Evaluation, Benchmarks, and Experimental Design

CS 6120 Natural Language Processing

Northeastern University

Si Wu

Some slides are based on Tatsunori Hashimoto's lecture slides on evaluation

Logistics

- Coding assignment 4 is released and due Nov 7th
- Final project data and experimental design due a week from now

- Today:
  - Evaluation
  - Metrics

Evaluation

Why do we need evaluation

- Track progress for a specific experiment
  - For example, if you are trying to improve machine translation performance, how do you know how much improvement you get at the end of each epoch when finetuning an LLM?
- Benchmark can drive progress in the field for a specific task
  - With a benchmark, you can compare your results with people who work on the same task

Categories of NLP evaluation

- By eval method:
  - **Automatic** (generally quantitative): metrics computed by algorithms
  - **Human eval** (qualitative, more subjective): crowdworkers, annotators, or sometimes, by interviewing people
- By task structure:
  - Closed-ended tasks: tasks with clear ground truth.
    - Classification: sentiment analysis, NER
  - Open-ended task
    - Generation: summarization

A little terminology clarification

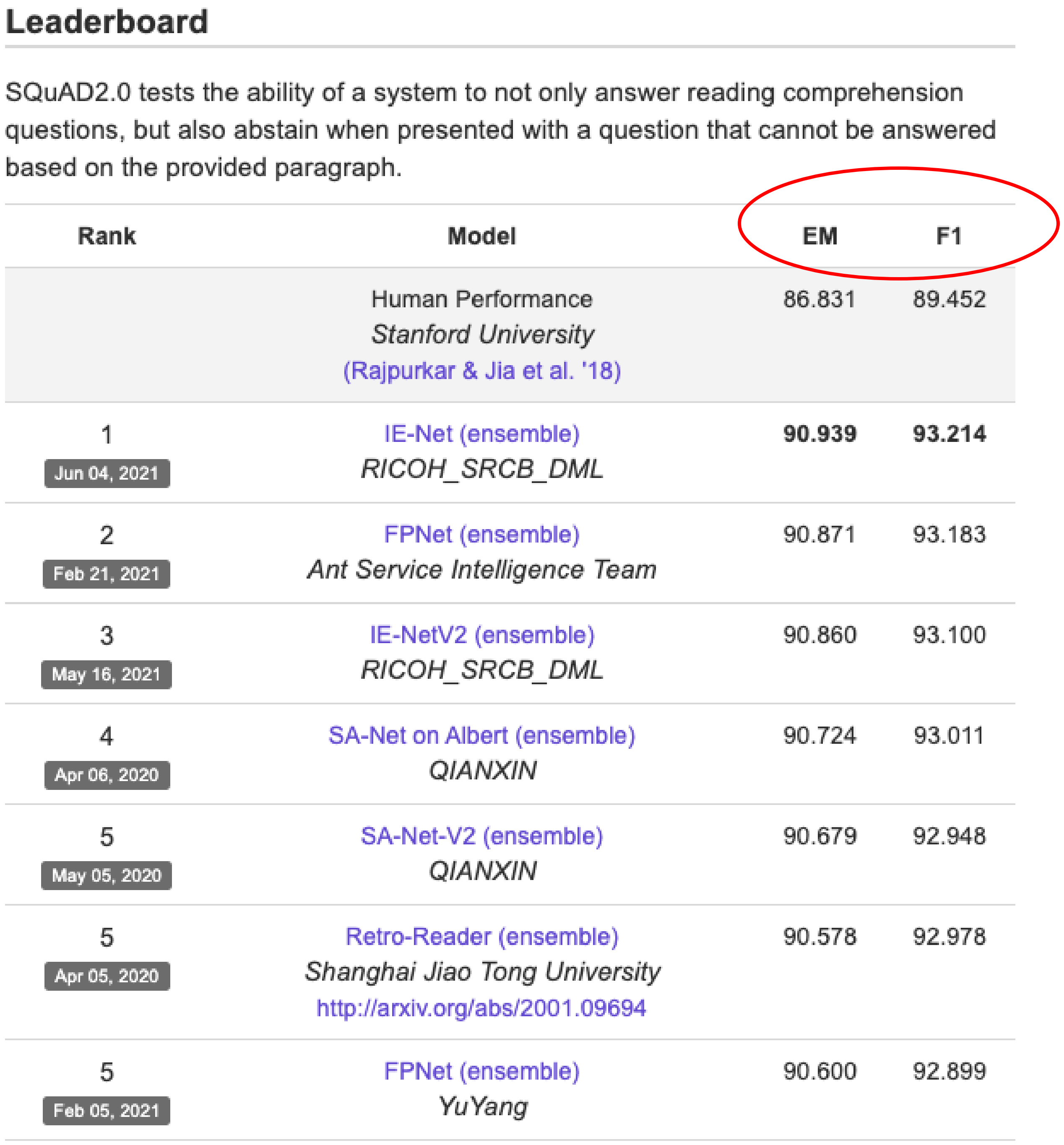
- **Metrics** are ways to measure performance on a task
    - For example: accuracy, F1, BLEU, ROUGE, F1 score, accuracy
    - It’s more reusable. Can be applied to many datasets and tasks
- **Benchmark**: usually consists of a dataset, some tasks, and evaluation protocol
    - Generally, it’s to test a model
    - For example: GLUE, SuperGLUE, SQuAD
    - Some benchmarks only consist of 1 task: SQuAD,
    - Others may consist many different tasks to measure general language capability: SuperGLUE

