How do we fix natural language understanding evaluation?

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.

This Talk

- What we learned from running the GLUE and SuperGLUE benchmarks GLUE language understanding benchmark
  Wang et al. '19a, Nangia & Bowman '19, Wang et al. '19b

- What's next for evaluation?
  Idle speculation '20

GLUE and SuperGLUE

GLUE

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

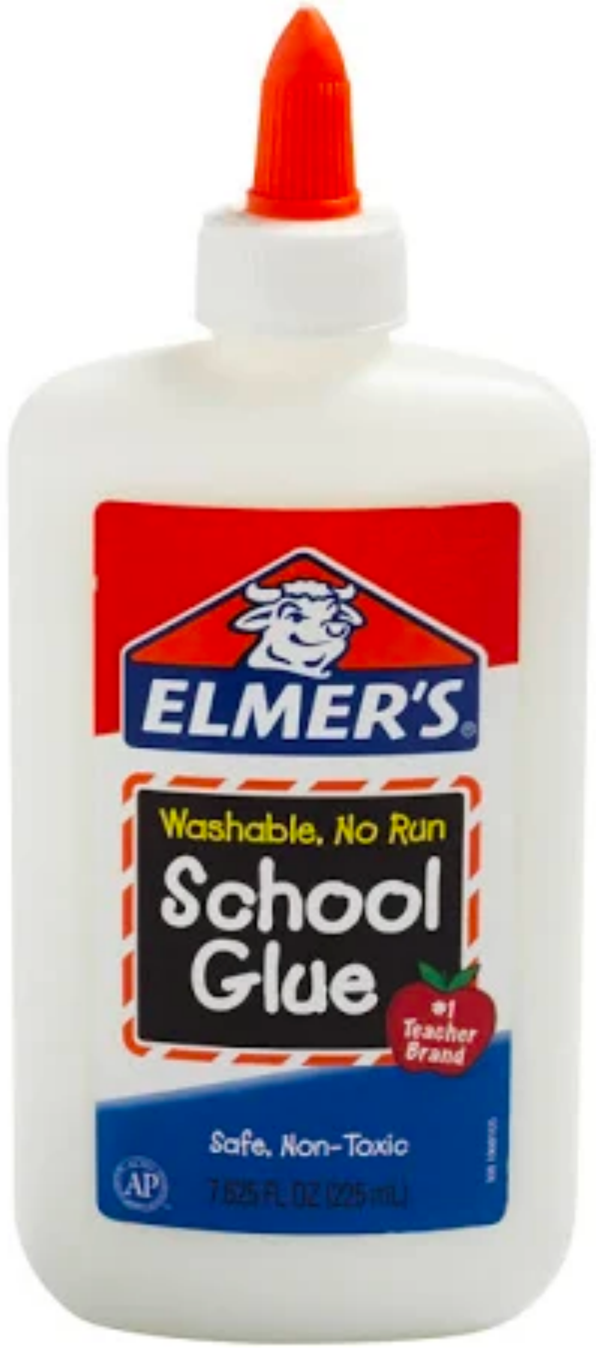
Nine English-language sentence understanding tasks based on existing data, with simple task APIs.

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?

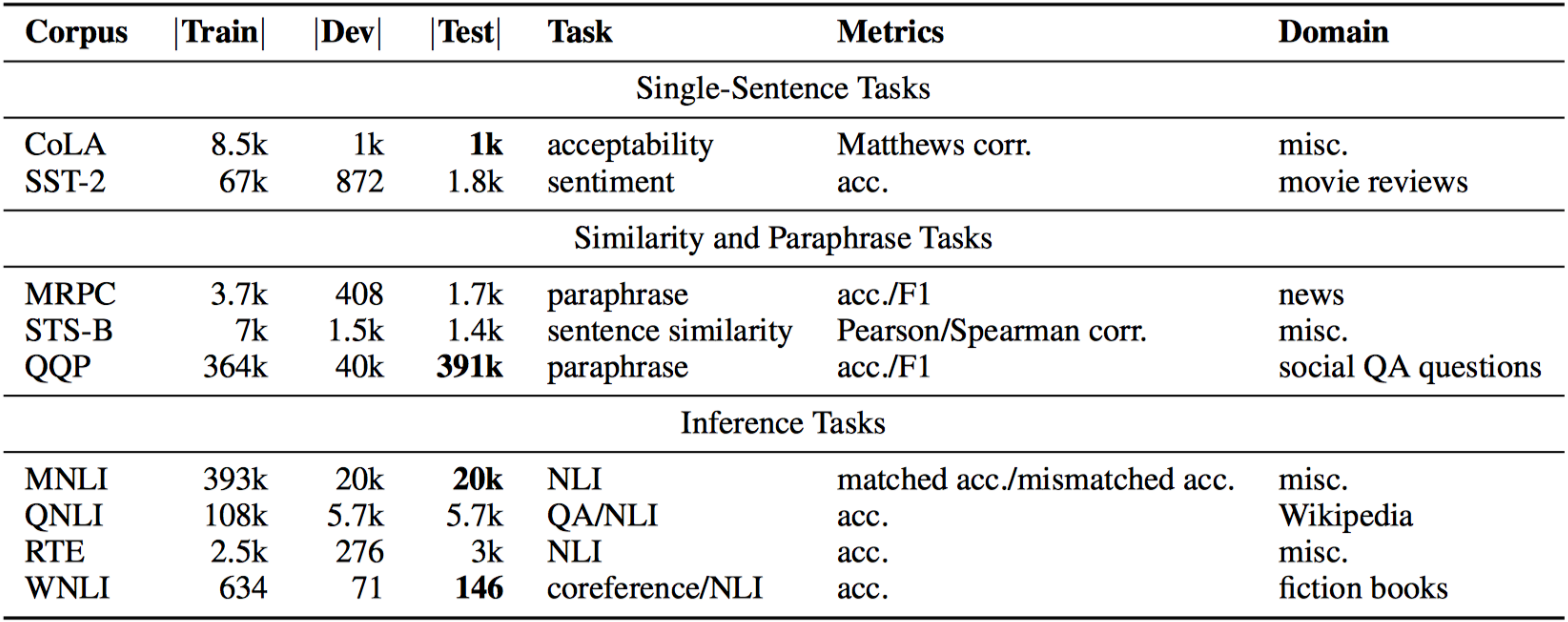
GLUE for non-NLP-specialist researchers:

- We provide a *single-number metric* that summarizes performance on problems of interest to researchers in NLU, alongside data, baseline results, and code.

