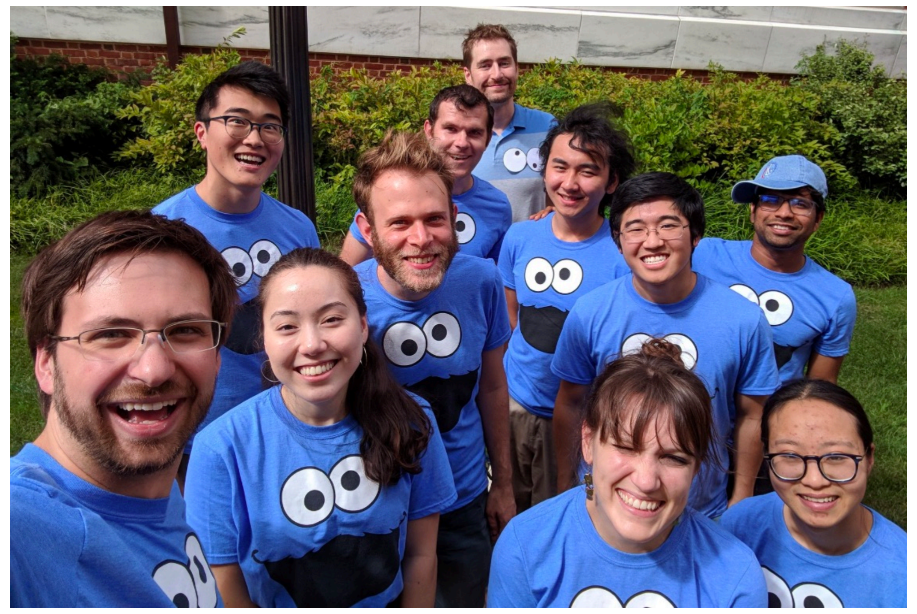


Intermediate Task	Avg	CoLA	SST							
		ELMo								
Random^E	70.5	38.5	87.7							
Single-Task^E	71.2	39.4	**90.6**							
CoLA^E	71.1	39.4	87.3							
SST^E	71.2	38.8	**90.6**							
MRPC^E	71.3	40.0	88.4							
QQP^E	70.8	34.3	88.6							
STS^E	71.6	39.9	88.4							
MNLI^E	72.1	38.9	89.0							
QNLI^E	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

- 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

Pretrained LSTMs



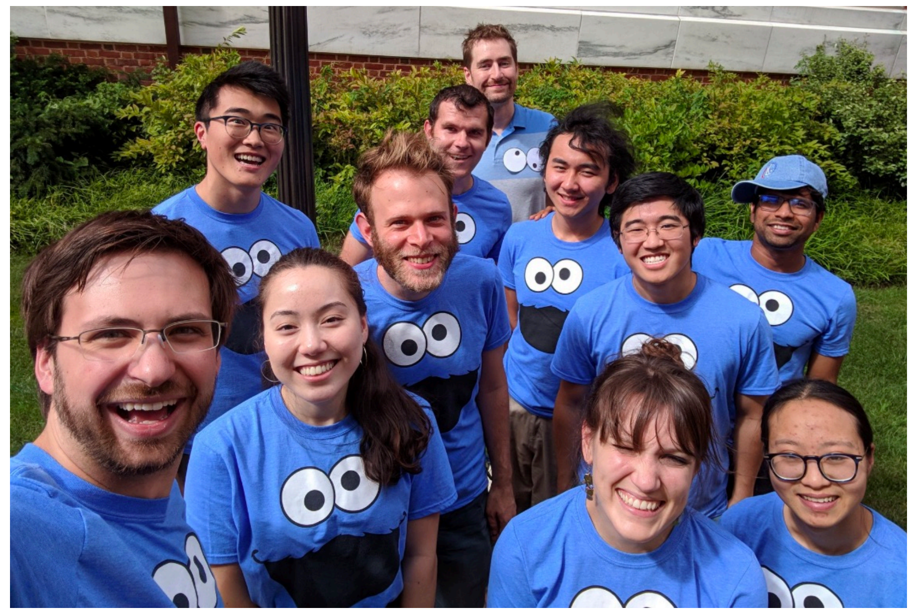
Pretr.	Avg	CoLA	SST	M						
Random	68.2	16.9	84.3	77.						
Single-Task	69.1	21.3	89.0	77.						
		GLUE Tasks as Pretraining Tasks								
CoLA	68.2	21.3	85.7	75.0/83.7	85.7/82.4	79.0/80.3	72.7	78.4	56.3	15.5*
SST	68.6	16.4	89.0	76.0/84.2	84.4/81.6	80.6/81.4	73.9	78.5	58.8	19.7*
MRPC	68.2	16.4	85.6	77.2/84.7	84.4/81.8	81.2/82.2	73.6	79.3	56.7	22.5*
QQP	68.0	14.7	86.1	77.2/84.5	84.7/81.9	81.1/82.0	73.7	78.2	57.0	45.1*
STS	67.7	14.1	84.6	77.9/85.3	81.7/79.2	81.4/82.2	73.6	79.3	57.4	43.7*
MNLI	69.1	16.7	88.2	78.9/85.2	84.5/81.5	81.8/82.6	74.8	**79.6**	58.8	36.6*
QNLI	67.9	15.6	84.2	76.5/84.2	84.3/81.4	80.6/81.8	73.4	78.8	58.8	**56.3**
RTE	68.1	18.1	83.9	77.5/85.4	83.9/81.2	81.2/82.2	74.1	79.1	56.0	39.4*
WNLI	68.0	16.3	84.3	76.5/84.6	83.0/80.5	81.6/82.5	73.6	78.8	58.1	11.3*
		Non-GLUE Pretraining Tasks								
DisSent WP	68.6	18.3	86.6	79.9/86.0	85.3/82.0	79.5/80.5	73.4	79.1	56.7	42.3*
LM WP	70.1	30.8	85.7	76.2/84.2	86.2/82.9	79.2/80.2	74.0	79.4	60.3	25.4*
LM BWB	**70.4**	30.7	86.8	79.9/86.2	**86.3/83.2**	80.7/81.4	74.2	79.0	57.4	47.9*
MT En-De	68.1	16.7	85.4	77.9/84.9	83.8/80.5	82.4/82.9	73.5	**79.6**	55.6	22.5*
MT En-Ru	68.4	16.8	85.1	79.4/86.2	84.1/81.2	82.7/83.2	74.1	79.1	56.0	26.8*
Reddit	66.9	15.3	82.3	76.5/84.6	81.9/79.2	81.5/81.9	72.7	76.8	55.6	53.5*
SkipThought	68.7	16.0	84.9	77.5/85.0	83.5/80.7	81.1/81.5	73.3	79.1	**63.9**	49.3*
		Multitask Pretraining								
MTL GLUE	68.9	15.4	**89.9**	78.9/86.3	82.6/79.9	**82.9/83.5**	**74.9**	78.9	57.8	38.0*
MTL Non-GLUE	69.9	30.6	87.0	**81.1/87.6**	86.0/82.2	79.9/80.6	72.8	78.9	54.9	22.5*
MTL All	**70.4**	**33.2**	88.2	78.9/85.9	85.5/81.8	79.7/80.0	73.9	78.7	57.4	33.8*
		Test Set Results								
LM BWB	66.5	29.1	86.9	75.0/82.1	82.7/63.3	74.0/73.1	73.4	68.0	51.3	65.1
MTL All	68.5	36.3	88.9	77.7/84.8	82.7/63.6	77.8/76.7	75.3	66.2	53.2	65.1

Overall result:

- Nothing works as well as language modeling.
- ...but everything works nearly as well as language modeling.
- Multi-task pretraining helps, but only slightly.