Single task benchmark: SQuAD2.0

The Stanford Question Answering Dataset

- **S**tanford **Qu**estion **A**nswering **D**ataset (SQuAD)
- A reading comprehension dataset
  - On English Wikipedia articles
- Questions are posed by crowdworkers
  - Answerable questions have a corresponding segment of text in the article
  - Otherwise unanswerable
  - SQuAD 2.0 added more adversarial questions that look like the answerable ones

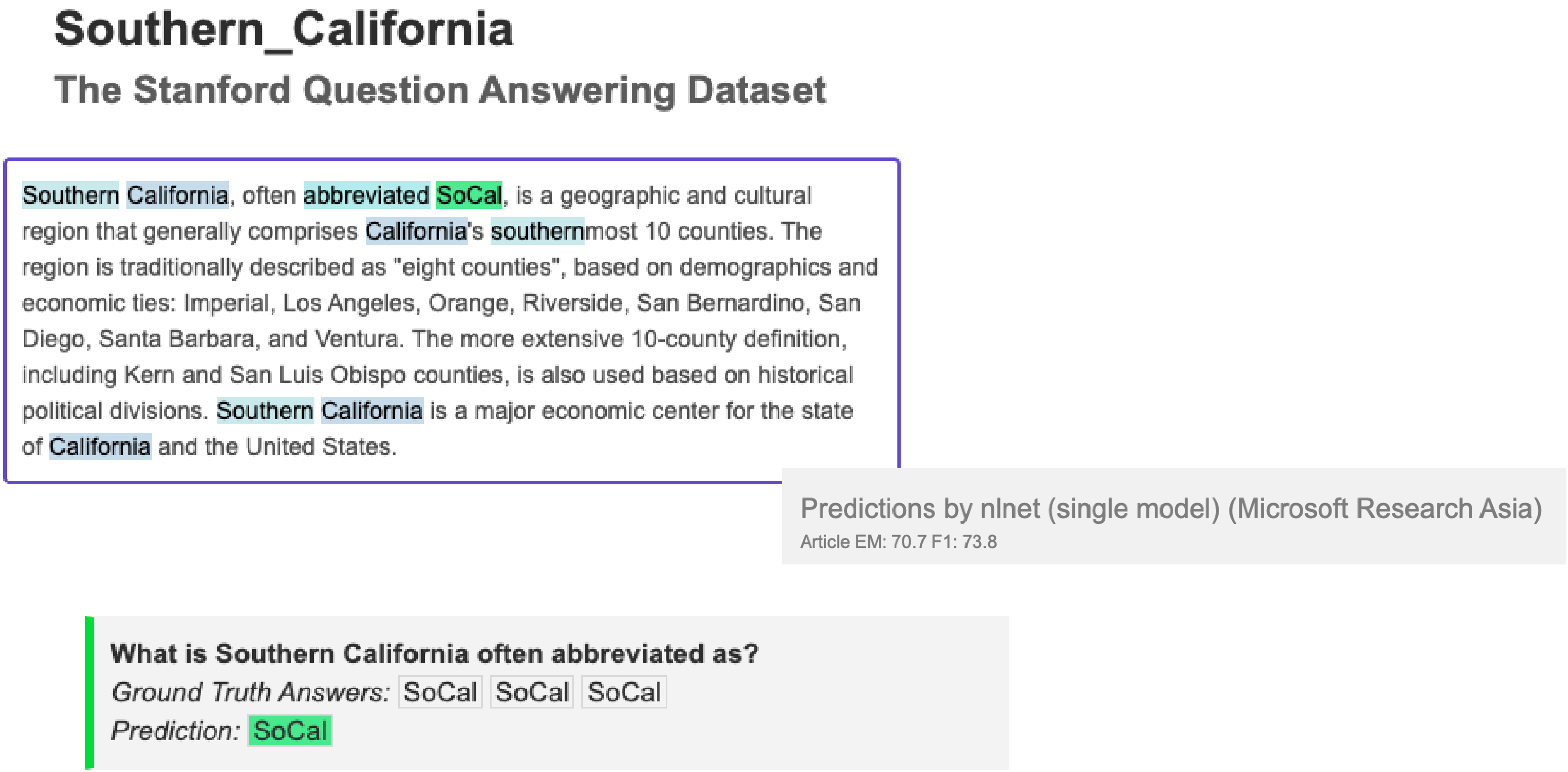
SQuAD2.0
The Stanford Question Answering Dataset



Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jun 04, 2021	IE-Net (ensemble) *RICOH_SRCB_DML*	**90.939**	**93.214**
2 Feb 21, 2021	FPNet (ensemble) *Ant Service Intelligence Team*	90.871	93.183
3 May 16, 2021	IE-NetV2 (ensemble) *RICOH_SRCB_DML*	90.860	93.100
4 Apr 06, 2020	SA-Net on Albert (ensemble) *QIANXIN*	90.724	93.011
5 May 05, 2020	SA-Net-V2 (ensemble) *QIANXIN*	90.679	92.948
5 Apr 05, 2020	Retro-Reader (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694	90.578	92.978
5 Feb 05, 2021	FPNet (ensemble) *YuYang*	90.600	92.899



Southern_California

The Stanford Question Answering Dataset

Southern California, often abbreviated SoCal, is a geographic and cultural region that generally comprises California's southernmost 10 counties. The region is traditionally described as "eight counties", based on demographics and economic ties: Imperial, Los Angeles, Orange, Riverside, San Bernardino, San Diego, Santa Barbara, and Ventura. The more extensive 10-county definition, including Kern and San Luis Obispo counties, is also used based on historical political divisions. Southern California is a major economic center for the state of California and the United States.

Predictions by nlnet (single model) (Microsoft Research Asia)