- We don't enforce any particular experimental design —that's up to the (expert) users.

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

GLUE: The Main Tasks

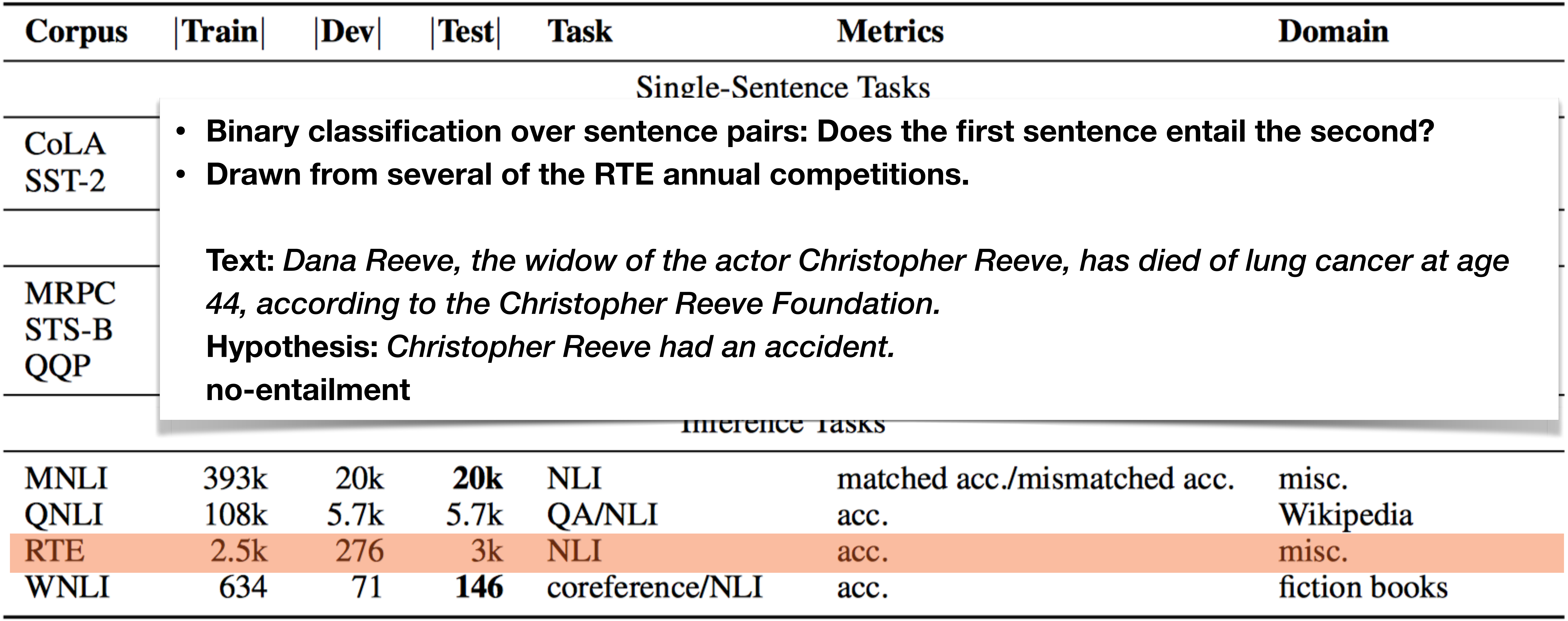


Corpus	|Train|	|Dev|	|Test|	Task	Metrics	Domain
Single-Sentence Tasks						
CoLA	8.5k	1k	**1k**	acceptability	Matthews corr.	misc.
SST-2	67k	872	1.8k	sentiment	acc.	movie reviews
Similarity and Paraphrase Tasks						
MRPC	3.7k	408	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.5k	1.4k	sentence similarity	Pearson/Spearman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



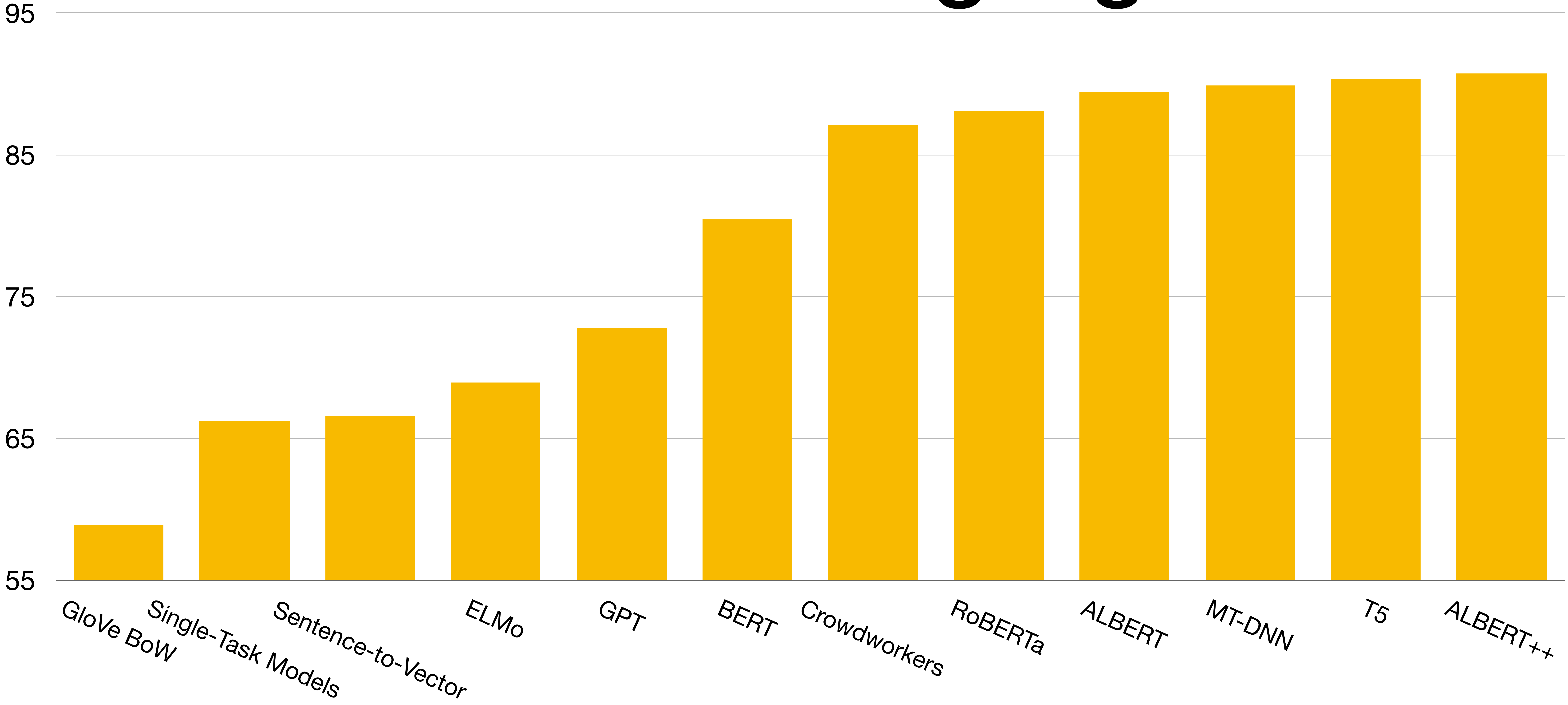
Corpus	|Train|	|Dev|	|Test|	Task	Metrics	Domain
				Single-Sentence Tasks		
CoLA						
SST-2						
MRPC						
STS-B						
QQP						
				Inference Tasks		
MNLI	393k	20k	**20k**	NLI	matched acc./mismatched acc.	misc.
QNLI	108k	5.7k	5.7k	QA/NLI	acc.	Wikipedia
RTE	2.5k	276	3k	NLI	acc.	misc.
WNLI	634	71	**146**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

