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
- Detour: A few things we've learned about modern models
  Warstadt et al. '19, Pruskachatkun et al. '20, Phang et al. '20
- What's next for evaluation?
  Idle speculation '20

GLUE: What is it?

GLUE

The General Language Understanding Evaluation (GLUE):

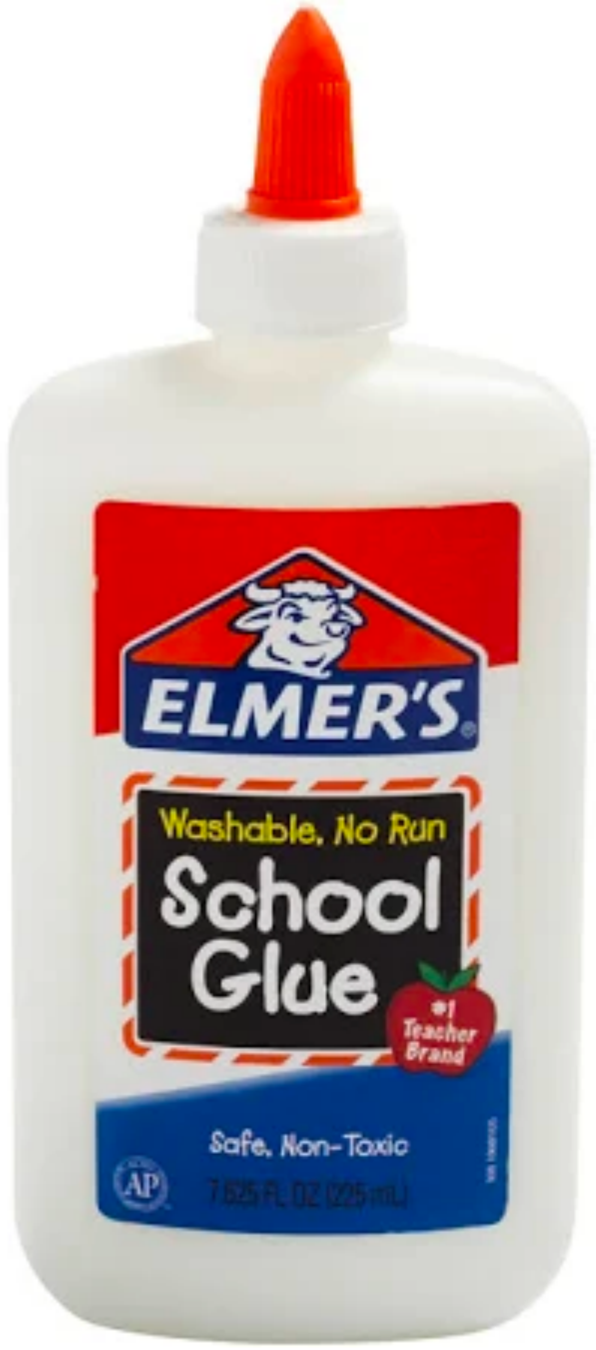
An open-ended competition and evaluation platform for general-purpose sentence encoders.

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:

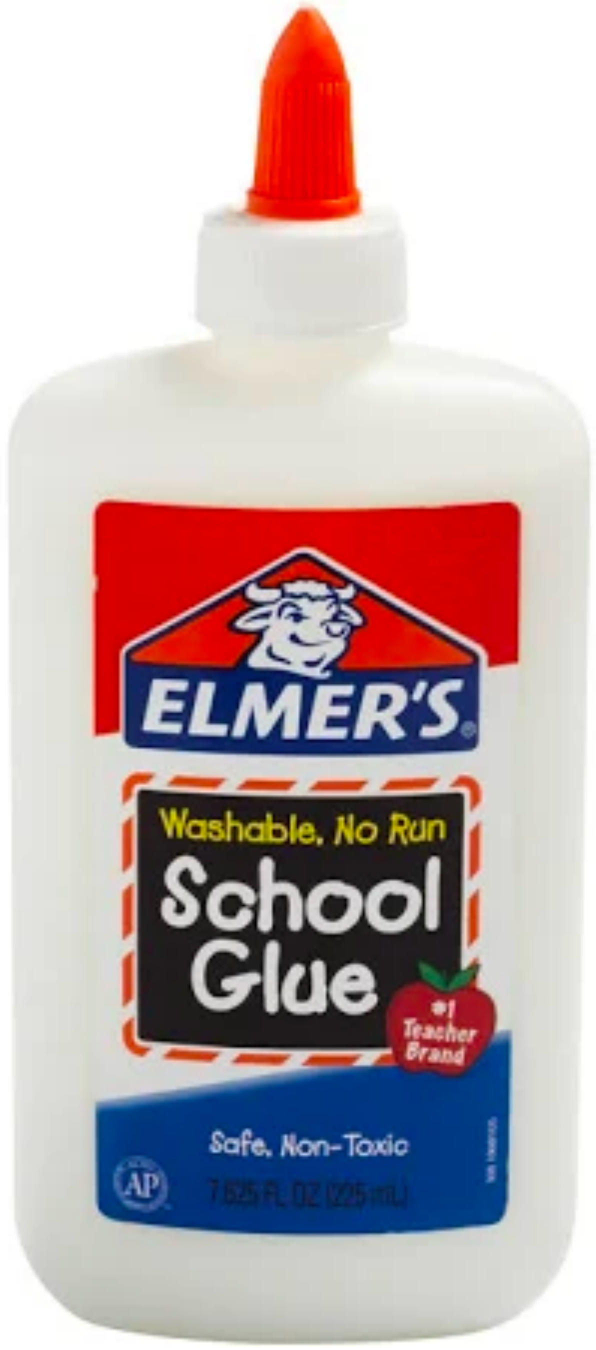
- Unsolved
- Varied training data volume
- Varied language style/genre