Article EM: 70.7 F1: 73.8

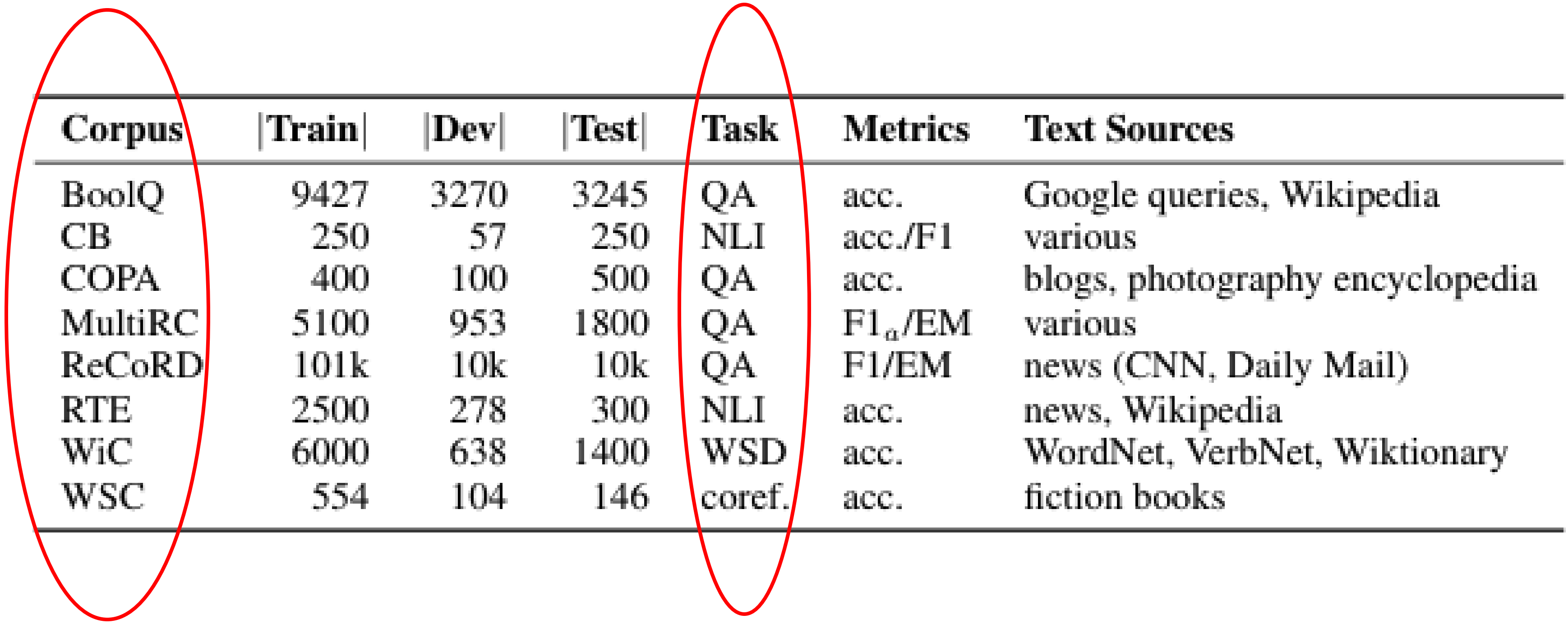
What is Southern California often abbreviated as?

Ground Truth Answers: SoCal SoCal SoCal

Prediction: SoCal

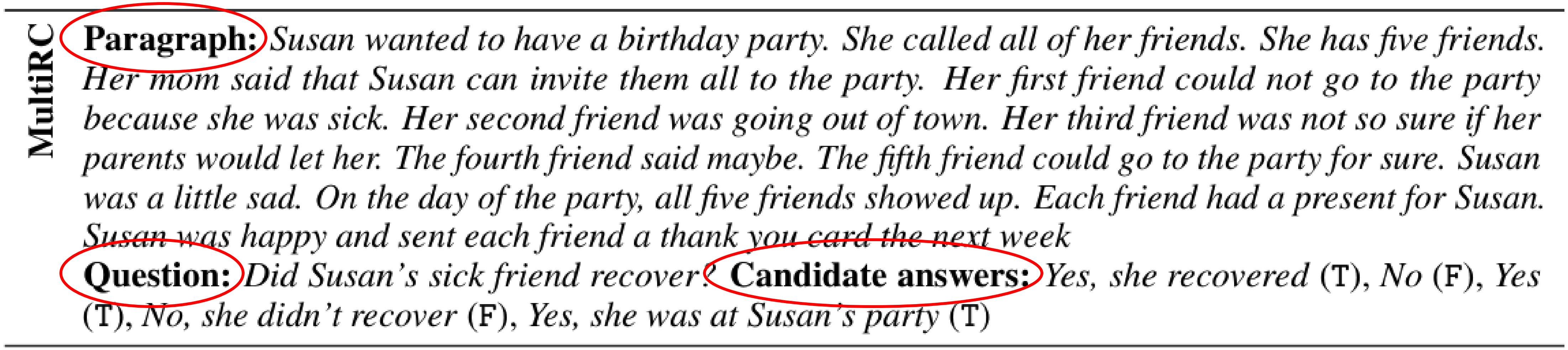
Multi-task benchmark: SuperGLUE

- “a new benchmark styled after GLUE with a new set of more **difficult language understanding tasks**, improved resources, and a new public leaderboard”