- **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

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

GLUE Score: Highlights



gluebenchmark.com

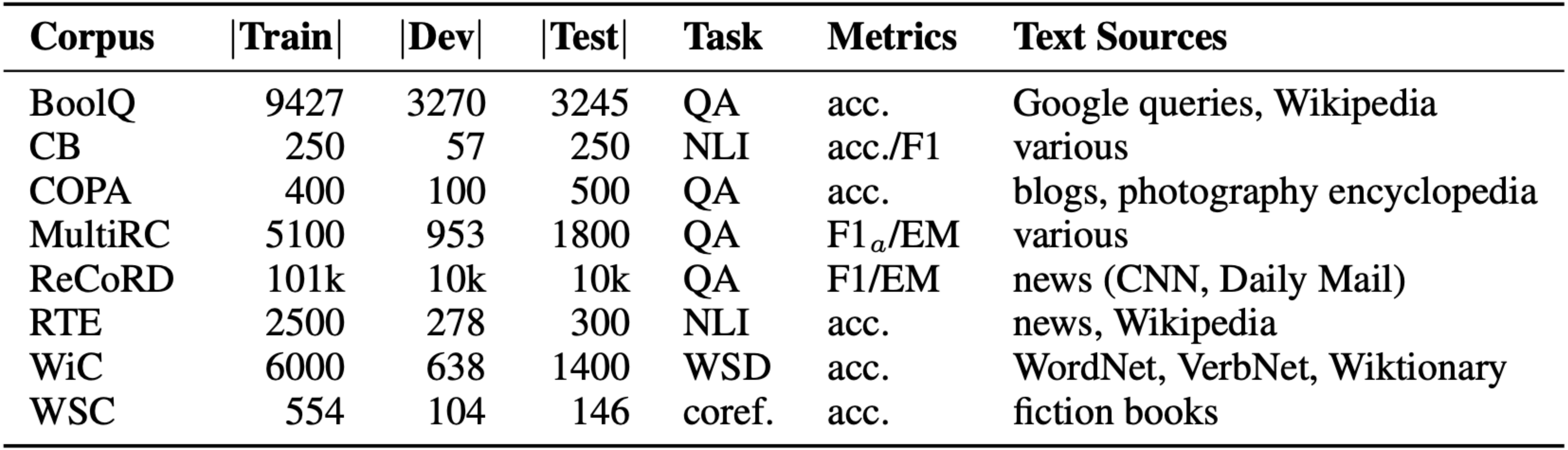
SuperGLUE

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
- …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

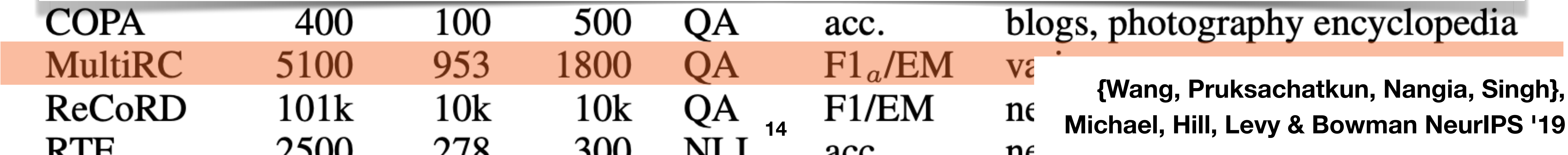
{Wang, Pruksachatkun, Nangia, Singh},
Michael, Hill, Levy & Bowman NeurIPS '19

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 Score: Highlights

95
82.5
70
57.5
45
GloVe Bag of Words
BERT
RoBERTa
NEZHA
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 likely than humans to act on irrelevant gender information.
- Mitigating these biases is a major open problem.

⚠ GLUE and SuperGLUE: Open Issues

The numbers are high, but we clearly h

10-point gap between humans and T5!

(See also: 100s of recent BERTology papers.)

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.*
entailment	neutral

GLUE and SuperGLUE: Open Issues

Our best systems aren't that robust to out-of-domain examples, so we haven't solved these *tasks.*

How sure are we that we've solved these *datasets*?

- GLUE-style datasets have relatively low inter-annotator agreement, and many instances are genuinely debatable. (see, e.g., Pavlick and Kwiatkowski.)
  - Does *I ate a burrito* entail *I ate a sandwich*?
- ML models are likely better than humans at predicting the *modal human response*.

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

Where are we now?

- Modern pre-trained transformers outperform non-expert humans on nearly all simple-output-space NLU tasks.
- These models still fail frequently, sometimes in bizarre ways.
- This makes for an unhealthy evaluation ecosystem, and lost opportunities for progress:
  - Generalist researchers are disincentivized to work on NLU.
  - NLU specialist researchers are left to choose methods without a clear empirical basis.