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.*

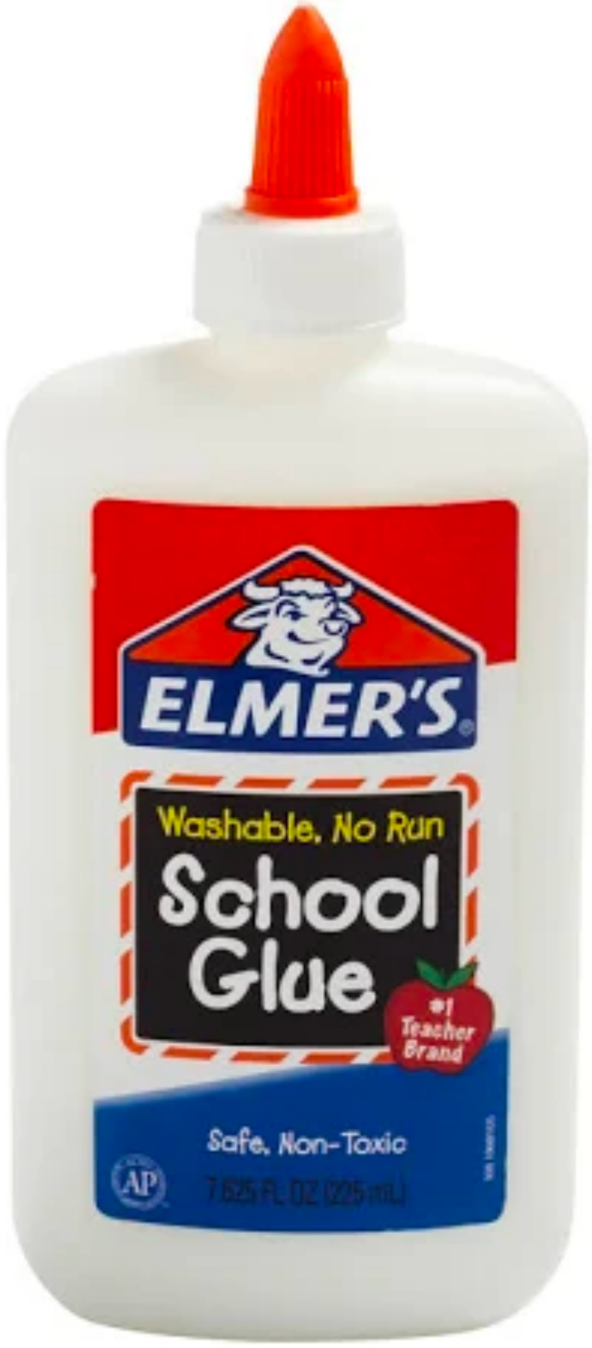


Wang, Singh, Michael, Hill, Levy & Bowman ICLR '19

GLUE

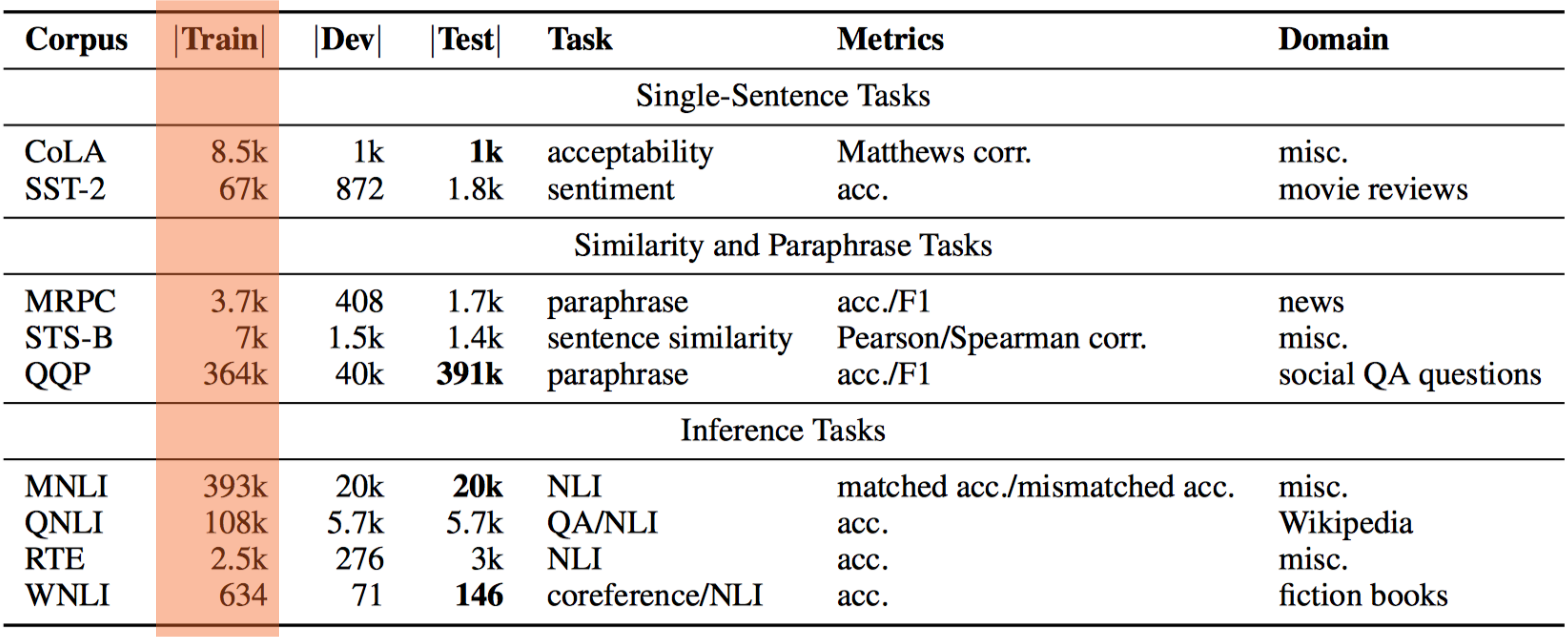
Simple leaderboard API: Upload predictions for a test set

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



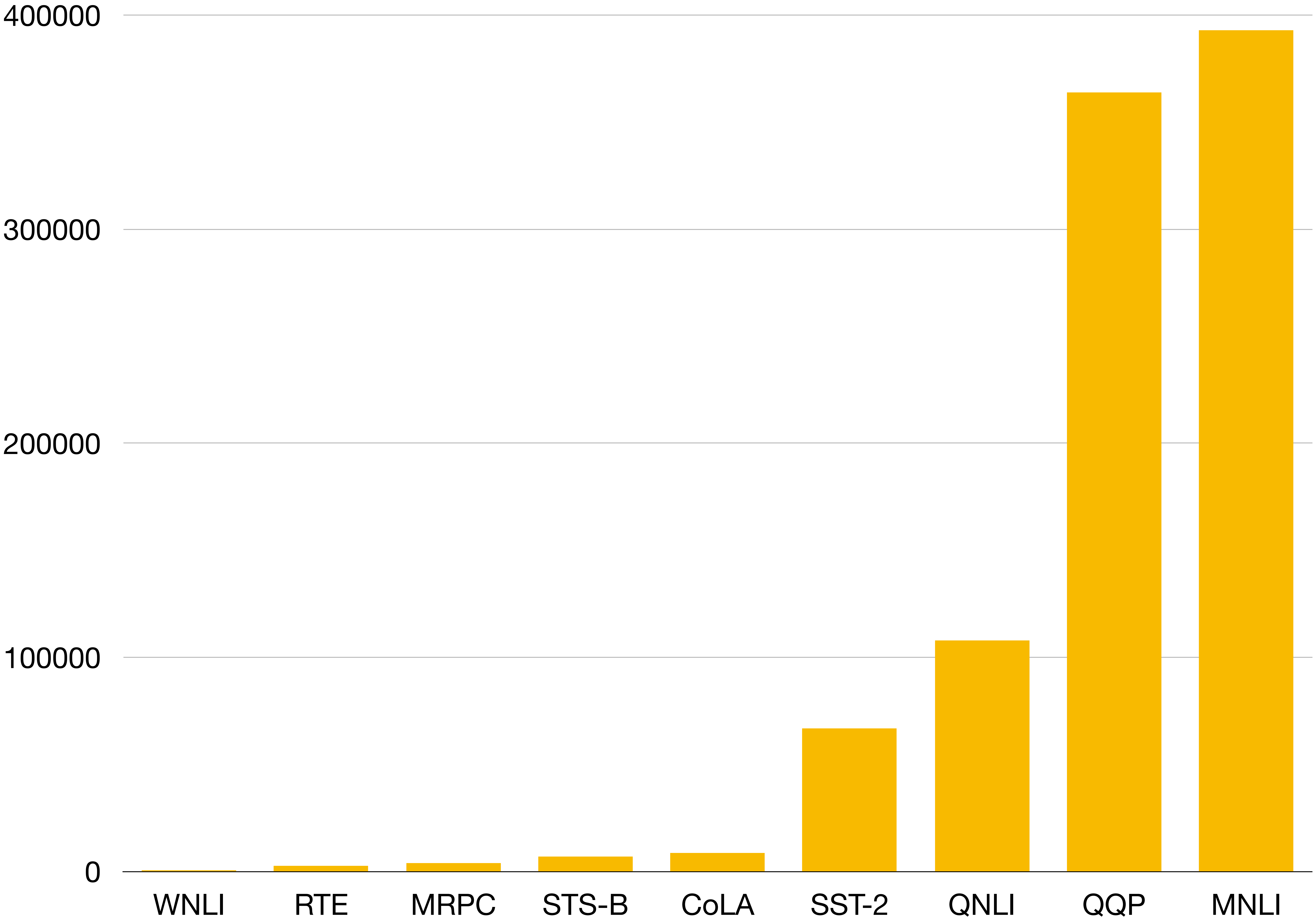
Wang, Singh, Michael, Hill, Levy & Bowman ICLR '19

GLUE: The Main Tasks

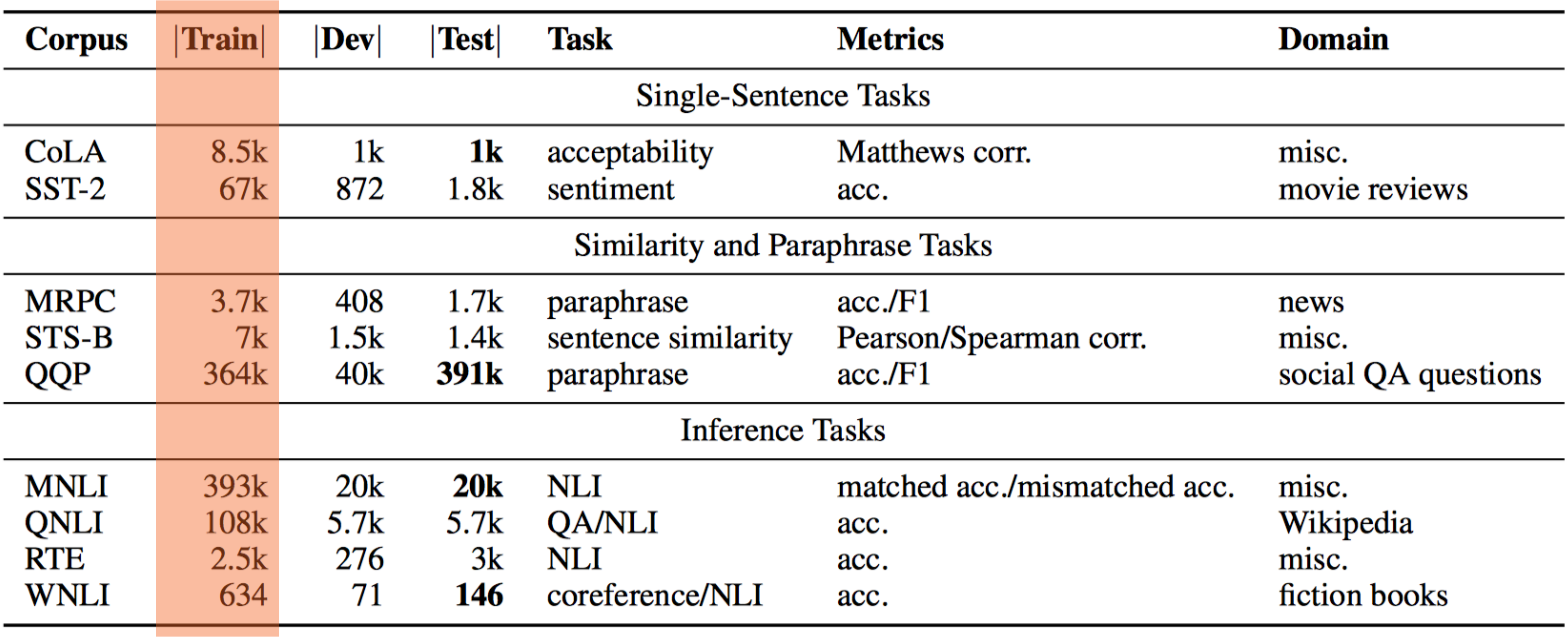


Corpus	**|Train|**	**|Dev|**	**|Test|**	**Task**	**Metrics**	**Domain**
				Single-Sentence Tasks		
CoLA	8.5k	1k	**1k**	acceptability	Matthews corr.	misc.
SST-2	67k	872	1.8k	sentiment	acc.	movie reviews
				Similarity and Paraphrase Tasks		
MRPC	3.7k	408	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.5k	1.4k	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