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

QA example from SuperGLUE



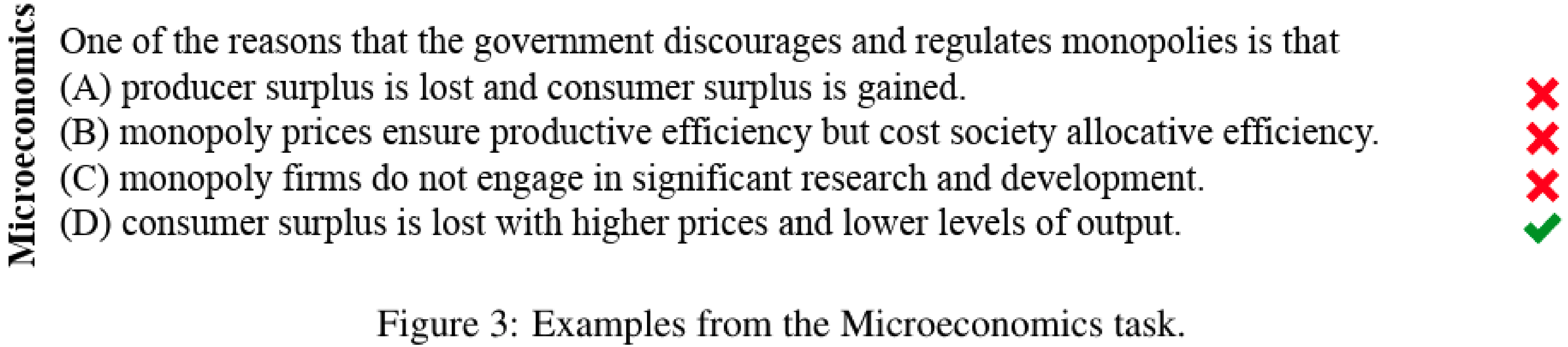
MultiRC

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

Multitask benchmark: MMLU

- Massive Multitask Language Understanding (MMLU)
- “The test covers 57 tasks including elementary mathematics, US history, computer science, law, and more.”
    - Multiple-choice questions on these topics
    - “The questions in the dataset were manually collected by graduate and undergraduate students from freely available sources online.”

An example from MMLU



Microeconomics

One of the reasons that the government discourages and regulates monopolies is that

(A) producer surplus is lost and consumer surplus is gained. ✗

(B) monopoly prices ensure productive efficiency but cost society allocative efficiency. ✗