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

Pretrained LSTMs

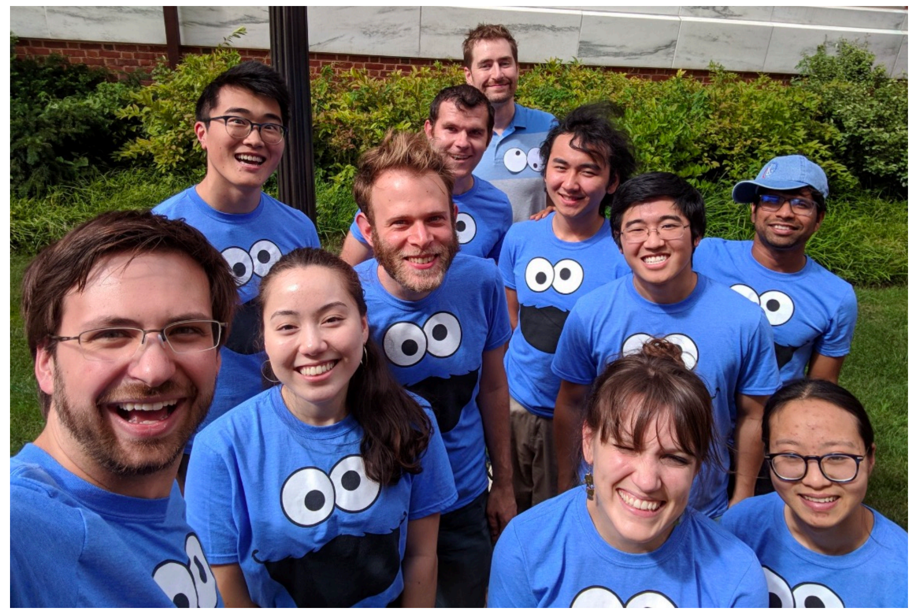


Pretr.	Avg	CoLA	SST	MRPC	QQP	STS	MNLI	QNLI	RTE	WNLI
					Baselines					
Random	68.2	16.9	84.3	77.7/85.6	83.0/80.6	81.7/82.6	73.9	**79.6**	57.0	31.0*
Single-Task	69.1	21.3	89.0	77.2/84.7	84.7/81.9	81.4/82.2	74.8	78.8	56.0	11.3*
					GLUE Tasks as Pretraining Tasks					
CoLA	68.2	21.3	85.7	75.0/83.7	85.7/82.4	79.0/80.3	72.7	78.4	56.3	15.5*
SST	68.6	16.4	89.0	76.0/84.2	84.4/81.6	80.6/81.4	73.9	78.5	58.8	19.7*
MRPC	68.2	16.4	85.6	77.2/84.7	84.4/81.8	81.2/82.2	73.6	79.3	56.7	22.5*
QQP	68.0	14.7	86.1	77.2/84.5	84.7/81.9	81.1/82.0	73.7	78.2	57.0	45.1*
STS	67.7	14.1	84.6	77.9/85.3	81.7/79.2	81.4/82.2	73.6	79.3	57.4	43.7*
MNLI	69.1	16.7	88.2	78.9/85.2	84.5/81.5	81.8/82.6	74.8	**79.6**	58.8	36.6*
QNLI	67.9	15.6	84.2	76.5/84.2	84.3/81.4	80.6/81.8	73.4	78.8	58.8	**56.3**
RTE	68.1	18.1	83.9	77.5/85.4	83.9/81.2	81.2/82.2	74.1	79.1	56.0	39.4*
WNLI	68.0	16.3	84.3	76.5/84.6	83.0/80.5	81.6/82.5	73.6	78.8	58.1	11.3*
					Non-GLUE Pretraining Tasks					
DisSent WP	68.6	18.3	86.6	79.9/86.0	85.3/82.0	79.5/80.5	73.4	79.1	56.7	42.3*
LM WP	70.1	30.8	85.7	76.2/84.2	86.2/82.9	79.2/80.2	74.0	79.4	60.3	25.4*
LM BWB	**70.4**	30.7	86.8	79.9/86.2	**86.3/83.2**	80.7/81.4	74.2	79.0	57.4	47.9*
MT En-De	68.1	16.7	85.4	77.9/84.9	83.8/80.5	82.4/82.9	73.5	**79.6**	55.6	22.5*
MT En-Ru	68.4	16.8	85.1	79.4/86.2	84.1/81.2	82.7/83.2	74.1	79.1	56.0	26.8*
Reddit	66.9	15.3	82.3	76.5/84.6	81.9/79.2	81.5/81.9	72.7	76.8	55.6	53.5*
SkipThought	68.7	16.0	84.9	77.5/85.0	83.5/80.7	81.1/81.5	73.3	79.1	**63.9**	49.3*
					Multitask Pretraining					
MTL GLUE	68.9	15.4	**89.9**	78.9/86.3	82.6/79.9	**82.9/83.5**	**74.9**	78.9	57.8	38.0*
MTL Non-GLUE	69.9	30.6	87.0	**81.1/87.6**	86.0/82.2	79.9/80.6	72.8	78.9	54.9	22.5*
MTL All	**70.4**	**33.2**	88.2	78.9/85.9	85.5/81.8	79.7/80.0	73.9	78.7	57.4	33.8*
					Test Set Results					
LM BWB	66.5	29.1	86.9	75.0/82.1	82.7/63.3	74.0/73.1	73.4	68.0	51.3	65.1
MTL All	68.5	36.3	88.9	77.7/84.8	82.7/63.6	77.8/76.7	75.3	66.2	53.2	65.1

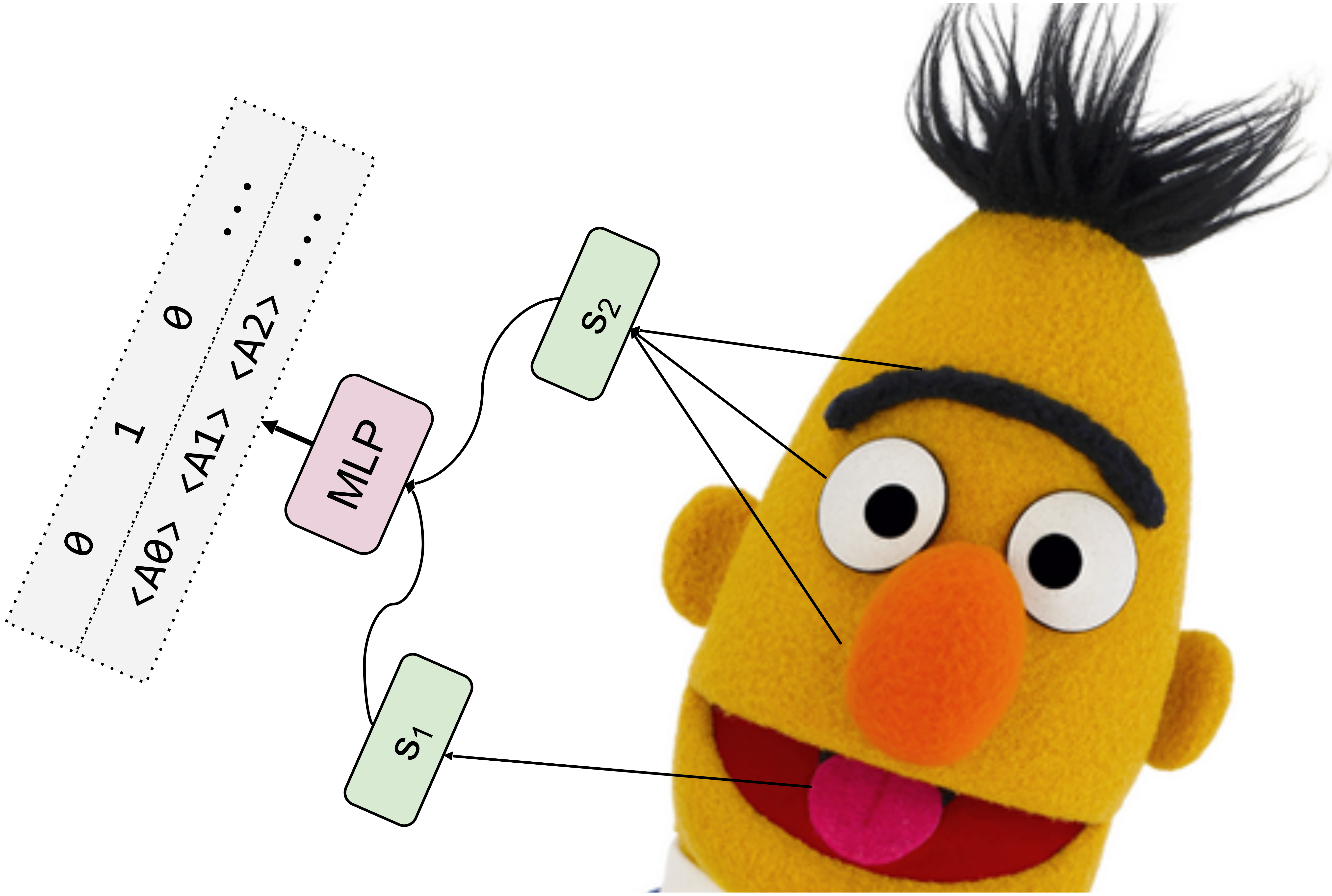
Correlations:

Task	Avg	CoLA	SST	STS	QQP	MNLI	QNLI
CoLA	0.86	1.00					
SST	0.60	0.25	1.00				
MRPC	0.39	0.21	0.34				
STS	-0.36	-0.60	0.01	1.00			
QQP	0.61	0.61	0.27	-0.58	1.00		
MNLI	0.54	0.16	0.66	0.40	0.08	1.00	
QNLI	0.43	0.13	0.26	0.04	0.27	0.56	1.00
RTE	0.34	0.08	0.16	-0.10	0.04	0.14	0.32
WNLI	-0.21	-0.21	-0.37	0.31	-0.37	-0.07	-0.26

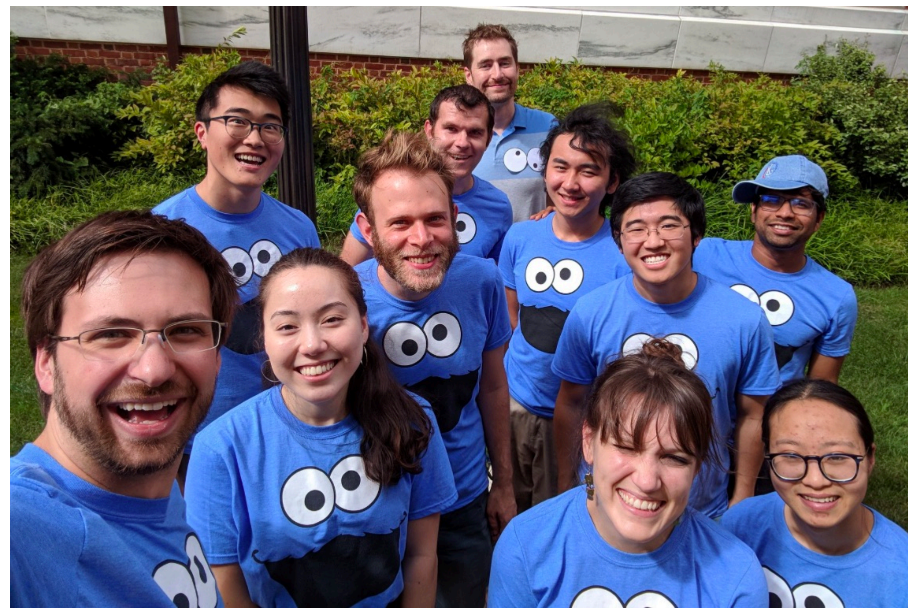
Wang, Hula, Xia, Pappagari, McCoy, Patel, Kim, Tenney, Huang, Yu, Jin, Chen, Van Durme, Grave, Pavlick and Bowman '19