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


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

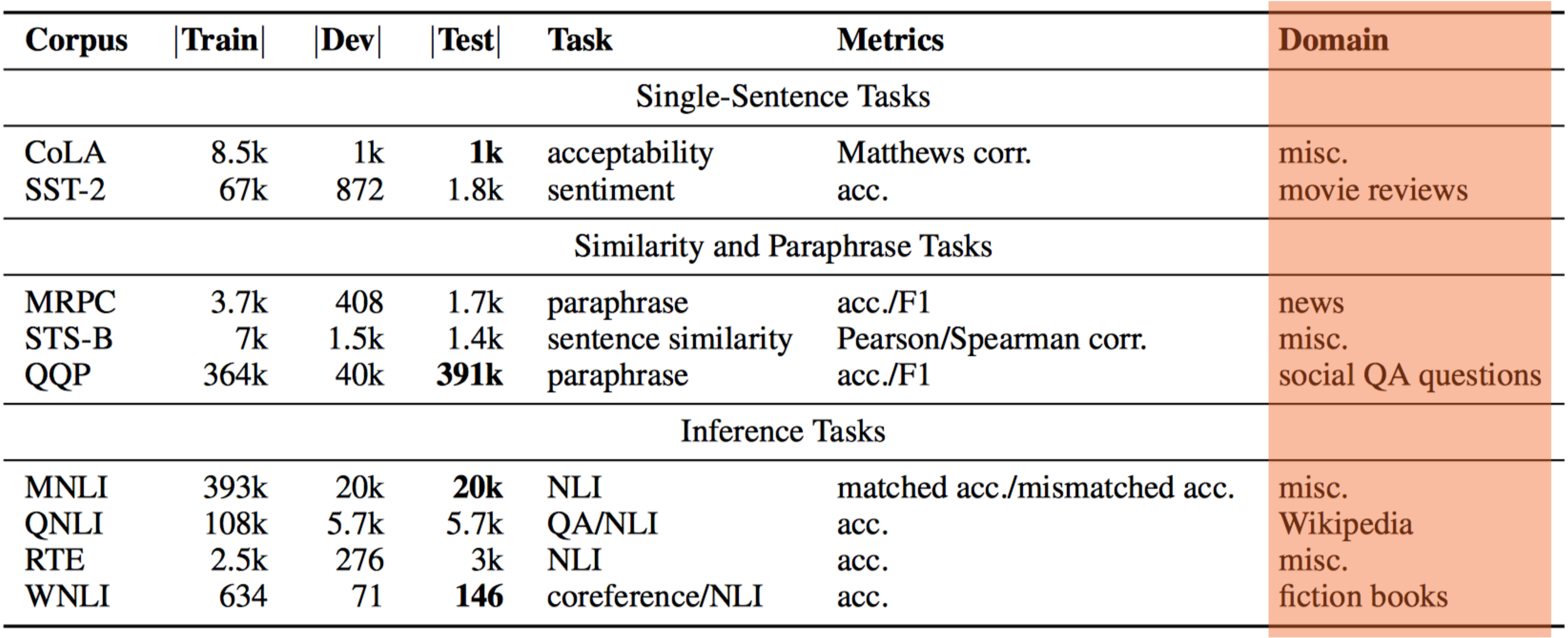
GLUE: The Main Tasks



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

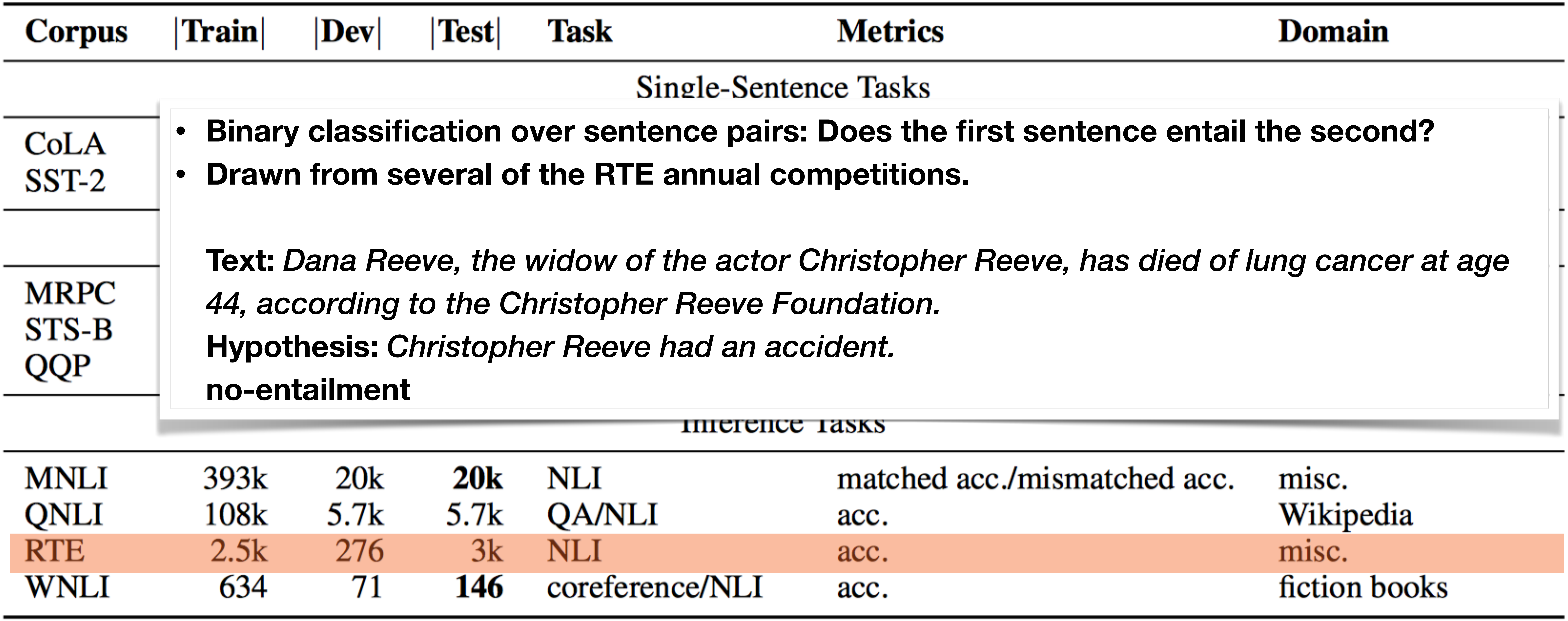
Wang, Singh, Michael, Hill, Levy & Bowman ICLR '19

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

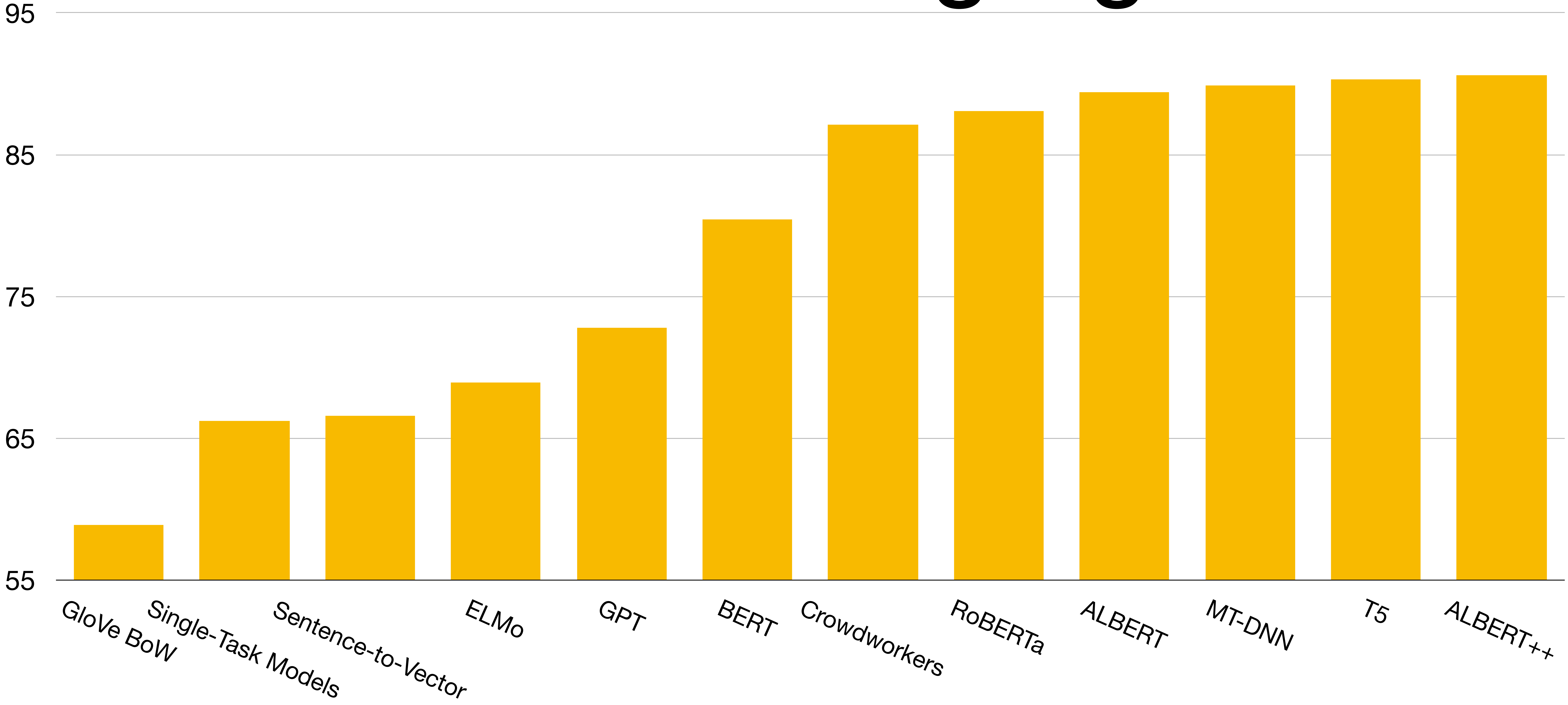


Corpus	|Train|	|Dev|	|Test|	Task	Metrics	Domain
				Single-Sentence Tasks		
CoLA						
SST-2						
MRPC						
STS-B						
QQP						
				Inference Tasks		
MNLI	393k	20k	**20k**	NLI	matched acc./mismatched acc.	misc.
QNLI	108k	5.7k	5.7k	QA/NLI	acc.	Wikipedia
RTE	2.5k	276	3k	NLI	acc.	misc.
WNLI	634	71	**146**	coreference/NLI	acc.	fiction books

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

GLUE: What methods work?