(C) monopoly firms do not engage in significant research and development. ✗

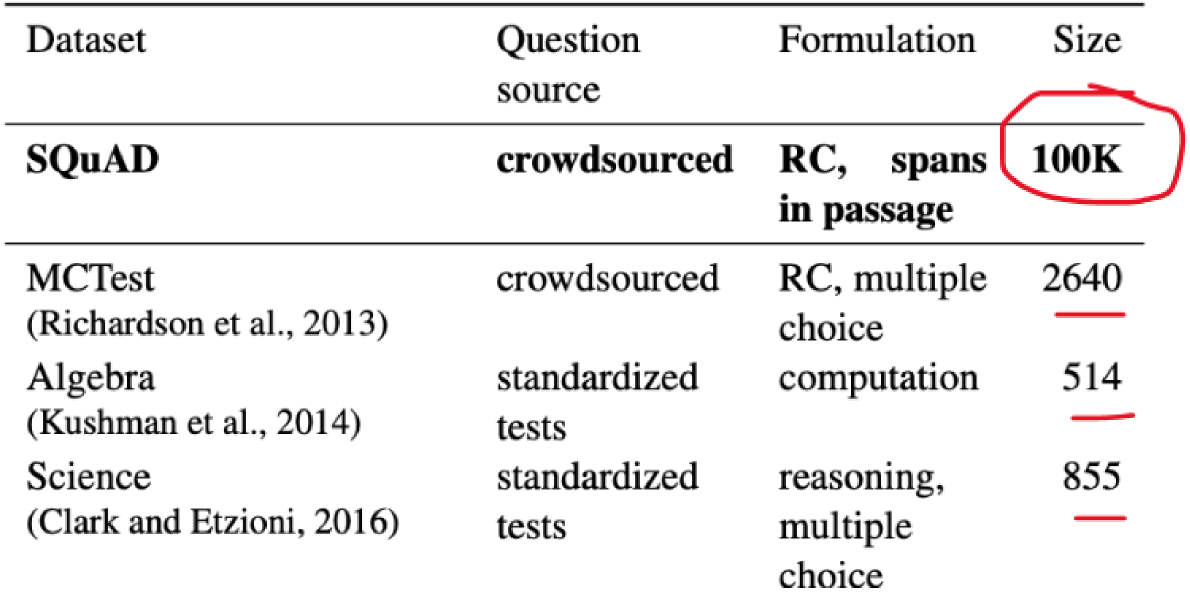
(D) consumer surplus is lost with higher prices and lower levels of output. ✓

Figure 3: Examples from the Microeconomics task.

What makes a good benchmark?

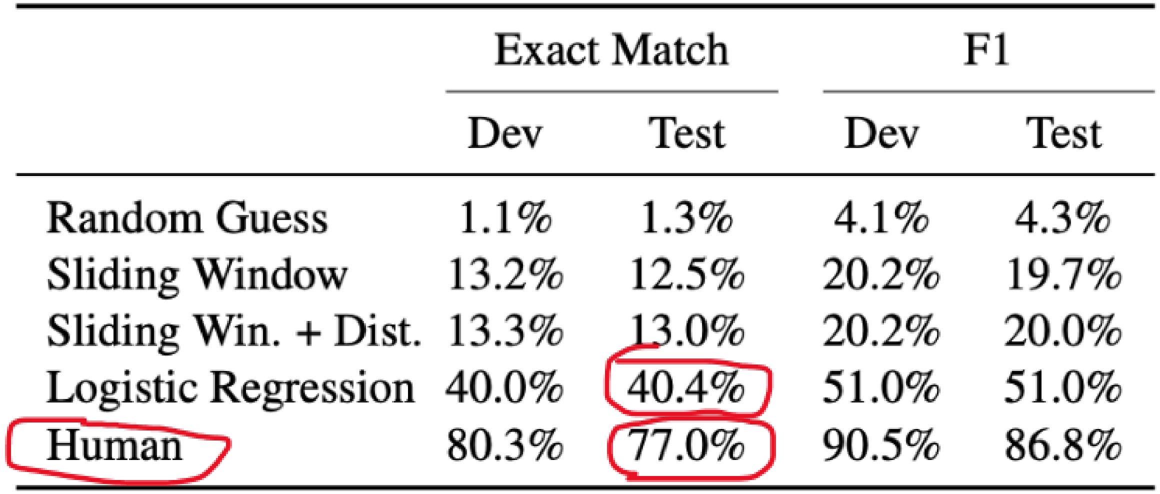
- **Example selection (scale, diversity)**
  - Benchmark should cover the phenomena of interest
  - Complex phenomena require many samples
- **Difficulty**
  - Doable for humans
  - Hard for baselines (at the time the benchmark was created)
- **Annotation quality and consistency**
  - ‘Correct’ behavior should be clear