So what's next?

Evaluation: What’s Next?

It 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?

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

- Good few-shot performance is probably useful, and it can be *sufficient* evidence of language understanding...
- …but adding artificial constraints in evaluation will yield missed opportunities.

Building a Better Evaluation Dataset

Instead, let's figure out how to build better evaluation datasets.

Building a Better Evaluation Dataset

What we want:

- A very high human performance ceiling. 99%+ accuracy?
- A very large size. 100,000+ items?
- Representative coverage of a maximally broad distribution of language-related phenomena.
- No incentive to build biased models.

Building a Better Evaluation Dataset

What we want:

- **A very high human performance ceiling. 99%+ accuracy?**
- A very large size. 100,000+ items?
- Representative coverage of a maximally broad distribution of language-related phenomena.
- No incentive to build biased models.

99%+ Human Performance?

- We'd like to be able to measure improvements against extremely strong baselines:
    - If you believe existing benchmarks like SuperGLUE, we are in the ballpark of human performance, and improving quickly.
- High agreement avoids the risk of spurious 'superhuman performance' results due to legitimate disagreements.

Building a Better Evaluation Dataset

What we want:

- A very high human performance ceiling. 99%+ accuracy?
- **A very large size. 100,000+ items?**
- Representative coverage of a maximally broad distribution of language-related phenomena.
- No incentive to build biased models.

100k+ Examples?

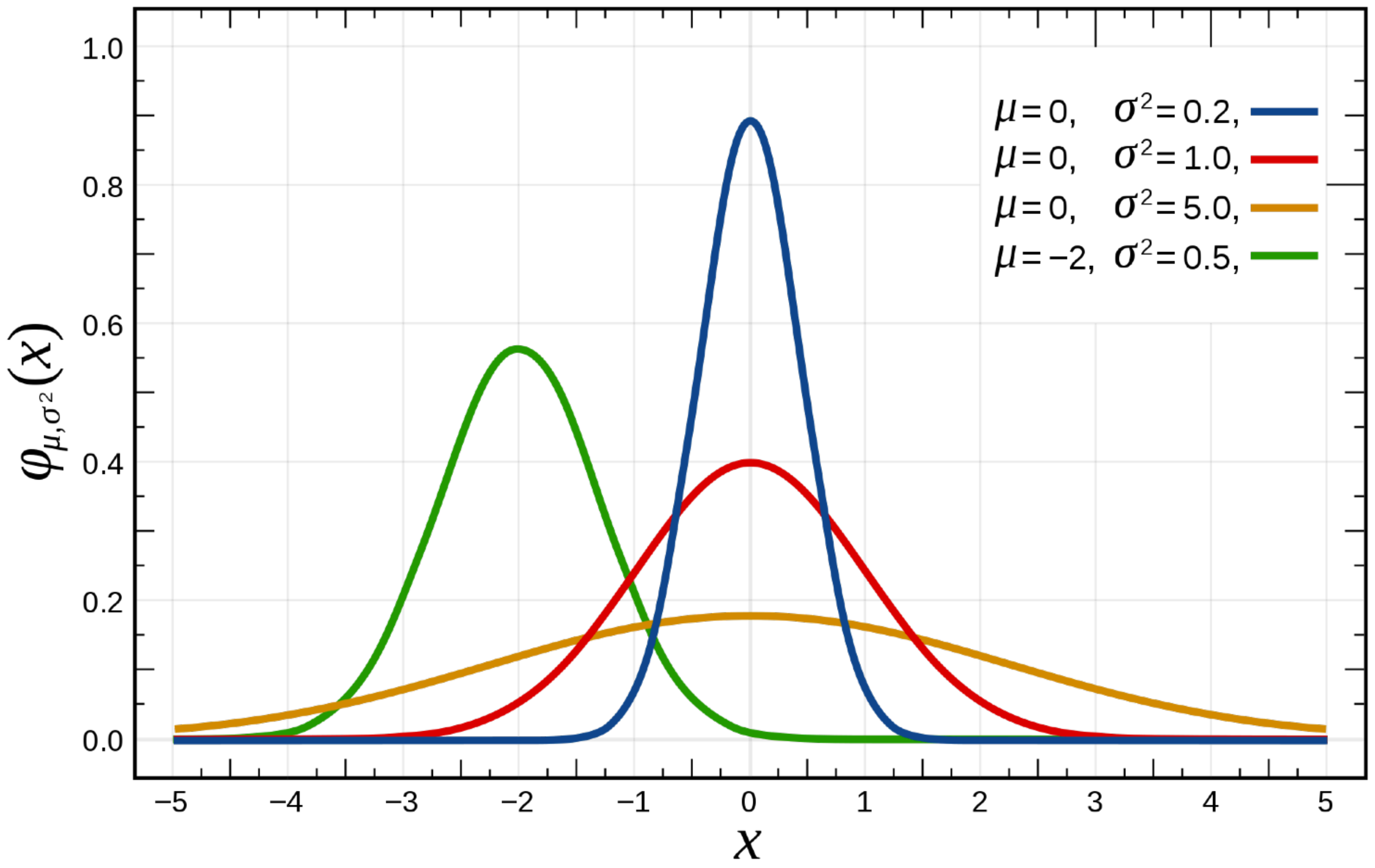
- We should expect to start spending more time in the long tail:
  - 98.7% => 98.9% is the new 70% => 74%

Building a Better Evaluation Dataset

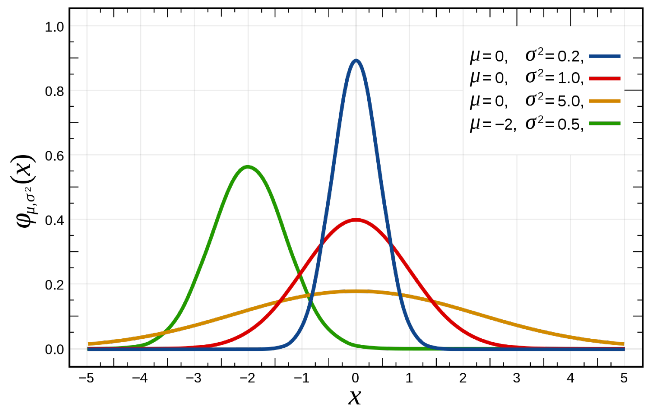
What we want:

- A very high human performance ceiling. 99%+ accuracy?
- A very large size. 100,000+ items?
- **Representative coverage of a maximally broad distribution of language-related phenomena.**
- No incentive to build biased models.