GLUE Score: Highlights

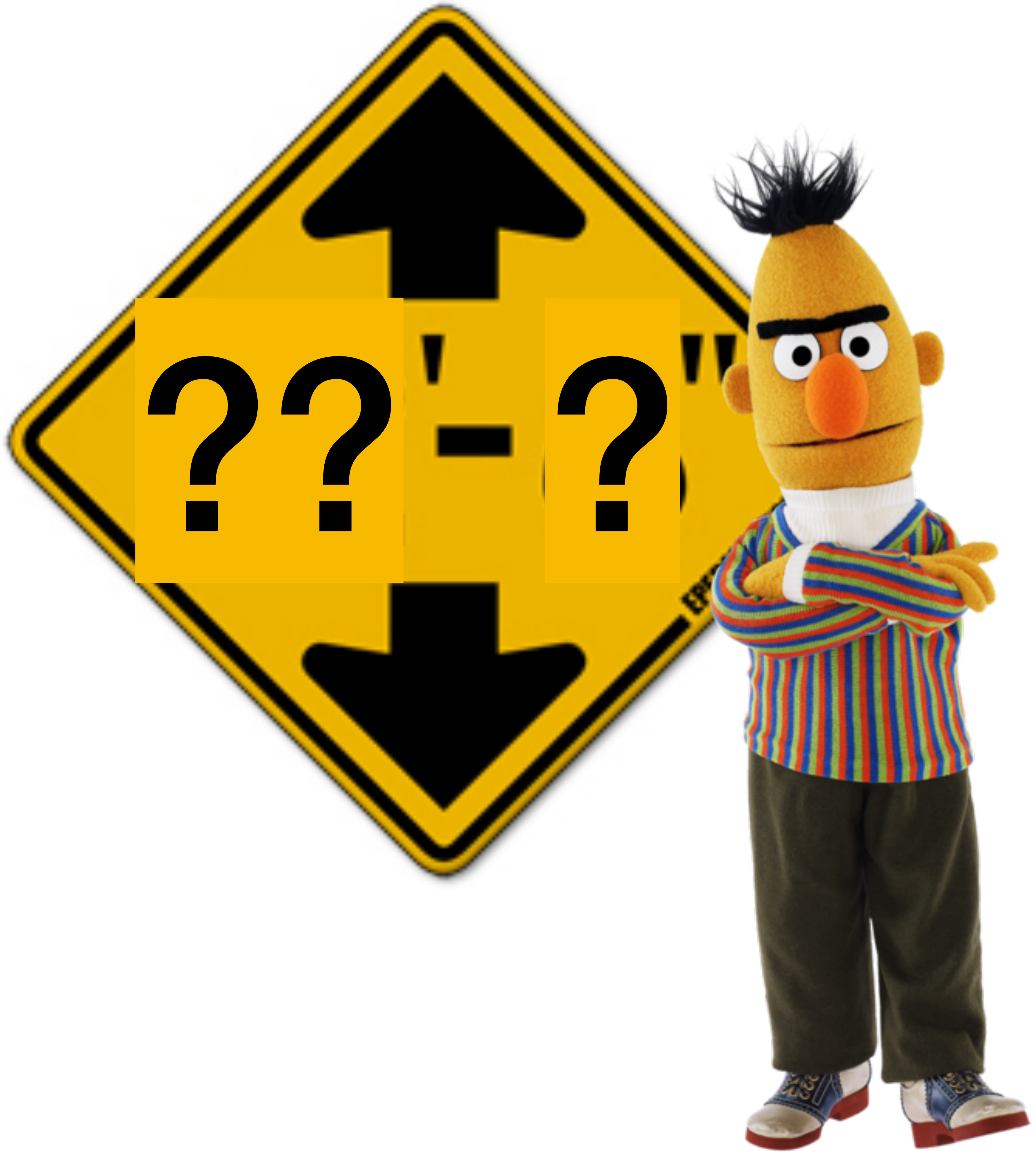


gluebenchmark.com

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

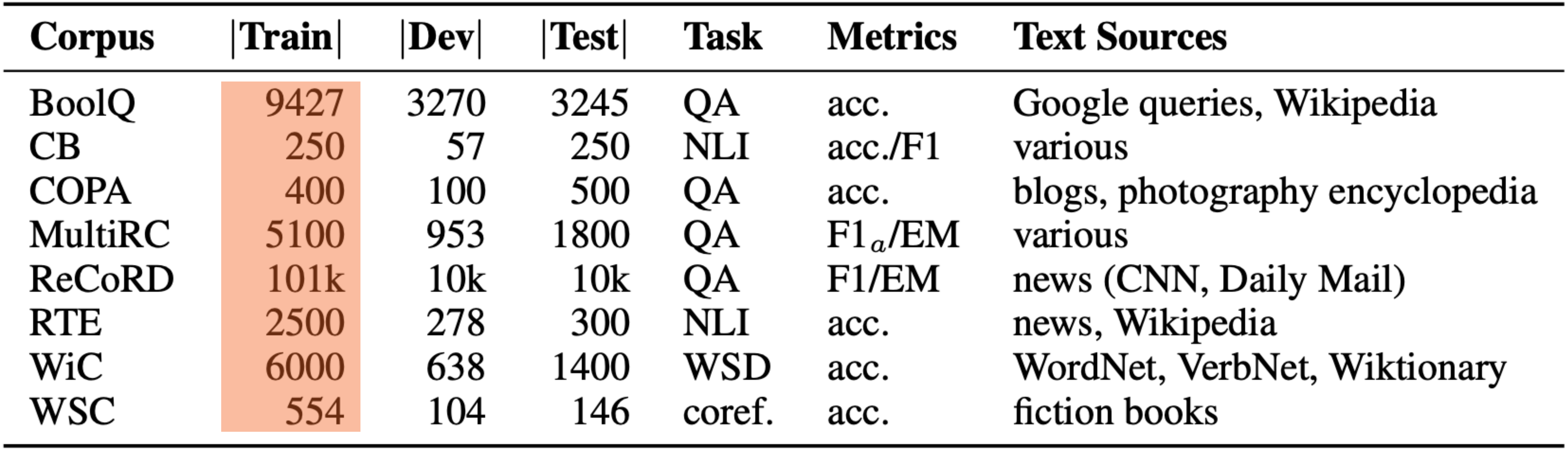
SuperGLUE

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.

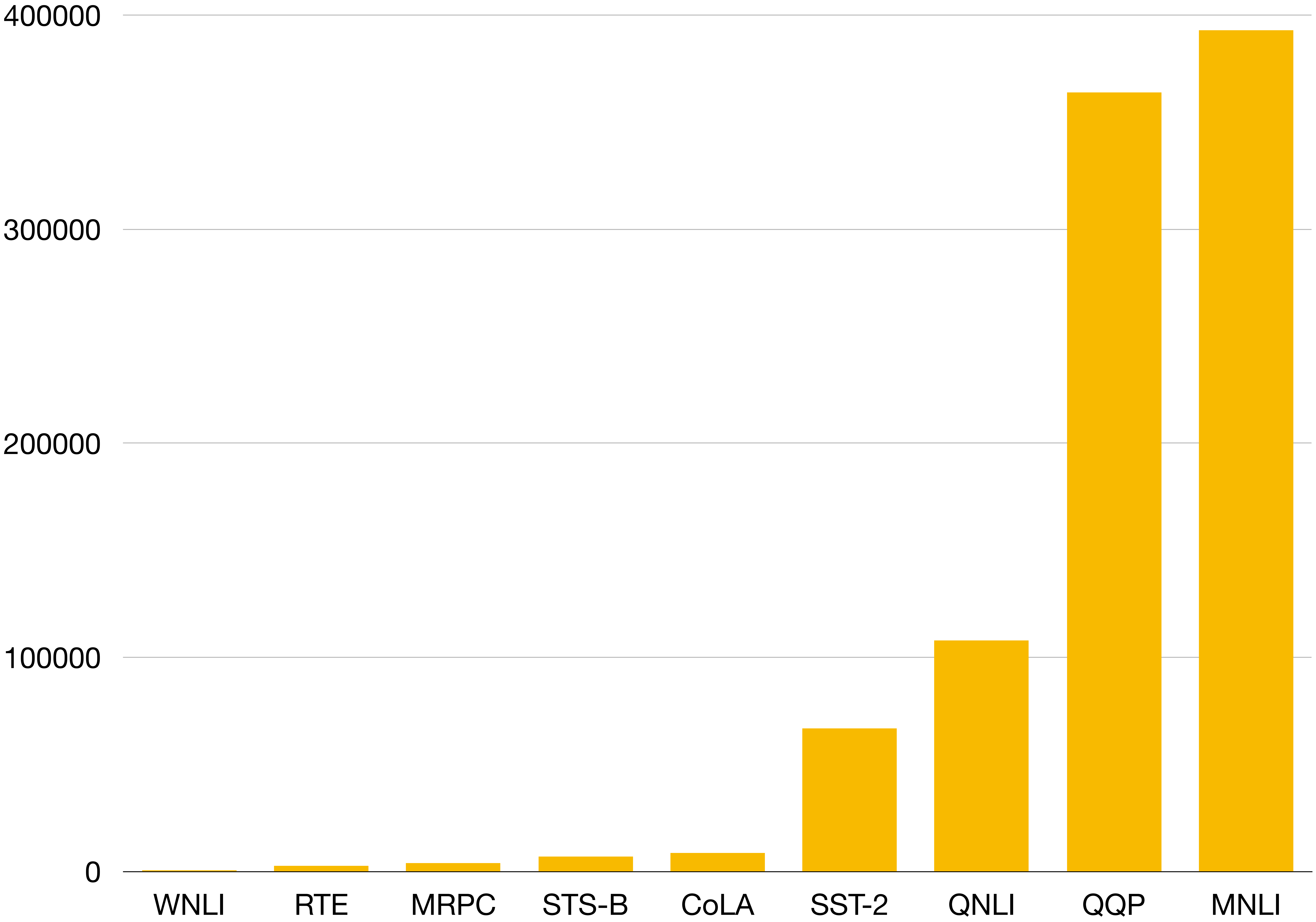
{Wang, Pruksachatkun, Nangia, Singh},
Michael, Hill, Levy & Bowman NeurIPS '19