Example of a good benchmark



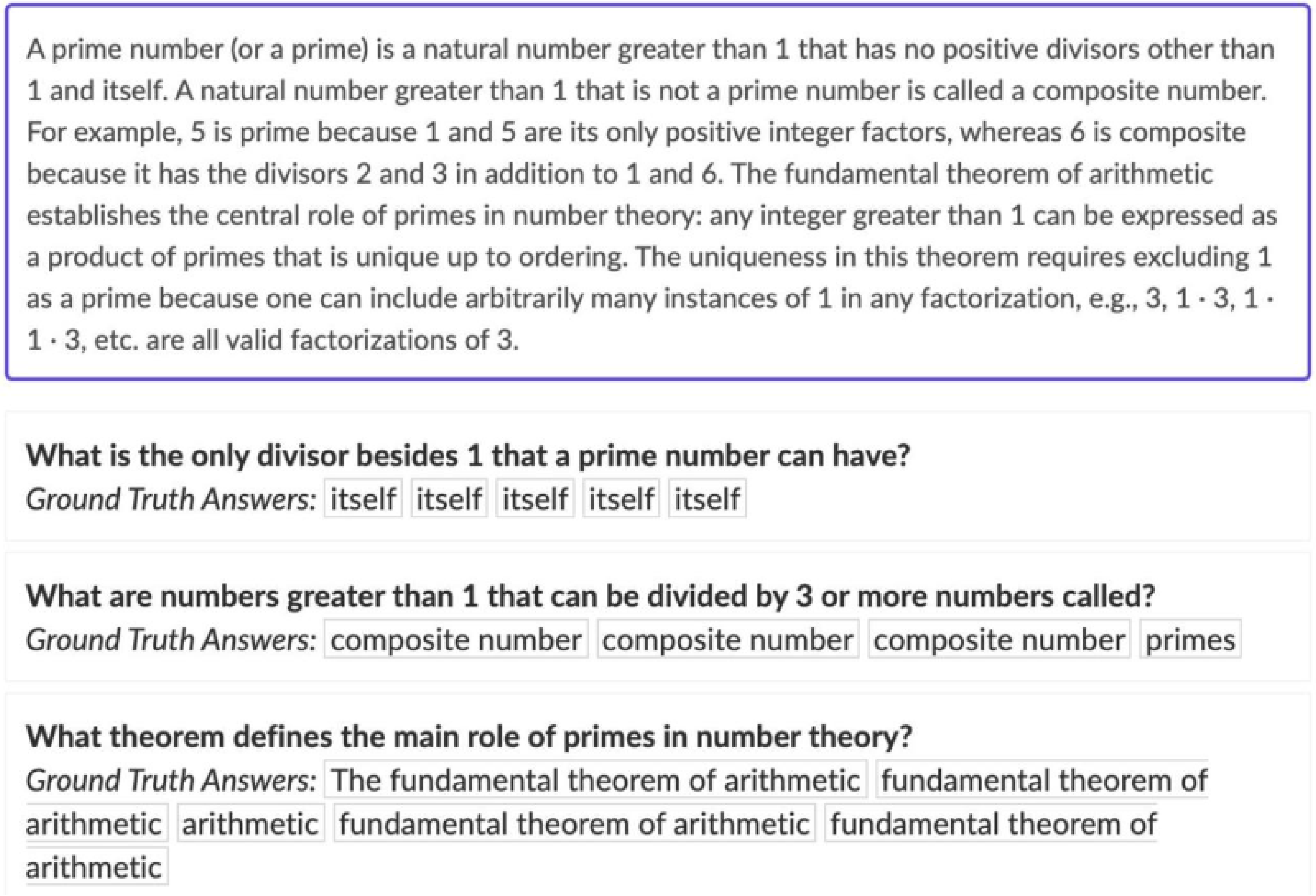
Dataset	Question source	Formulation	Size
SQuAD	**crowdsourced**	**RC, spans in passage**	**100K**
MCTest (Richardson et al., 2013)	crowdsourced	RC, multiple choice	2640
Algebra (Kushman et al., 2014)	standardized tests	computation	514
Science (Clark and Etzioni, 2016)	standardized tests	reasoning, multiple choice	855

Scale (and inclusion of training data)



	Exact Match		F1	
	Dev	Test	Dev	Test
Random Guess	1.1%	1.3%	4.1%	4.3%
Sliding Window	13.2%	12.5%	20.2%	19.7%
Sliding Win. + Dist.	13.3%	13.0%	20.2%	20.0%
Logistic Regression	40.0%	40.4%	51.0%	51.0%
Human	80.3%	77.0%	90.5%	86.8%