Representative Coverage of a Broad Distribution



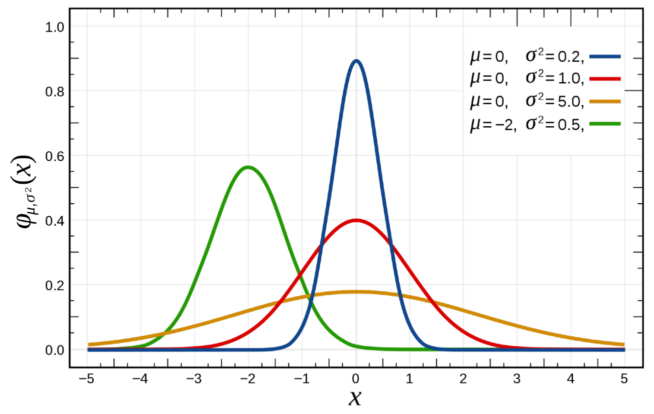
What is P(X)?



Representative Coverage of a Broad Distribution

There is no satisfying *natural distribution* of NLU test examples:

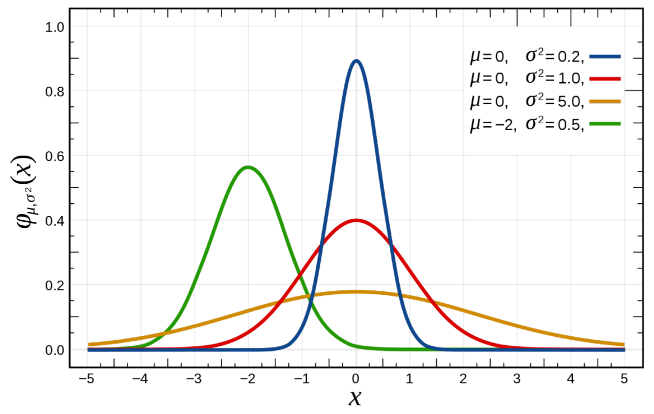
- Tasks built on random samples from the web/books/etc. don't tend to be compelling benchmarks *for NLU*.
  - Language modeling as an example: Mostly tests specific knowledge. Cases that isolate understanding are rare (see LAMBADA).
- Applied tasks like open question answering *can* draw on user data, but this usually introduces undesirable assumptions.
  - Natural Questions draws on Google queries, but these are heavily influenced by what users expect existing systems to understand.
- Many multiple-input tasks like NLI and RC have no obvious preexisting source distribution.



Representative Coverage of a Broad Distribution

What if we let experts write all the examples?

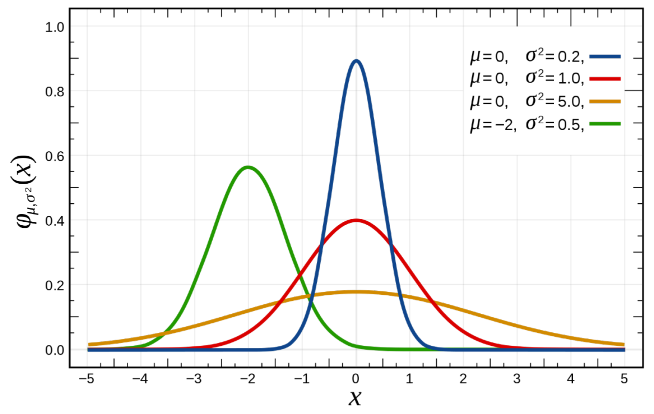
- Example: FraCaS for textual entailment
- Pitfall: Intentionally or unintentionally, we'll focus the evaluation almost exclusively on phenomena that we know how to characterize.
- We shouldn't expect good results on such a dataset to translate to any other evaluation dataset.



Representative Coverage of a Broad Distribution

What if we let crowdworkers write all the examples? (Based on some seed text.)

- Example: MNLI for textual entailment, SQuAD for QA
- This gets us a broader distribution, but...
- Pitfall: annotation artifacts

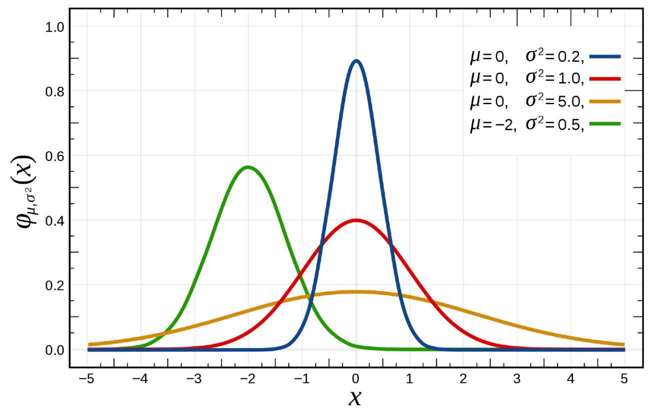


Representative Coverage of a Broad Distribution

What if we let an ML model pick the distribution?

- Example: Adversarial NLI and DynaBench for textual entailment, HellaSWAG for language modeling

- Pitfall: Incentive to create new models with *different* error patterns from current models, even when this doesn't improve overall performance. This likely slows progress.



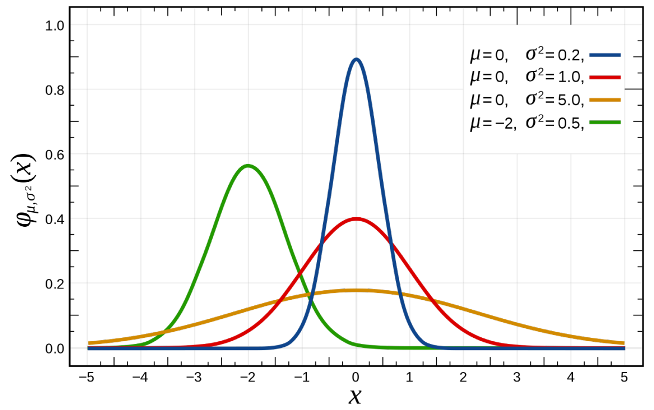
Representative Coverage of a Broad Distribution

What if we let crowdworkers pick the distribution, *with limited help from experts*?