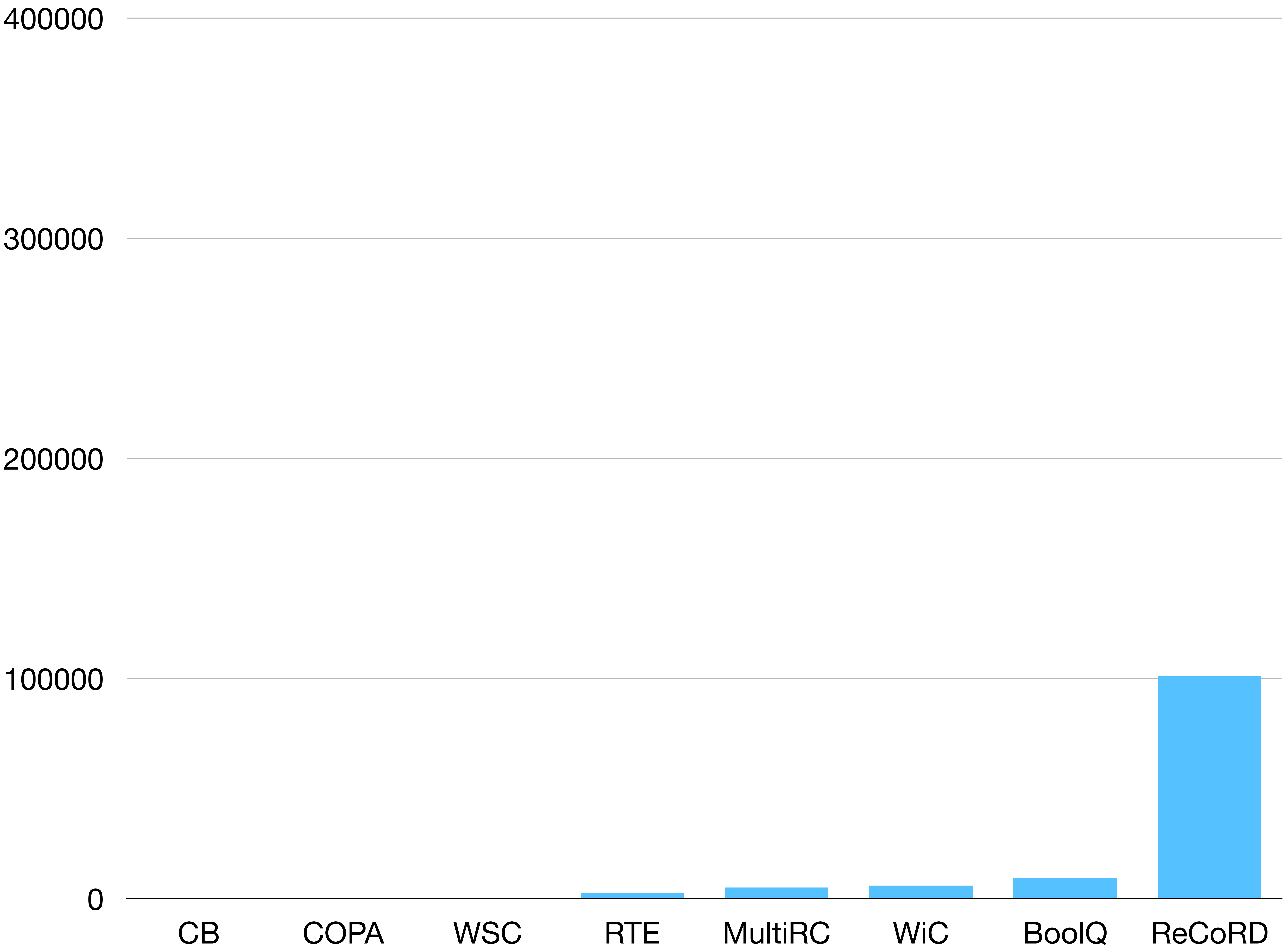
SuperGLUE: The Main Tasks



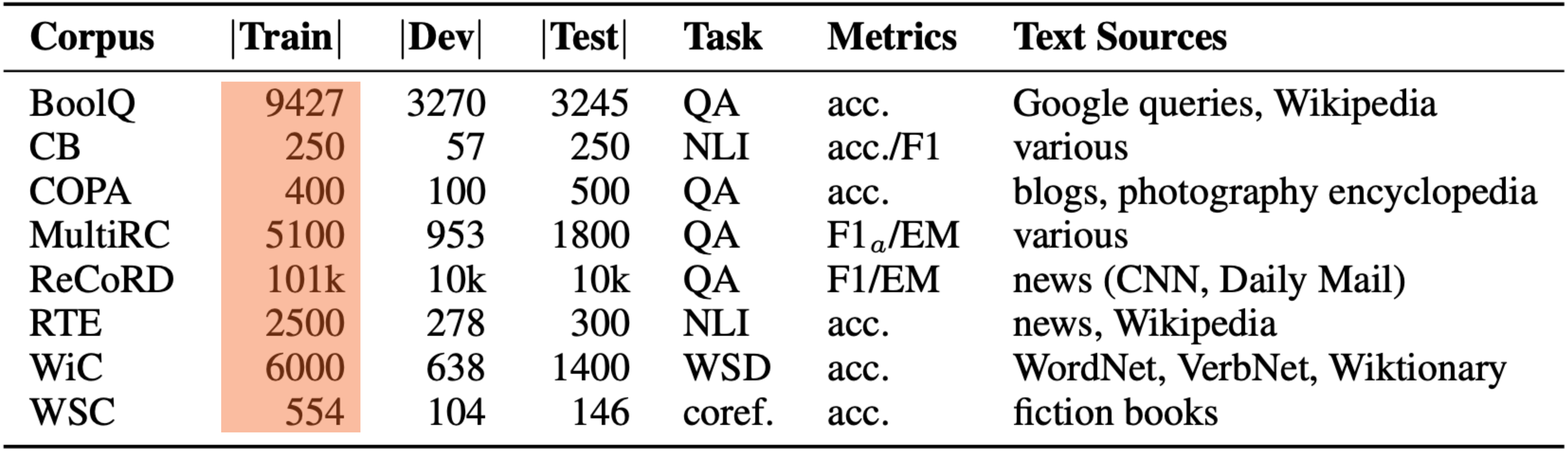
Corpus	|Train|	|Dev|	|Test|	Task	Metrics	Text Sources
BoolQ	9427	3270	3245	QA	acc.	Google queries, Wikipedia
CB	250	57	250	NLI	acc./F1	various
COPA	400	100	500	QA	acc.	blogs, photography encyclopedia
MultiRC	5100	953	1800	QA	F1_a/EM	various
ReCoRD	101k	10k	10k	QA	F1/EM	news (CNN, Daily Mail)
RTE	2500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


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

MultiRC

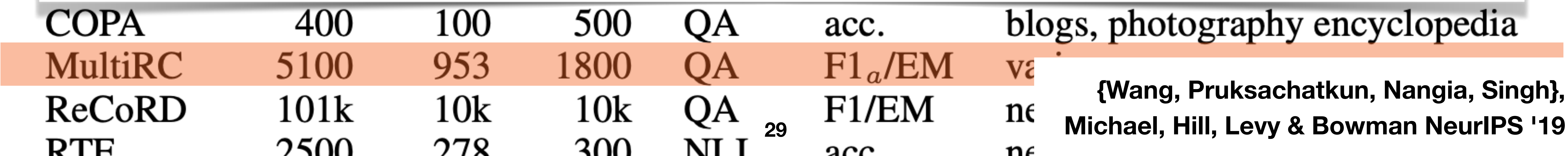
Khashabi et al. '18

- **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?*

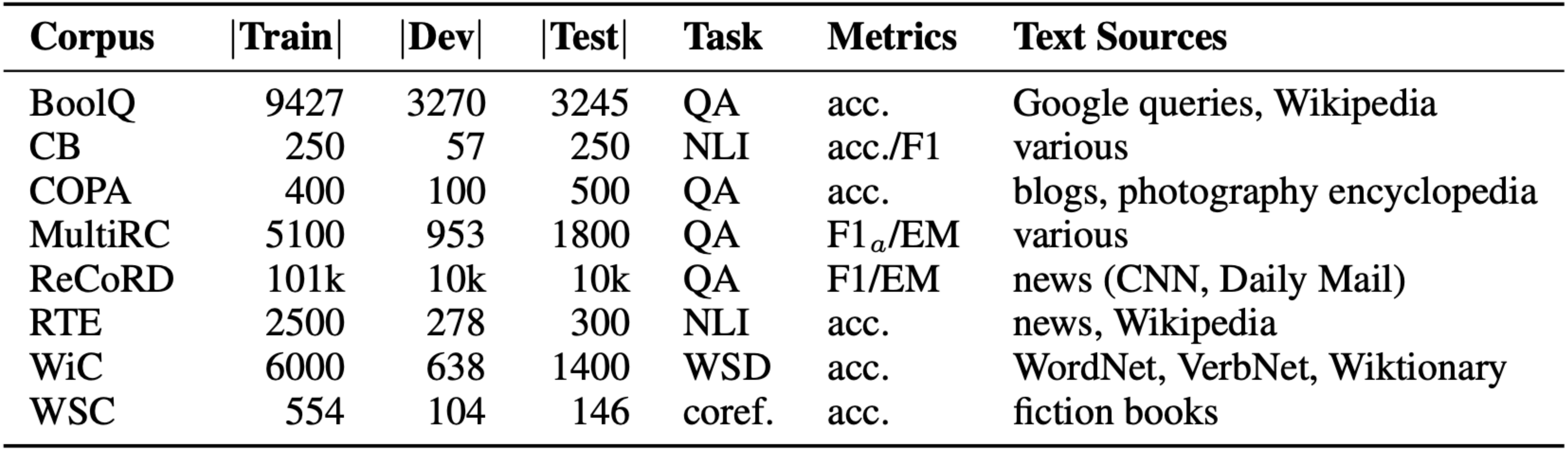
Answers: *Yes, she recovered (**T**), No (**F**), Yes (**T**), No, she didn't recover (**F**), Yes, she was at Susan's party (**T**)*



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

95
82.5
70
57.5
45
GloVe Bag of Words
BERT
RoBERTa
RoBERTa++
T5
Human Crowdworkers

super.gluebenchmark.com

⚠ 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

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 pizza with some friends.*
I ate olives.	*I ate some friends.*
<u>entailment</u>	<u>neutral</u>

GLUE and SuperGLUE: Open Issues

How sure are we that we've solved these NLU tasks *for IID test sets*?