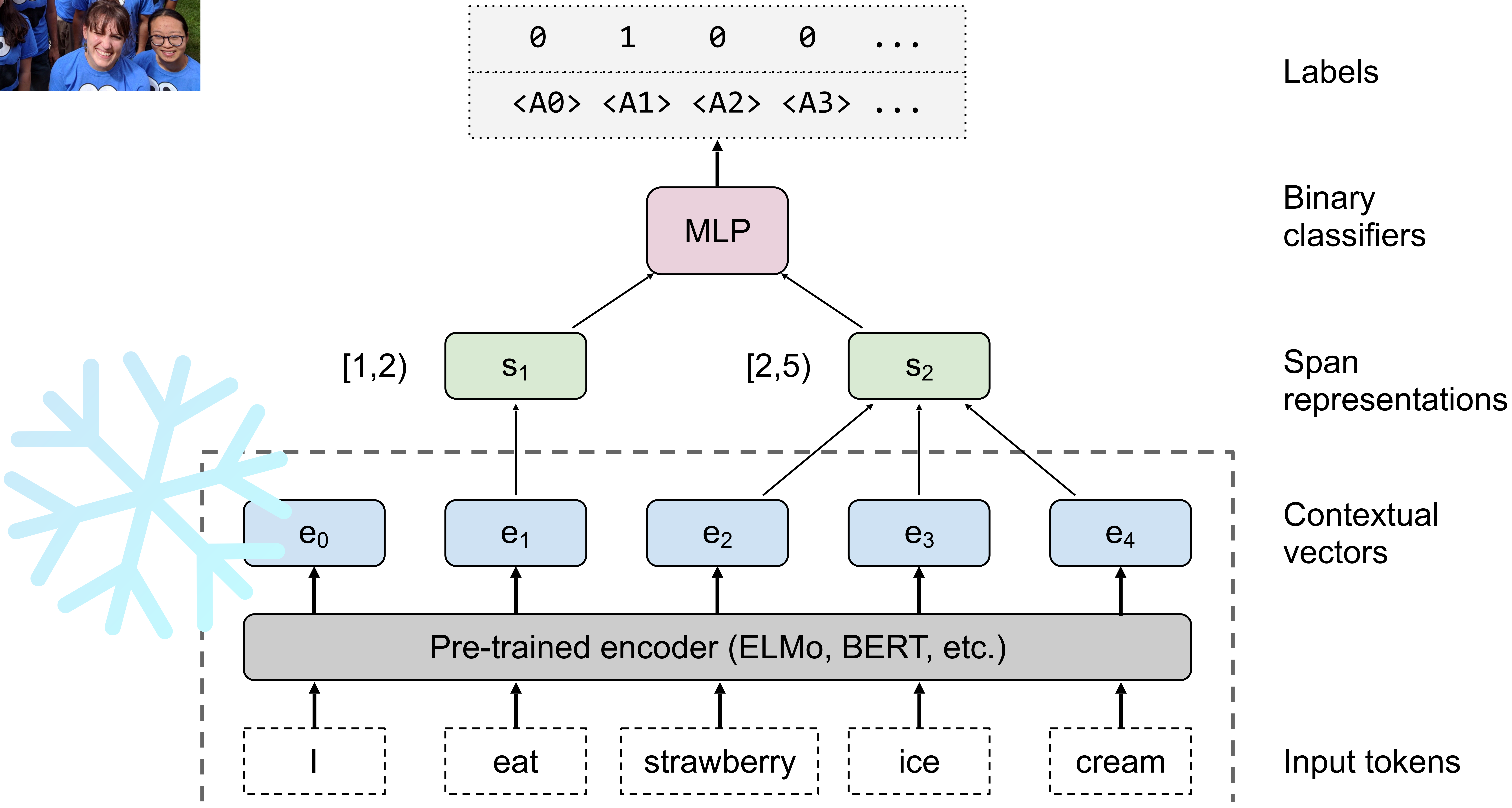
Another View: Edge Probing



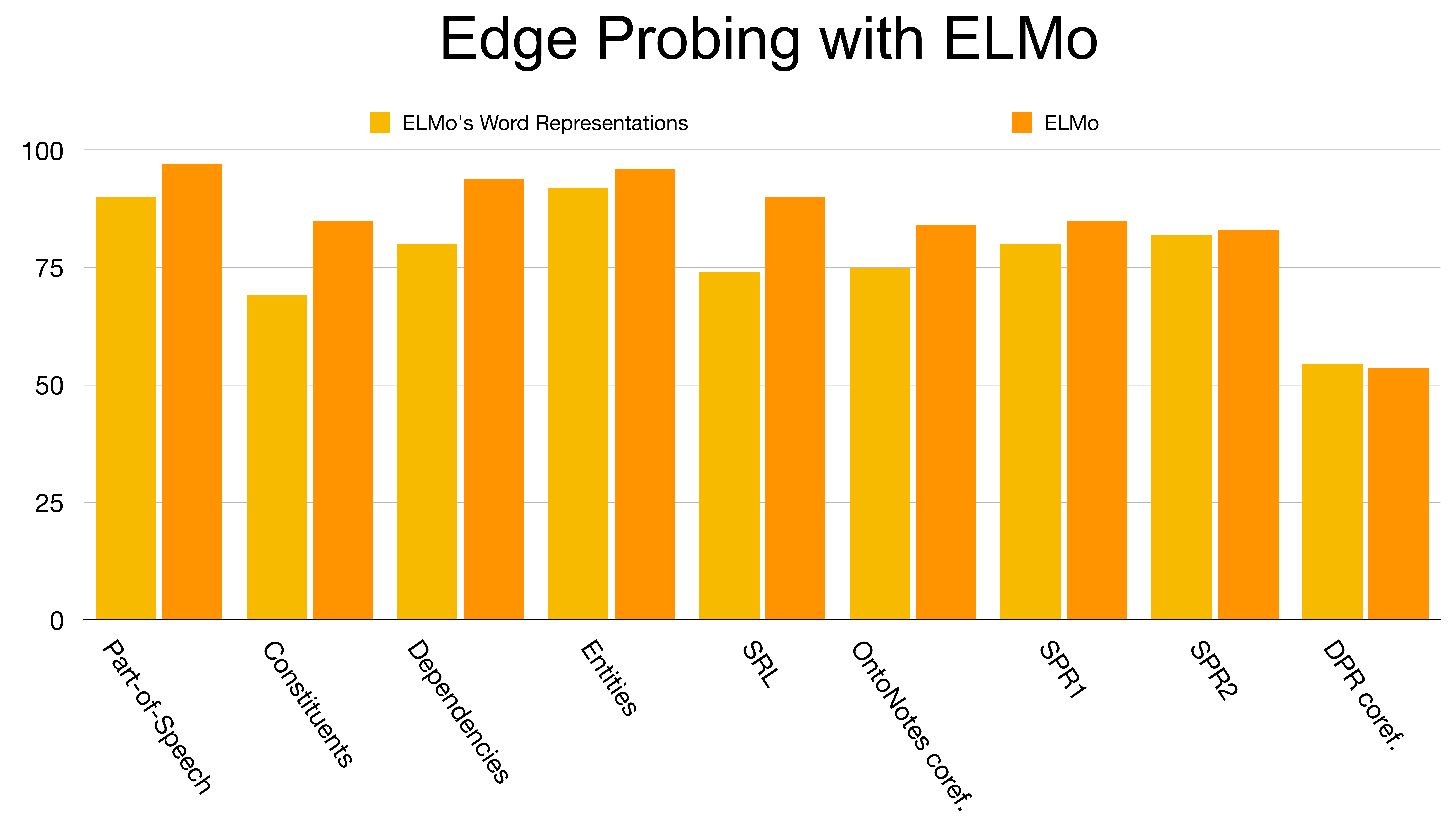
Tenney, Xia, Chen, Wang, Poliak, McCoy, Kim, Van Durme, Bowman, Das, & Pavlick '19