- Example: OCNLI for textual entailment, ORB for QA, Gardner et al. contrast sets for many other tasks

- Pitfall: Experts can still steer research toward well understood topics...

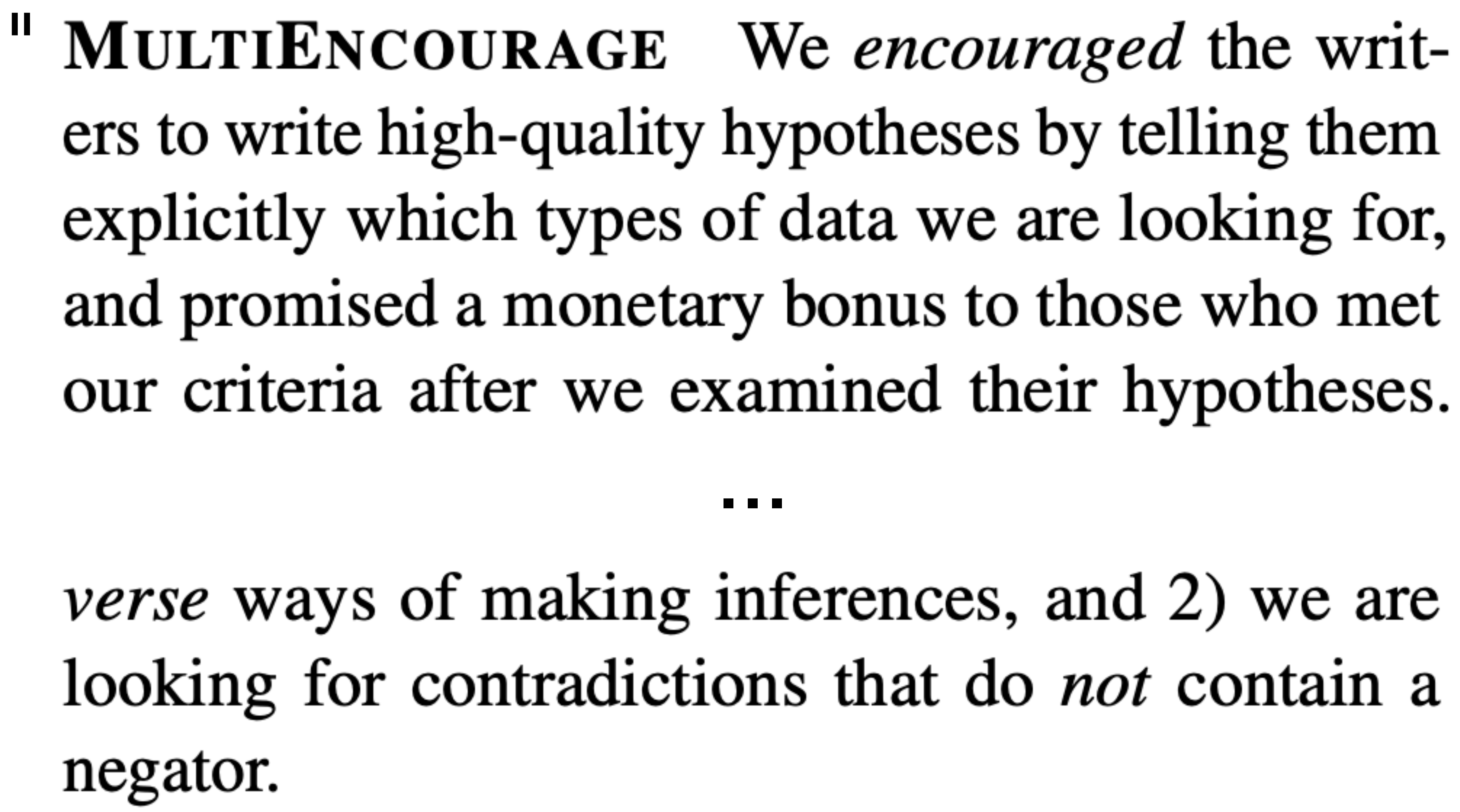
- ...but with the right constraints, this may be able to mitigate annotation artifacts without losing the diversity that comes from crowdsourced data.



Representative Coverage of a Broad Distribution

What if we let crowdworkers pick the distribution, *with limited help from experts*?

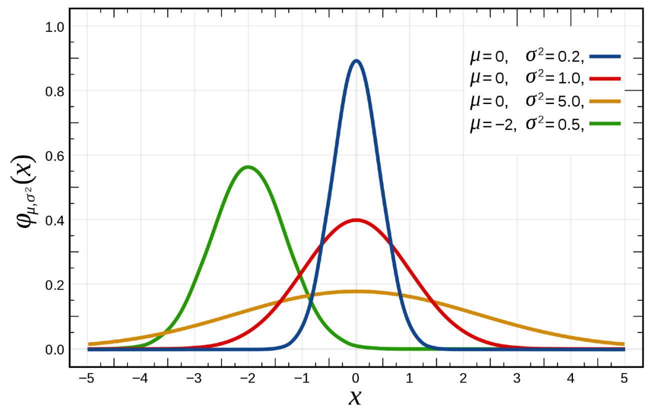
- Example: OCNLI for textual entailment



" **MULTIENCOURAGE** We *encouraged* the writers to write high-quality hypotheses by telling them explicitly which types of data we are looking for, and promised a monetary bonus to those who met our criteria after we examined their hypotheses.

...