Two relevant facts:

- Popular datasets for NLI, QA, etc. involve lots of phenomena that we know models aren't great at.

- Popular datasets for NLI, QA, etc. have relatively low inter-annotator agreement, and some instances are genuinely subjective. ML models are likely better than humans at predicting the modal human response. (see, e.g., Pavlick and Kwiatkowski)

Are subjectivity and low-agreement making ML models look artificially good?

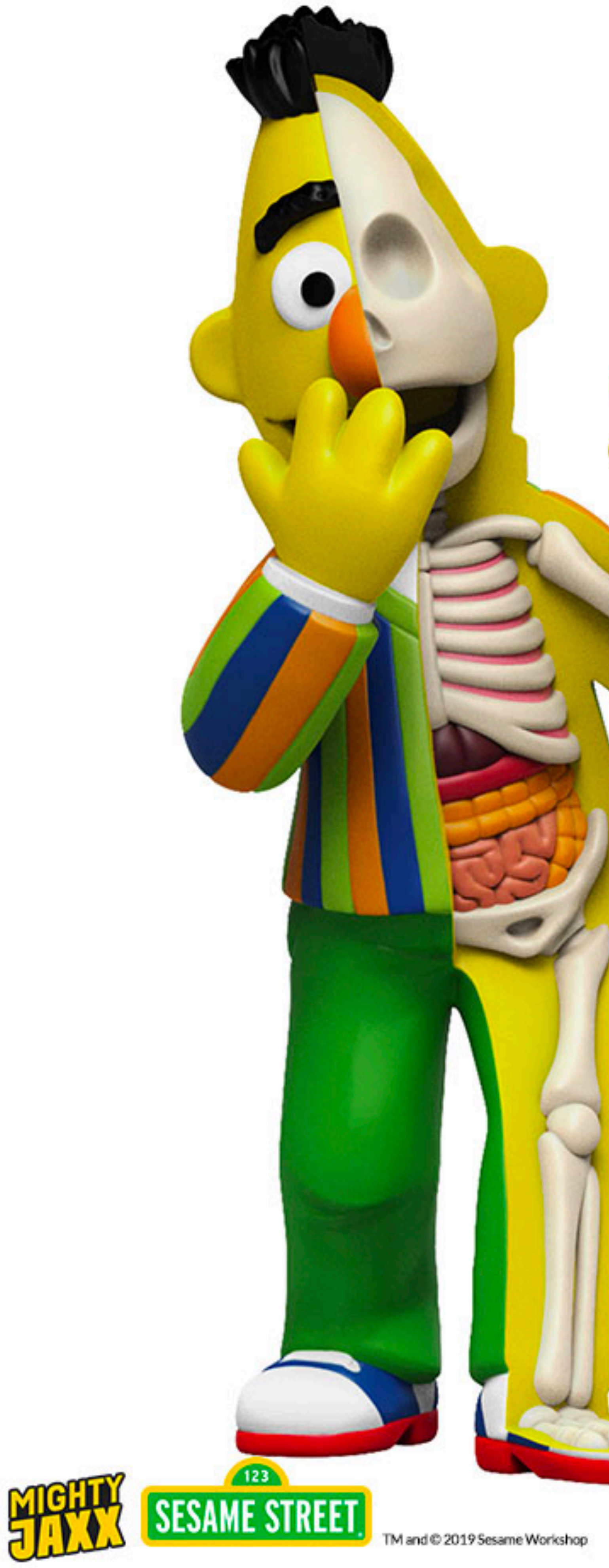
Why does BERT* work so well?
What does BERT know?

*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?

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?

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

Case study question: Does BERT know where *NPI* words like *any* can o

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

- Well-characterized in the linguistics literature.
- Based on complex long-distance dependencies with few local cues, so not trivial to learn.

*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

Do we train on in-domain data?

What performance metric do we use?

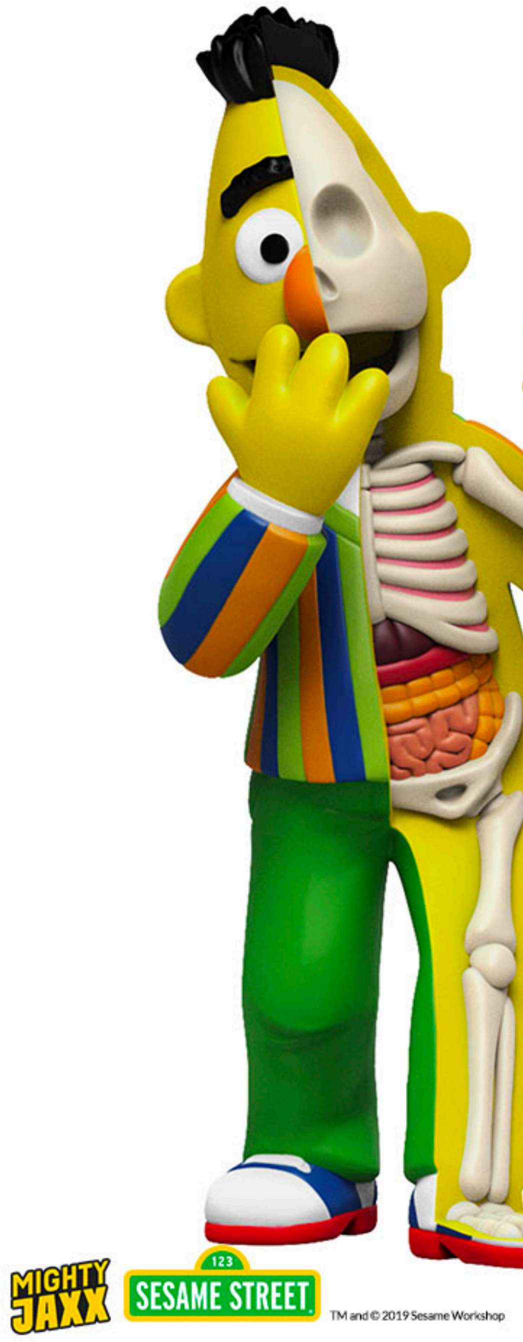
Do we use BERT's language modeling head at test time?

Do we fine-tune BERT when training the classifier?

*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



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

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

BERT does better than chance, but not especially well.

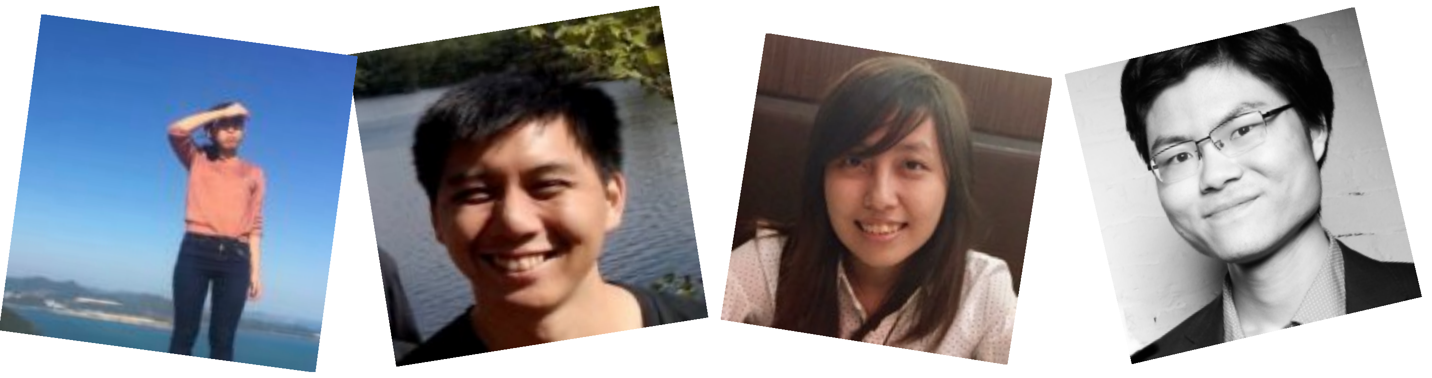
BERT does better than chance, but not especially well.

BERT has complete and perfect knowledge of NPI licensing.