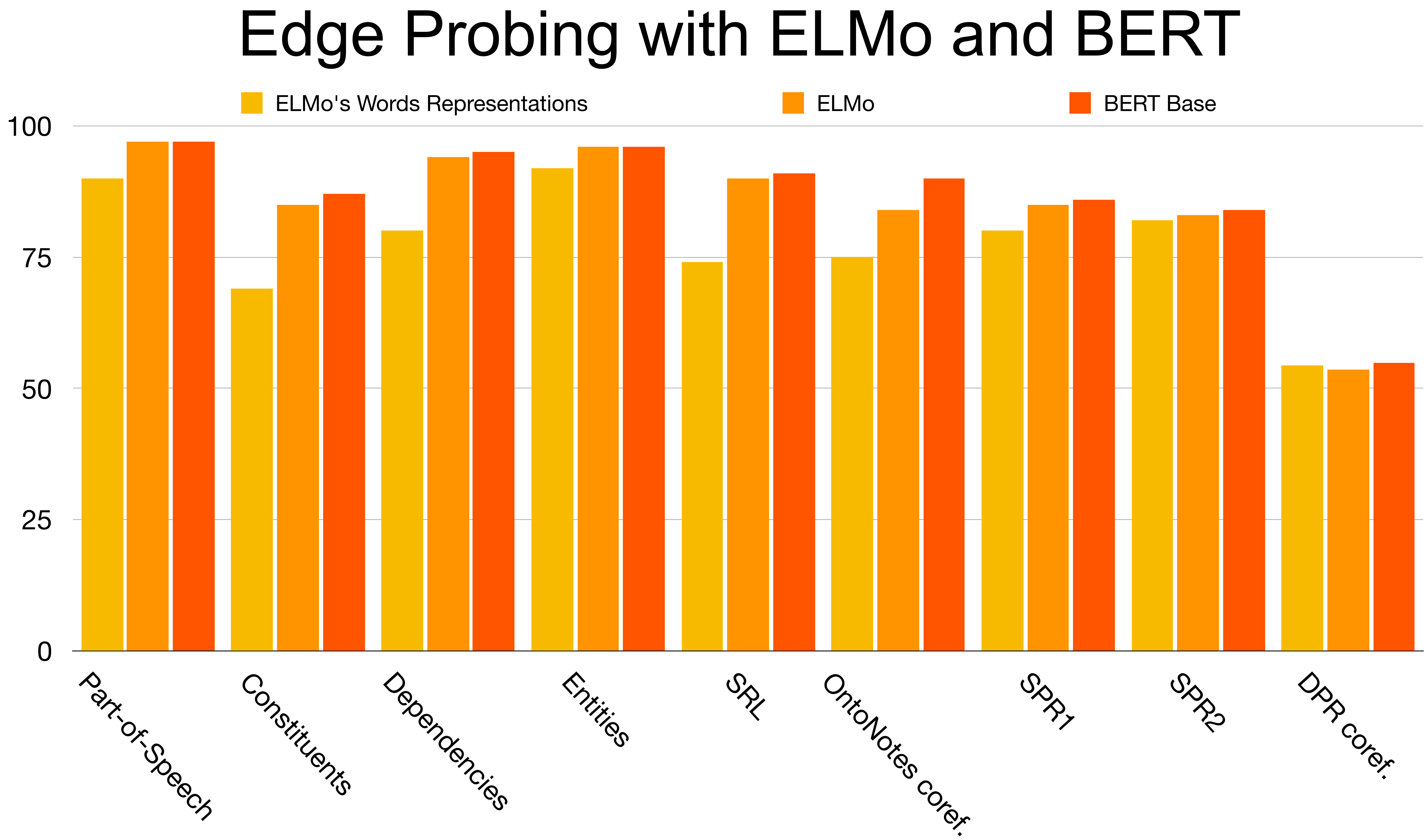


Another View: Edge Probing

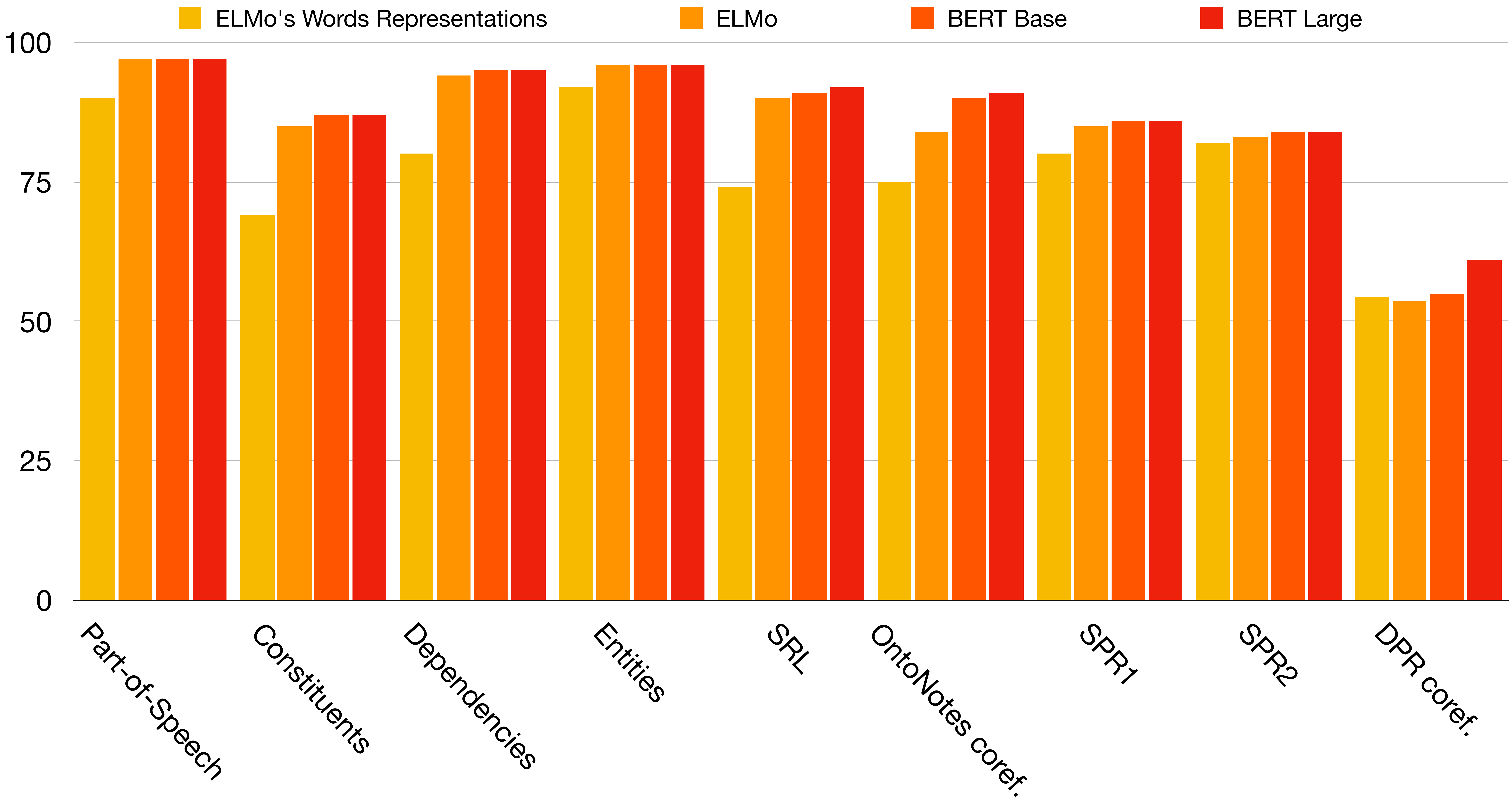


Tenney, Xia, Chen, Wang, Poliak, McCoy, Kim, Van Durme, Bowman, Das, & Pavlick '19


Edge Probing with ELMo
ELMo's Word Representations
ELMo
100
75
50
25
0
Part-of-Speech
Constituents
Dependencies
Entities
SRL
OntoNotes coref.
SPR1
SPR2
DPR coref.


Edge Probing with ELMo and BERT
ELMo's Words Representations
ELMo
BERT Base
100
75
50
25
0
Part-of-Speech
Constituents
Dependencies
Entities
SRL
OntoNotes coref.
SPR1
SPR2
DPR coref.

Edge Probing with ELMo and BERT



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 **BERT.**
  - 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!
- Keep an eye on super.gluebenchmark.com for future developments in this area.

Open Questions

Plenty of open questions!

- What will it take to scale these successes down to hundreds (or tens) of training examples?
- 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?
- Is unsupervised pretraining helping us or hurting us on issues of socially-relevant bias? How do we minimize this bias?

Thanks!

Questions:

bowman@nyu.edu

Try SuperGLUE:

super.gluebenchmark.com

But wait! There's more!

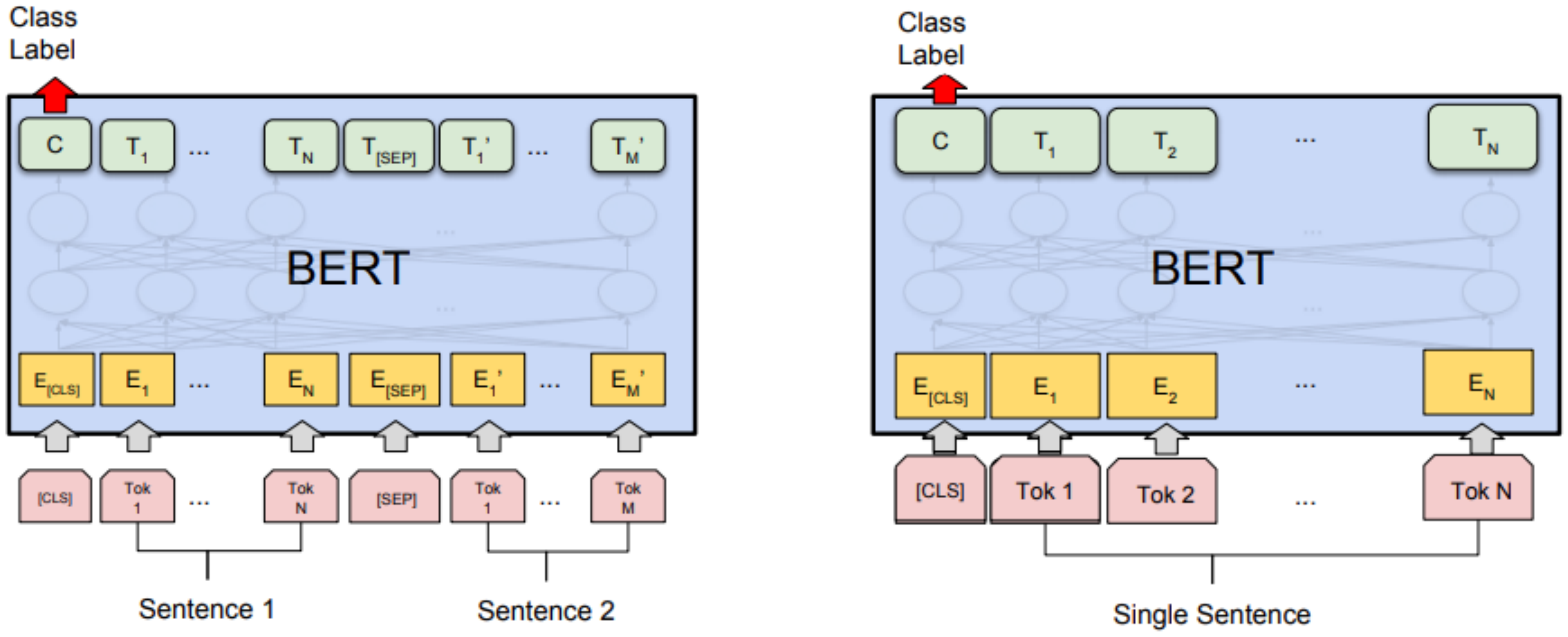
Final Pointers

A bit more analysis worth mentioning:

- BERT can do fairly well at acceptability judgments involving *binding*, *unusual argument structures*, and *some embedded VPs and clauses*, but struggles with *Wh-movement* and *gaps*.
  Warstadt and Bowman '19