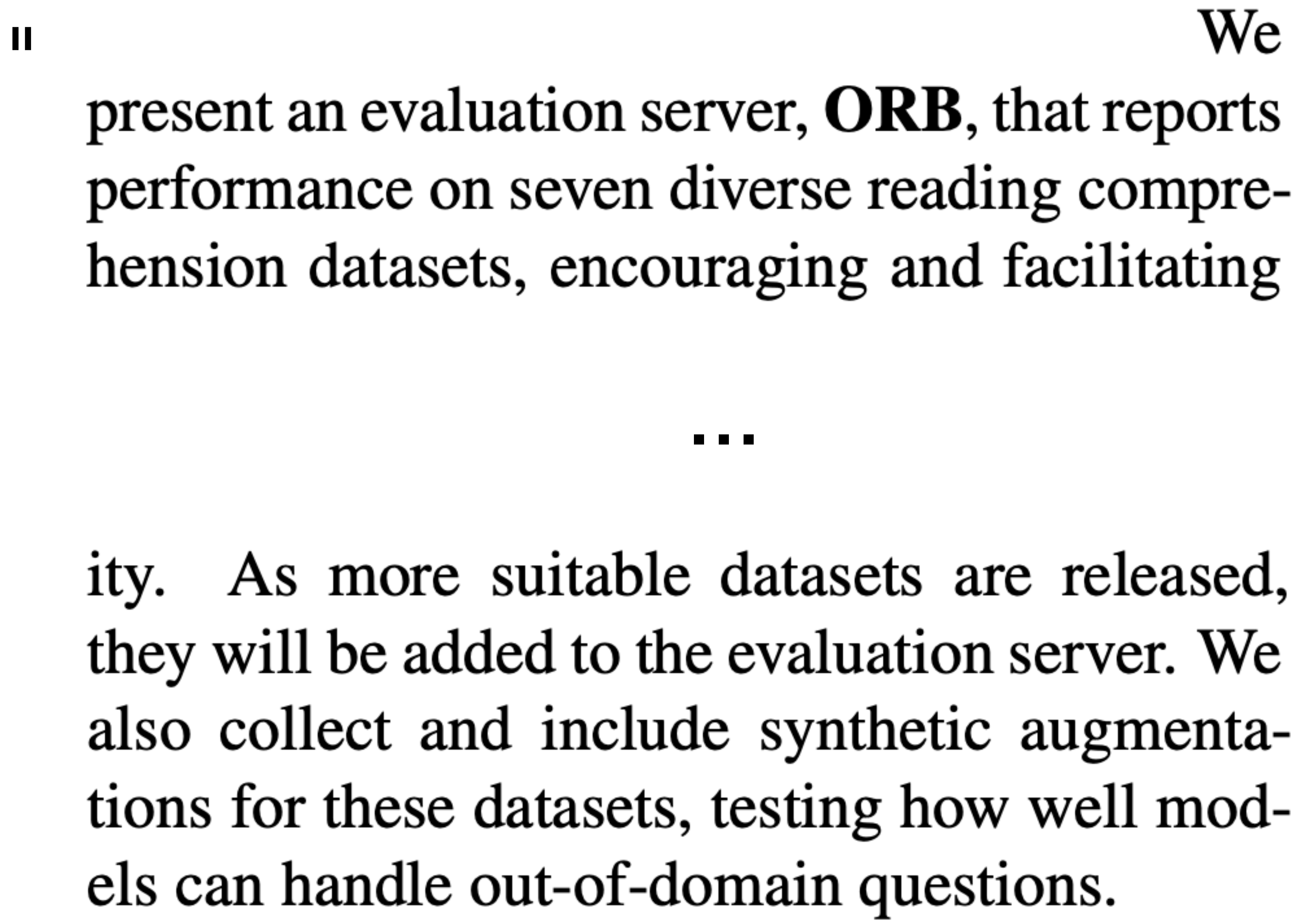
verse ways of making inferences, and 2) we are looking for contradictions that do *not* contain a negator.



Representative Coverage of a Broad Distribution

What if we let crowdworkers pick the distribution, *with limited help from experts*?

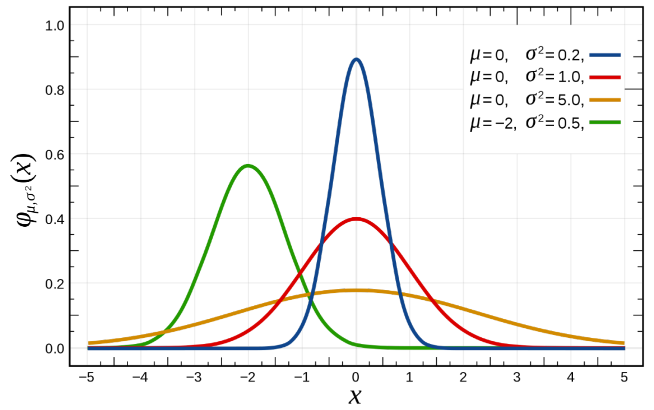
- Example: ORB for QA



" We present an evaluation server, **ORB**, that reports performance on seven diverse reading comprehension datasets, encouraging and facilitating

...

ity. As more suitable datasets are released, they will be added to the evaluation server. We also collect and include synthetic augmentations for these datasets, testing how well models can handle out-of-domain questions.



Representative Coverage of a Broad Distribution

What if we let crowdworkers pick the distribution, *with limited help from experts*?

- Example: Gardner et al. contrast sets for many other tasks
  - Kaushik et al.: Add evidence of the causal factors behind labels by *minimally editing* crowdsourced datapoints such that their labels change, add those edited examples to the dataset.

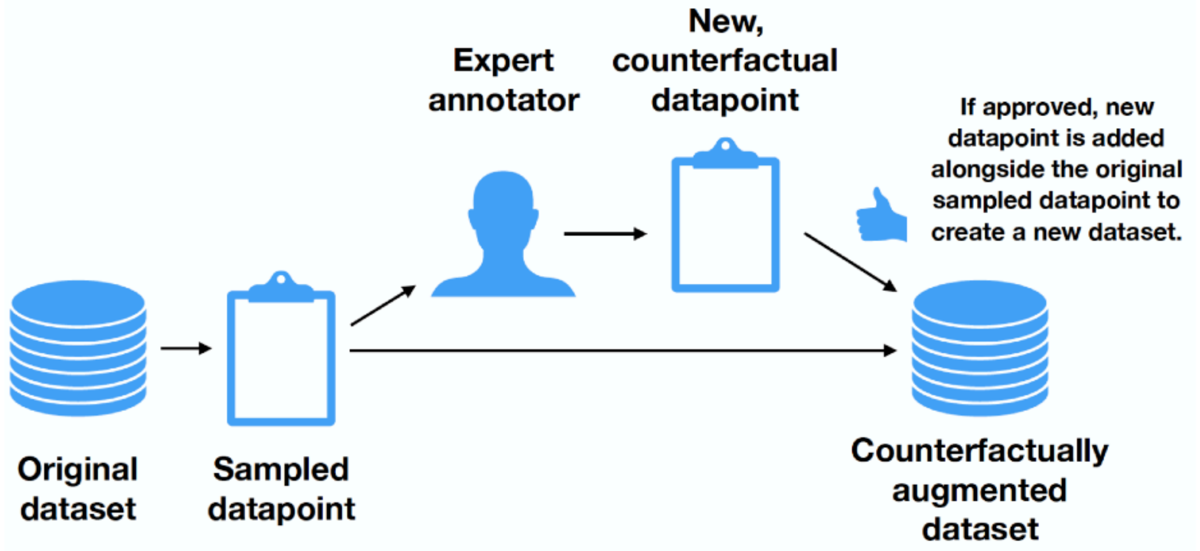

Original dataset
Sampled datapoint
Expert annotator
New, counterfactual datapoint
If approved, new datapoint is added alongside the original sampled datapoint to create a new dataset.
Counterfactually augmented dataset