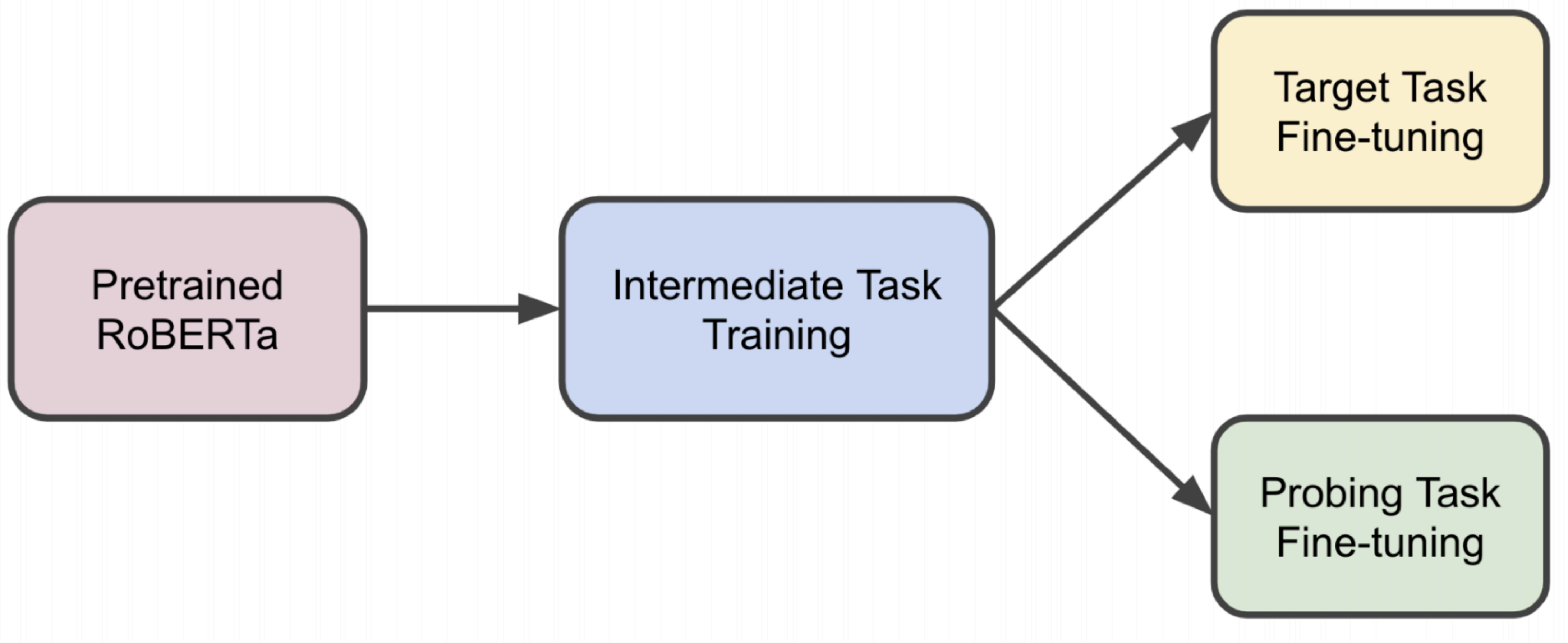
BERT does better than chance, but not especially well.

What's the Role of Multitask and Transfer Learning?

What's the Role of Multitask and Transfer Learning?



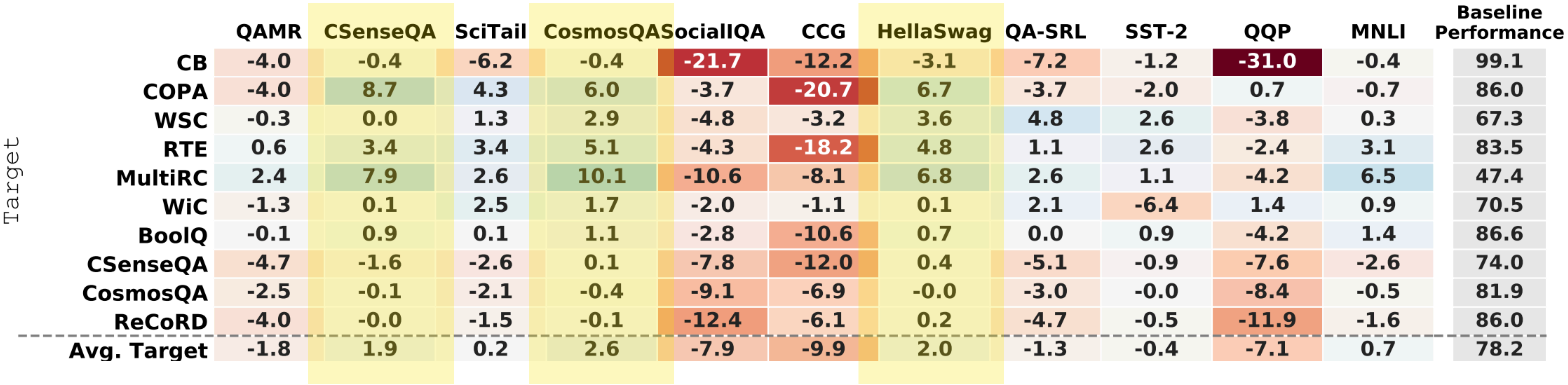
- Several of the strongest models on (Super)GLUE use some form of *intermediate-task training*:
    - Pretrain a model on unlabeled data
    - *Fine-tune it on a large labeled* intermediate *dataset*
    - Fine-tune it again on a smaller target labeled dataset
- What tasks work well as intermediate tasks?
- Can probing studies give us a clue as to why?



{Pruksachatkun, Phang, Liu, Htut}, Zhang, Pang, Vania, Kann & Bowman ACL '20

When does Intermediate-Task Transfer Learning Work?

RoBERTa with Intermediate-Task Training on...

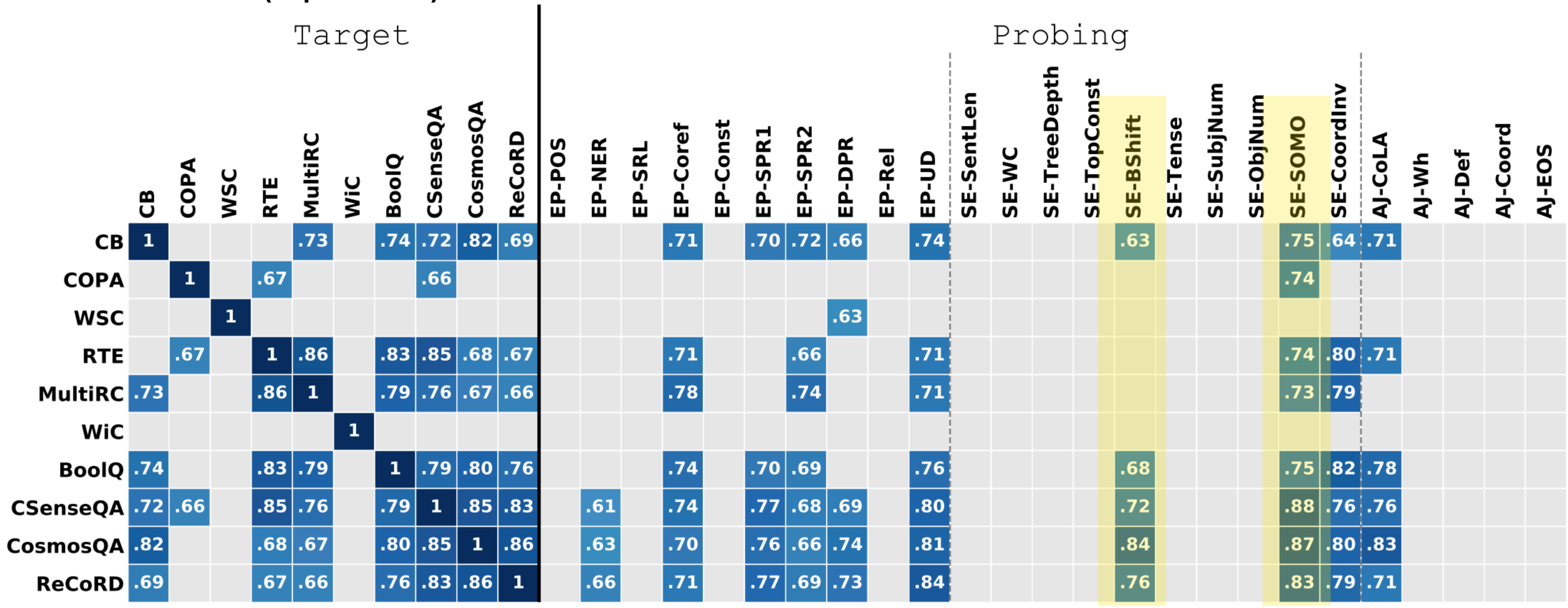


Target	QAMR	CSenseQA	SciTail	CosmosQA	SocialIQA	CCG	HellaSwag	QA-SRL	SST-2	QQP	MNLI	Baseline Performance
CB	-4.0	-0.4	-6.2	-0.4	-21.7	-12.2	-3.1	-7.2	-1.2	-31.0	-0.4	99.1
COPA	-4.0	8.7	4.3	6.0	-3.7	-20.7	6.7	-3.7	-2.0	0.7	-0.7	86.0
WSC	-0.3	0.0	1.3	2.9	-4.8	-3.2	3.6	4.8	2.6	-3.8	0.3	67.3
RTE	0.6	3.4	3.4	5.1	-4.3	-18.2	4.8	1.1	2.6	-2.4	3.1	83.5
MultiRC	2.4	7.9	2.6	10.1	-10.6	-8.1	6.8	2.6	1.1	-4.2	6.5	47.4
WiC	-1.3	0.1	2.5	1.7	-2.0	-1.1	0.1	2.1	-6.4	1.4	0.9	70.5
BoolQ	-0.1	0.9	0.1	1.1	-2.8	-10.6	0.7	0.0	0.9	-4.2	1.4	86.6
CSenseQA	-4.7	-1.6	-2.6	0.1	-7.8	-12.0	0.4	-5.1	-0.9	-7.6	-2.6	74.0
CosmosQA	-2.5	-0.1	-2.1	-0.4	-9.1	-6.9	-0.0	-3.0	-0.0	-8.4	-0.5	81.9
ReCoRD	-4.0	-0.0	-1.5	-0.1	-12.4	-6.1	0.2	-4.7	-0.5	-11.9	-1.6	86.0
Avg. Target	-1.8	1.9	0.2	2.6	-7.9	-9.9	2.0	-1.3	-0.4	-7.1	0.7	78.2

{Pruksachatkun, Phang, Liu, Htut},
Zhang, Pang, Vania, Kann & Bowman
ACL '20

What can Probing Tasks Tell us?