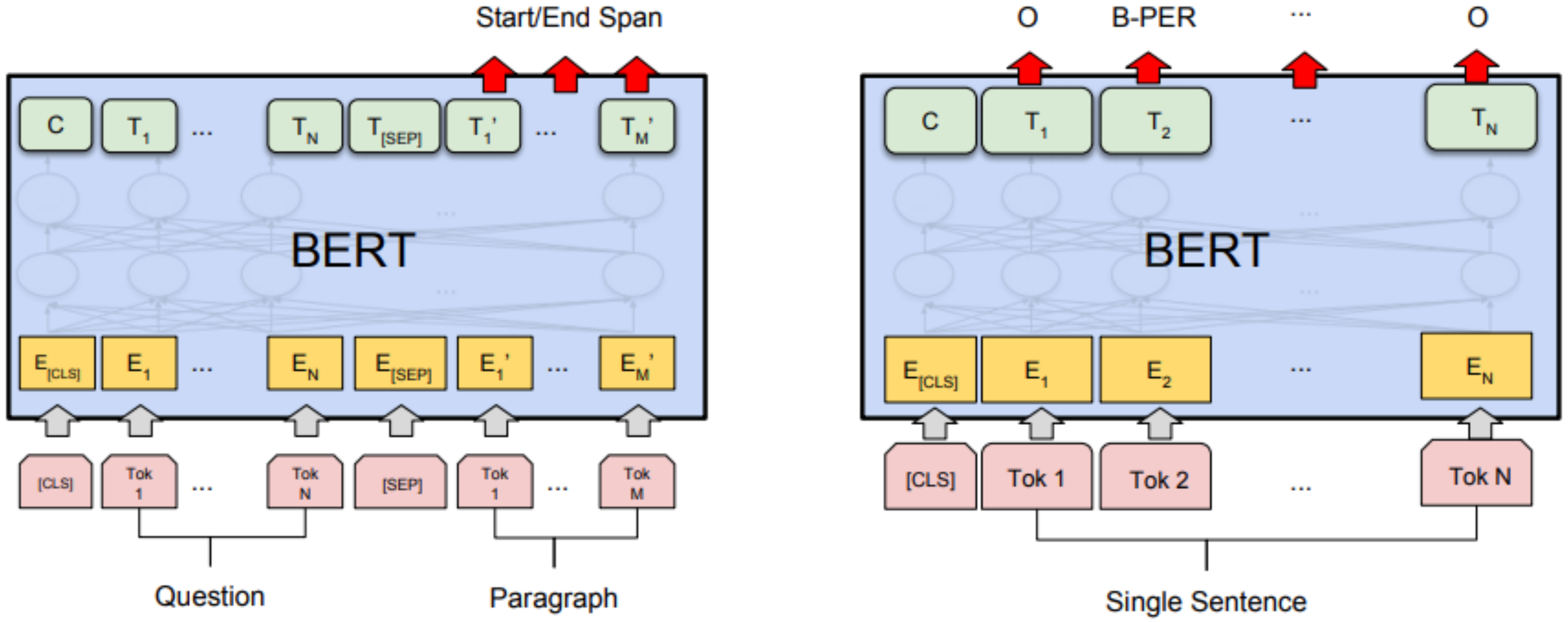
- BERT reaches near-human performance on a variant of the Marvin and Linzen's subject–verb agreement tests, even where past large LMs have failed.
  Goldberg '19; @Thom_Wolf '19, Twitter

- Multilingual variants of BERT, trained on monolingual and parallel data, are showing promise on cross-lingual transfer: Train a task model on English, and test it on Urdu.
  Lample and Conneau '19



(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

(b) Single Sentence Classification Tasks: SST-2, CoLA

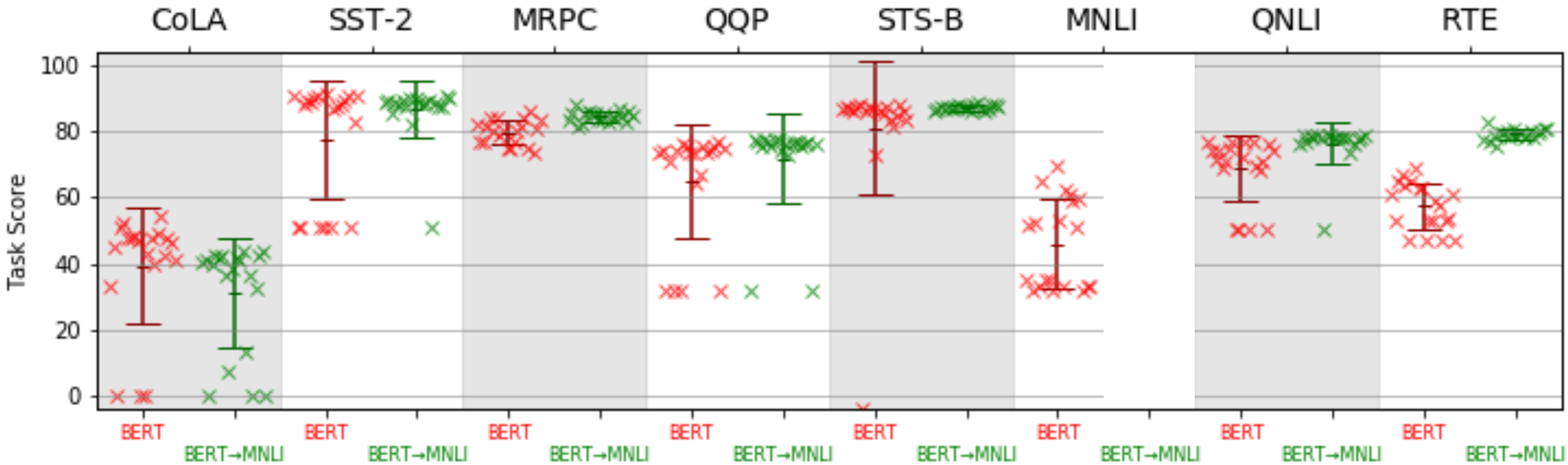


(c) Question Answering Tasks: SQuAD v1.1

(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

Devlin et al. '18

Muppets on STILTs?



Development Set Results with 1k training examples per task.

Phang, Févry & Bowman '18

Five More Views

- gradient minimal pairs: ceiling

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

General-Purpose Representation Learning

Words:

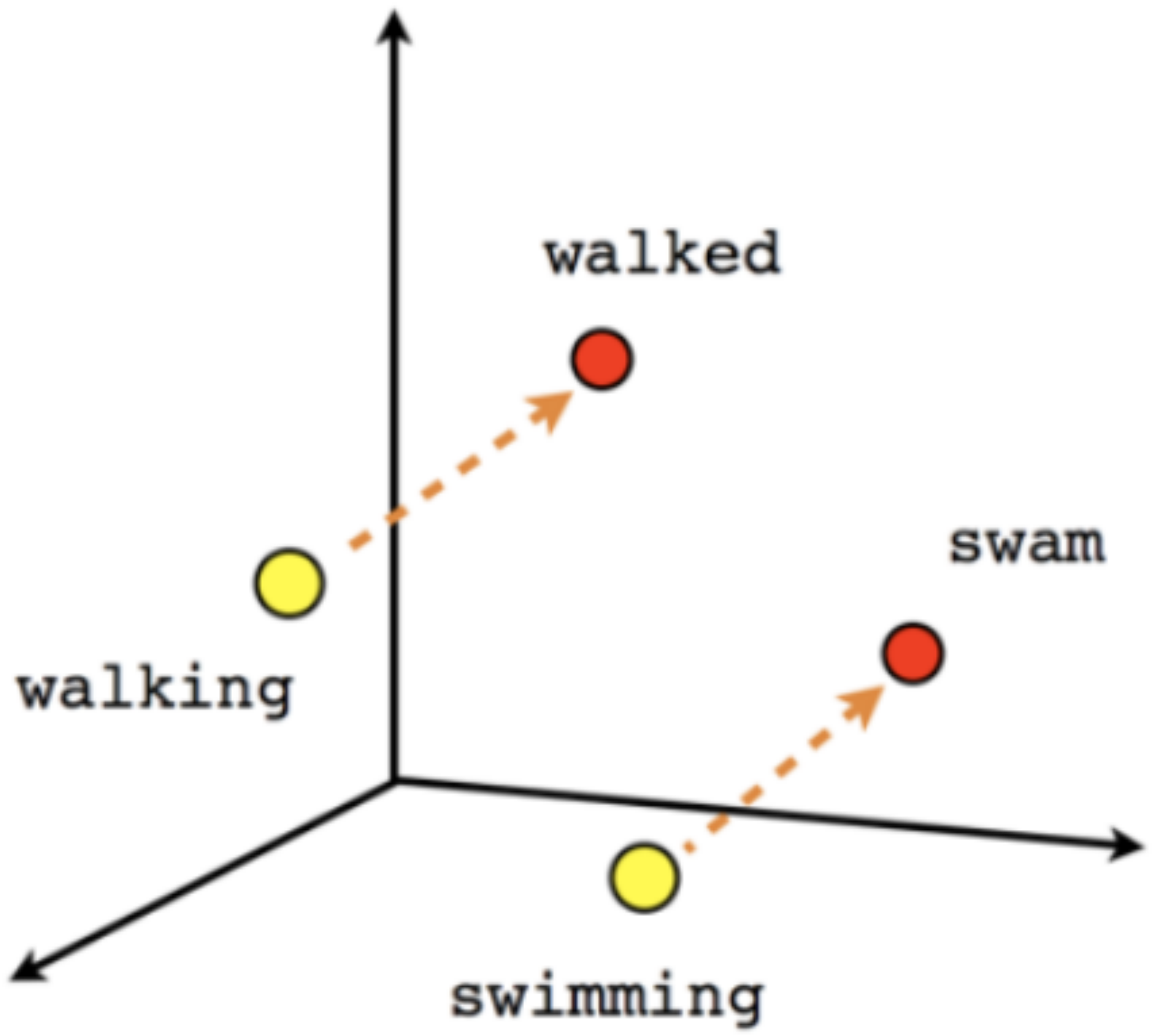
- Distributional *word embeddings*:
  SENNA, word2vec, GloVe, fastText, etc.

Images:

- ImageNet-trained deep CNNs

Sentences:

- Slow start, but dramatic progress over the last eighteen months!



Where might this be valuable?

Scenario 1: *An engineer wants to solve some English sentence classification task for which no data exists.*

Examples:

- Intent detection for a new Alexa skill
- Relation classification for information extraction
- Customer service ticket classification for a new business

...

Where might this be valuable?

Scenario 1: *An engineer wants to solve some English sentence classification task for which no data exists.*

Standard approach:

- Pay to annotate 10k–1m examples at \$0.05–0.50 each

- Train a BiLSTM-based classification model over word embeddings

With effective sentence representations:

- Train a model over the outputs of an existing encoder.

➔ Comparable performance with ~1–10% the data.

Where might this be valuable?

Scenario 2: *An engineer wants to solve some English sentence understanding task for which ample labeled data exists, but performance is still inadequate.*

Examples:

- English–Chinese translation

  ...

Where might this be valuable?

Scenario 2: ***An engineer wants to solve some English sentence understanding task for which ample labeled data exists, but performance is still inadequate.***

Standard approach:

- Train large attention-based NN model over word embeddings

With effective sentence representations:

- Use a general-purpose encoder as the input layer(s) of the model

➔ Prior knowledge of English makes learning more effective

A Too Brief, Very Self Interested History

- 2014: Sentence-to-vector pretraining becomes established as a task
  Dai and Le '14, Kiros et al. '15, Hill et al. '16, Wieting et al. '16, Conneau et al. '17, Subramanian et al. '18...

- 2017: First contextualized word vector pretraining methods appear
  Peters et al. '17, McCann et al. '17 (CoVe), Peters et al. '18 (ELMo)

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

- Summer '18: First major successes with unsupervised pretraining and fine-tuning
  Radford et al. '18 (OpenAI GPT), Devlin et al. '18 (BERT)

- Spring '19: The updated SuperGLUE benchmark launches
  Wang et al. '19

Today!

Plus some analysis!

Choice of Plausible Alternatives

Roemelle et al. '11

- **Multiple choice QA: Which is the most plausible cause (or consequence) of some event?**

Premise: *My body cast a shadow over the grass.* **Question:** *What's the CAUSE for this?*
Alternative 1: *The sun was rising.* **Alternative 2:** *The grass was cut.* **Correct Alternative:** 1

Corpus	|Train|	|Dev|	|Test|	Task	Metrics	Text Sources
CB	250	57	250	NLI	acc./F1	various
COPA	400	100	500	SC	acc.	online blogs, photography encyclopedia
MultiRC	5100	953	1800	QA	$F1_m/F1_a$	various
RTE	2500	278	300	NLI	acc.	
WiC	6000	638	1400	WSD	acc.	

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

Word-in-Context Sense Matching

Pilehvar and Camacho-Collados et al. '19

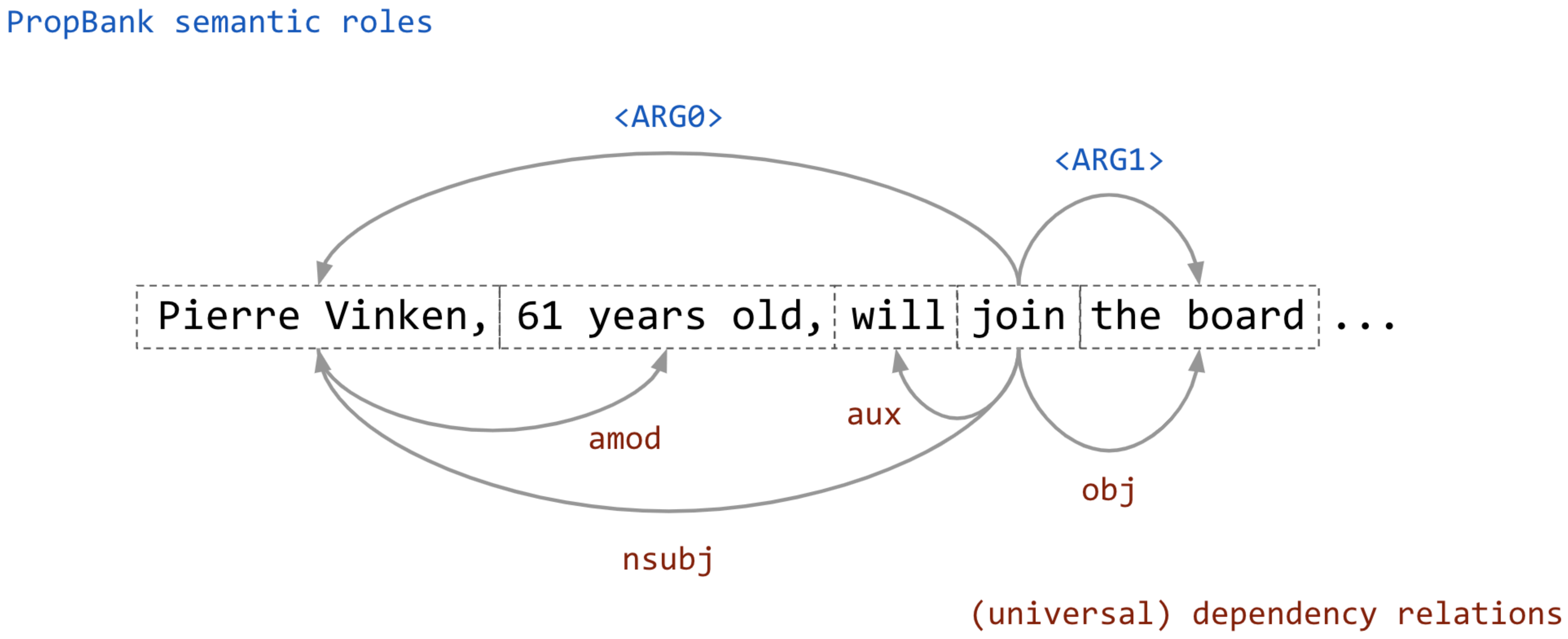
- **Two-way classification: Do two uses of a word follow the same sense?**

Context 1: *Room and board.* **Context 2:** *He nailed boards across the windows.*
Sense match: `False`

CB	250	57	250	NLI	acc./F1	various
COPA	400	100	500	SC	acc.	online blogs, photography encyclopedia
MultiRC	5100	953	1800	QA	$F1_m/F1_a$	various
RTE	2500	278	300	NLI	acc.	news, Wikipedia
WiC	6000	638	1400	WSD	acc.	WordNet, VerbNet, Wiktionary
WSC	554	104	146	coref.	acc.	