Figure 1: Pipeline for collecting and leveraging counterfactually-altered data

  - Caveat: Initial attempts at this used crowdworkers, who didn't produce diverse enoguh edits. (Khashabi et al., Huang et al.)
  - Gardner: If you recruit NLP experts as annotators, you get clean and diverse-ish edits.

Building a Better Evaluation Dataset

What we want:

- A very high human performance ceiling. 99%+ accuracy?
- A very large size. 100,000+ items?
- Representative coverage of a maximally broad distribution of language-related phenomena.
- **No incentive to build biased models.**

Managing Bias in Benchmarks

Two problems:

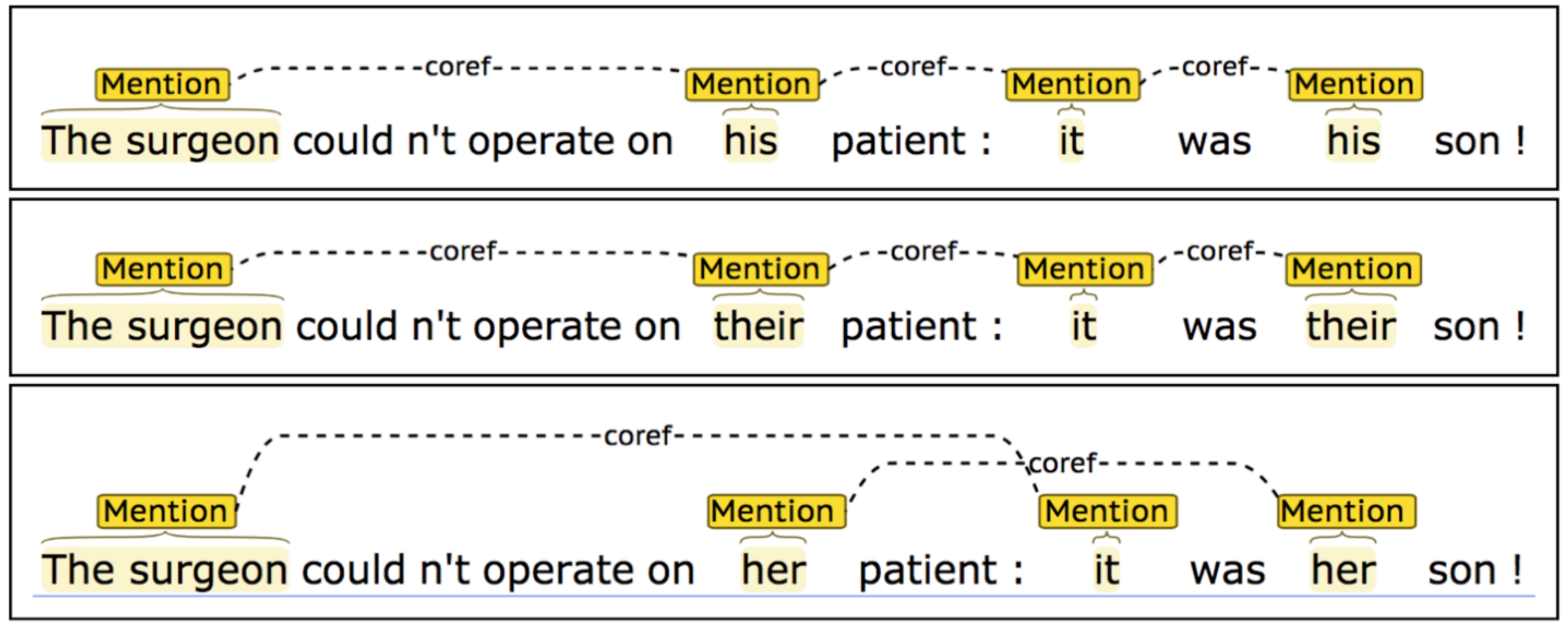
- What counts as bias? Who decides?
    - Race-Criminality associations? Sex-Gender? Caste-Occupation?
    - It's essentially impossible to find a single standard that is broadly appropriate across tasks and cultural contexts, even if we limit ourselves to US English NLU.
- How do we prevent a test set from rewarding biased associations?
    - May be tractable, but little progress so far.

See Blodgett et al. for useful discussion.

Managing Bias in Benchmarks

Awkward compromise: *Bolt-on* bias metrics

- Build separate benchmark datasets that *measure* specific categories of model bias in specific settings.
  - Examples: WinoGender for gender bias in coreference, CrowS-Pairs for several bias categories in MLM predictions



The surgeon could n't operate on his patient : it was his son !

The surgeon could n't operate on their patient : it was their son !

The surgeon could n't operate on her patient : it was her son !

- System building requires multi-objective optimization on NLU benchmarks and bias benchmarks.
- Not a great fit with *leaderboardism*; mismatches between benchmarks can hide relevant biases.

Managing Bias in Benchmarks

What we want:

- A very high human performance ceiling. 99%+ accuracy?
- A very large size. 100,000+ items?
- Balanced coverage of a maximally broad distribution of language-related phenomena.
- No incentive to build biased models.

Evaluation: What's Next?

Some open frontiers:

- Crowdsourcing workflows and incentives: How do we get high agreement and high diversity at scale?

- Expert workflows: How do we best enable dataset creators to patch gaps in datasets?

- Bias mitigation: How do we scale up the creation of useful bias metrics? How do we facilitate widespread use?

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

ML^2 Machine Learning
for Language

Sam Bowman

@sleepinyourhat