(SuperGLUE+)

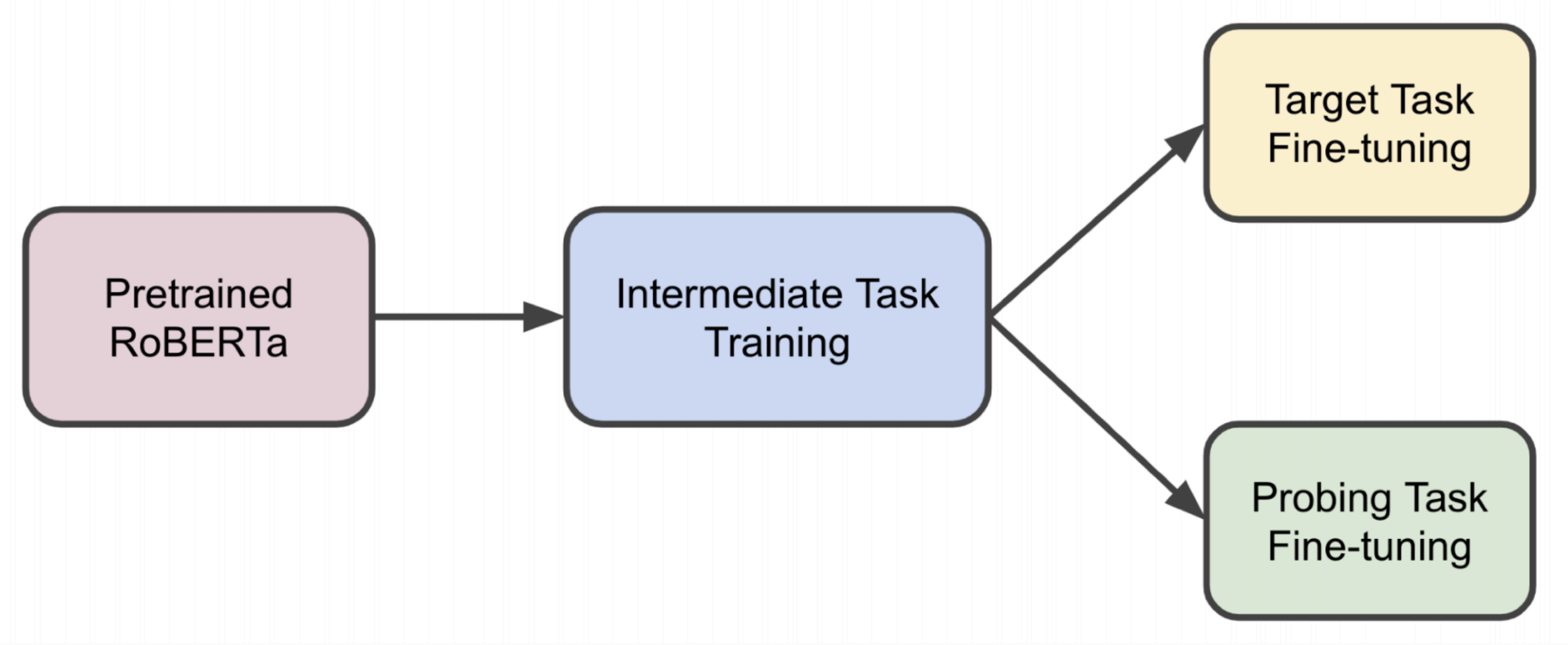


	Target										Probing																								
	CB	COPA	WSC	RTE	MultiRC	WiC	BoolQ	CSenseQA	CosmosQA	ReCoRD	EP-POS	EP-NER	EP-SRL	EP-Coref	EP-Const	EP-SPR1	EP-SPR2	EP-DPR	EP-Rel	EP-UD	SE-SentLen	SE-WC	SE-TreeDepth	SE-TopConst	SE-BShift	SE-Tense	SE-SubjNum	SE-ObjNum	SE-SOMO	SE-CoordInv	AJ-CoLA	AJ-Wh	AJ-Def	AJ-Coord	AJ-EOS
CB	1				.73		.74	.72	.82	.69				.71		.70	.72	.66		.74					.63				.75	.64	.71				
COPA		1		.67				.66																					.74						
WSC			1															.63																	
RTE		.67		1	.86		.83	.85	.68	.67				.71			.66			.71									.74	.80	.71				
MultiRC	.73			.86	1		.79	.76	.67	.66				.78			.74			.71									.73	.79					
WiC						1																													
BoolQ	.74			.83	.79		1	.79	.80	.76				.74		.70	.69			.76					.68				.75	.82	.78				
CSenseQA	.72	.66		.85	.76		.79	1	.85	.83		.61		.74		.77	.68	.69		.80					.72				.88	.76	.76				
CosmosQA	.82			.68	.67		.80	.85	1	.86		.63		.70		.76	.66	.74		.81					.84				.87	.80	.83				
ReCoRD	.69			.67	.66		.76	.83	.86	1		.66		.71		.77	.69	.73		.84					.76				.83	.79	.71				

{Pruksachatkun, Phang, Liu, Htut}, Zhang, Pang, Vania, Kann & Bowman ACL '20

Ongoing Work: Stay Tuned

- Since there are signs of *catastrophic forgetting*, does it help to mix pretraining updates in during intermediate-task training?
    - Tentatively: No. Why?
- How much do these results vary across different pretrained models?



Does this Work with *Crosslingual* Transfer?
(English intermediate and target training; Non-English evaluation)

		Target tasks									
		XNLI	PAWS-X	POS	NER	XQuAD	MLQA	TyDiQA	BUCC	Tatoeba	Avg.
	Metric	*acc.*	*acc.*	*F1*	*F1*	*F1 / EM*	*F1 / EM*	*F1 / EM*	*F1*	*acc.*	–
	# langs.	15	7	33	40	11	7	9	5	37	–
	XLM-R	80.1	86.5	75.7	62.8	76.1 / 60.0	70.1 / 51.5	75.7 / 61.0	71.5	31.0	67.2
No MLM	ANLI^{+}	- 0.8	+ 0.4	- 0.9	- 0.8	- 0.6 / - 0.1	- 0.6 / - 0.8	+ 2.2 / + 3.1	+20.1	+49.8	+ 7.7
	QQP	- 1.4	- 2.1	- 5.6	- 6.9	- 3.8 / - 3.8	- 3.9 / - 4.4	- 0.6 / - 0.2	+20.2	+51.7	+ 5.3
	SQuAD	- 1.4	+ 0.7	- 1.6	+ 0.2	+ 1.1 / + 1.3	+ 1.9 / + 2.5	+ 5.6 / + 7.4	+19.7	+46.9	+ 8.3
	HellaSwag	- 0.3	+ 0.8	- 0.7	- 1.0	- 0.3 / + 0.1	- 0.1 / + 0.2	+ 1.9 / + 1.3	+20.4	+49.9	+ 7.9
	CCG	- 2.6	- 3.4	- 1.5	- 0.7	- 1.5 / - 1.3	- 1.6 / - 1.5	+ 0.4 / + 0.7	+ 5.5	+38.9	+ 3.7
	CosmosQA	- 2.9	+ 1.5	- 1.2	- 0.9	+ 0.2 / + 0.3	+ 0.4 / + 0.5	+ 2.7 / + 3.8	+13.2	+28.8	+ 4.7
	CSQA	- 2.9	- 0.6	- 1.7	- 0.5	+ 0.2 / + 0.4	+ 1.6 / + 1.6	+ 3.0 / + 4.1	+11.3	+33.1	+ 4.9
	Multi-task	- 1.6	- 0.2	- 2.3	- 2.4	- 2.6 / - 3.1	- 1.4 / - 1.7	+ 1.9 / + 1.9	+18.4	+48.3	+ 6.4
		XTREME Benchmark Scores[†]									
	XLM-R (Hu et al., 2020)	79.2	86.4	72.6	**65.4**	76.6 / 60.8	71.6 / 53.2	65.1 / 45.0	66.0	57.3	68.1
	XLM-R (Ours)	79.5	86.2	74.0	62.6	76.1 / 60.0	70.2 / 51.2	75.5 / 61.0	64.5	31.0	66.1
	Our Best Models[‡]	**80.4**	**87.7**	**74.4**	63.4	**77.2 / 61.3**	**72.3 / 53.5**	**81.2 / 68.4**	**71.9**	**82.7**	**74.2**
	Human	92.8	97.5	97.0	-	91.2 / 82.3	91.2 / 82.3	90.1 / -	-	-	-

Phang, Htut, Pruksachatkun, Liu, Vania, Kann, Calixto & Bowman, arXiv 2020

Interim Conclusions

- Modern pre-trained transformers, especially with intermediate-task training, outperform non-expert humans on nearly all established NLU evaluation tasks.

- These models still fail frequently, sometimes in bizarre ways, and we're only just starting to understand why they work.

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...
- ...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%.
- Aim for *100k+* test examples, so we can still productively compare models with *near*-perfect accuracy.
- 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?

See cited papers for full project details. This presentation does not reflect the position of the sponsoring organizations.

Thanks!

SCHMIDT **FUTURES**

SAMSUNG

NSF

intuit®

NVIDIA®

NYU

ML² Machine Learning for Language

Sam Bowman

@sleepinyourhat