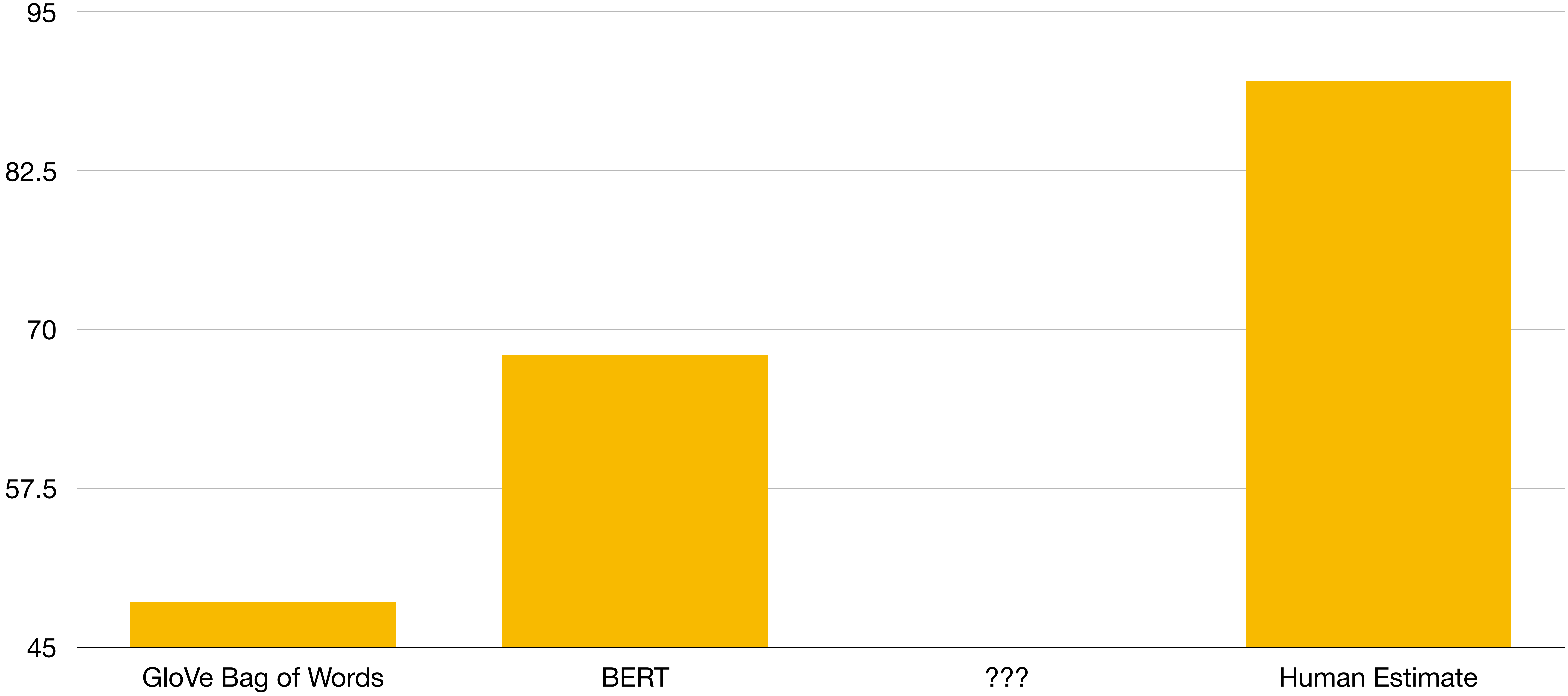
{Wang, Pruksachatkun, Nangia}, Singh, Michael, Hill, Levy & Bowman '19

Another View: Edge Probing



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

BERT on STILTs



Phang, Févry & Bowman '18

BERT on STILTs

Phang, Févry & Bowman '18

Muppets on STILTs?

Training Set Size	Avg	A.Ex	CoLA *8.5k*	SST *67k*	MRPC *3.7k*	QQP *364k*	STS *7k*	MNLI *393k*	QNLI *108k*	RTE *2.5k*
Development Set Scores										
BERT	80.8	78.4	**62.1**	92.5	89.0/92.3	**91.5/88.5**	90.3/90.1	**86.2**	89.4	70.0
BERT→QQP	80.9	78.5	56.8	93.1	88.7/92.0	~~91.5/88.5~~	90.9/90.7	86.1	89.5	74.7
BERT→MNLI	82.4	80.5	59.8	**93.2**	**89.5/92.3**	91.4/88.4	**91.0/90.8**	~~86.2~~	**90.5**	**83.4**
BERT→SNLI	81.4	79.2	57.0	92.7	88.5/91.7	91.4/88.4	90.7/90.6	86.1	89.8	80.1
BERT→Real/Fake	77.4	74.3	52.4	92.1	82.8/88.5	90.8/87.5	88.7/88.6	84.5	88.0	59.6
BERT, Best of Each	**82.6**	**80.8**	**62.1**	**93.2**	**89.5/92.3**	**91.5/88.5**	**91.0/90.8**	**86.2**	**90.5**	**83.4**
GPT	75.4	72.4	**50.2**	**93.2**	80.1/85.9	89.4/85.9	86.4/86.5	**81.2**	82.4	58.1
GPT→QQP	76.0	73.1	48.3	93.1	83.1/88.0	~~89.4/85.9~~	87.0/86.9	80.7	82.6	62.8
GPT→MNLI	76.7	74.2	45.7	92.2	**87.3/90.8**	89.2/85.3	88.1/88.0	~~81.2~~	**82.6**	**67.9**
GPT→SNLI	76.0	73.1	41.5	91.9	86.0/89.9	89.9/86.6	**88.7/88.6**	81.1	82.2	65.7
GPT→Real/Fake	76.6	73.9	49.5	91.4	83.6/88.6	**90.1/86.9**	87.9/87.8	81.0	82.5	66.1
GPT, Best of Each	**77.5**	**75.9**	**50.2**	**93.2**	**87.3/90.8**	**90.1/86.9**	**88.7/88.6**	**81.2**	**82.6**	**67.9**
ELMo	63.8	59.4	15.6	84.9	69.9/80.6	86.4/82.2	64.5/64.4	69.4	73.0	50.9
ELMo→QQP	64.8	61.7	16.6	87.0	73.5/82.4	~~86.4/82.2~~	71.6/72.0	63.9	73.4	52.0
ELMo→MNLI	66.4	62.8	16.4	87.6	73.5/83.0	87.2/83.1	**75.2/75.8**	~~69.4~~	72.4	**56.3**
ELMo→SNLI	66.4	62.7	14.8	**88.4**	**74.0/82.5**	**87.3/83.1**	74.1/75.0	69.7	**74.0**	56.0
ELMo→Real/Fake	66.9	63.3	**27.3**	87.8	72.3/81.3	87.1/83.1	70.3/70.6	**70.3**	73.7	54.5
ELMo, Best of Each	**68.0**	**64.8**	**27.3**	**88.4**	**74.0/82.5**	**87.3/83.1**	**75.2/75.8**	**70.3**	**74.0**	**56.3**

Phang, Févry & Bowman '18