Large headroom to human perf



A prime number (or a prime) is a natural number greater than 1 that has no positive divisors other than 1 and itself. A natural number greater than 1 that is not a prime number is called a composite number. For example, 5 is prime because 1 and 5 are its only positive integer factors, whereas 6 is composite because it has the divisors 2 and 3 in addition to 1 and 6. The fundamental theorem of arithmetic establishes the central role of primes in number theory: any integer greater than 1 can be expressed as a product of primes that is unique up to ordering. The uniqueness in this theorem requires excluding 1 as a prime because one can include arbitrarily many instances of 1 in any factorization, e.g., 3, 1 · 3, 1 · 1 · 3, etc. are all valid factorizations of 3.

What is the only divisor besides 1 that a prime number can have?
Ground Truth Answers: itself itself itself itself itself

What are numbers greater than 1 that can be divided by 3 or more numbers called?
Ground Truth Answers: composite number composite number composite number primes

What theorem defines the main role of primes in number theory?
Ground Truth Answers: The fundamental theorem of arithmetic fundamental theorem of arithmetic arithmetic fundamental theorem of arithmetic fundamental theorem of arithmetic

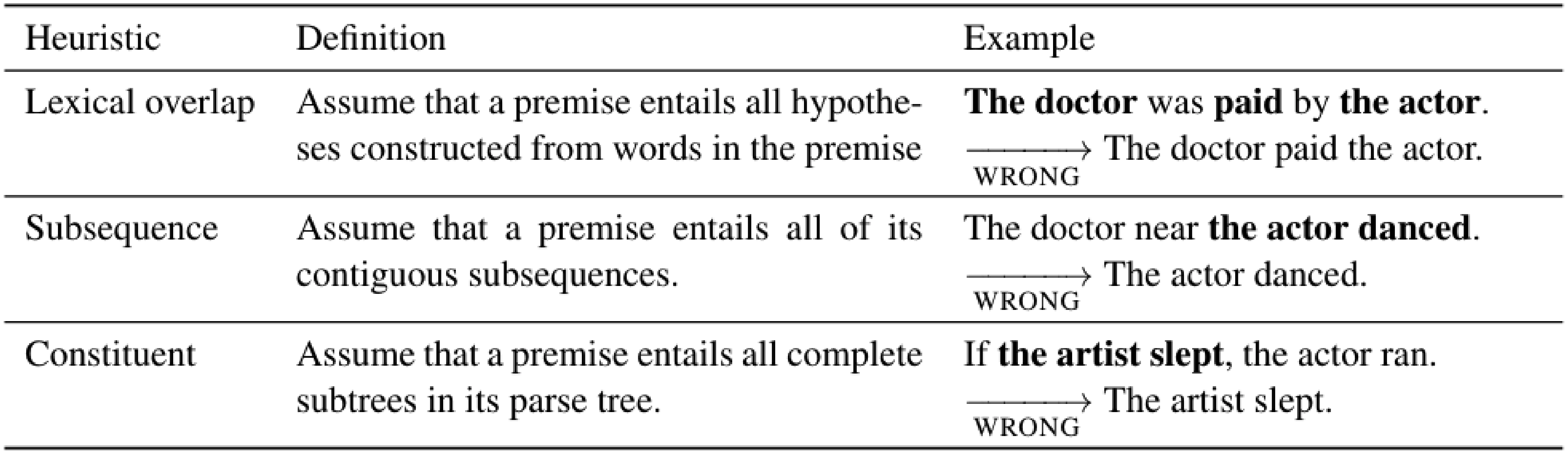
Easy, relatively clean automatic evaluation

Diagnostic tests

- Sometimes we want to make sure a model is not getting good results using simple heuristics
  - As opposed to actually understand the question/language
  - For example: for entailment, it thinks A→B is true if there are lot of overlapping words.
- We can create a diagnostic test set to examine if a specific undesirable behavior exists in the model

An example of diagnostic test

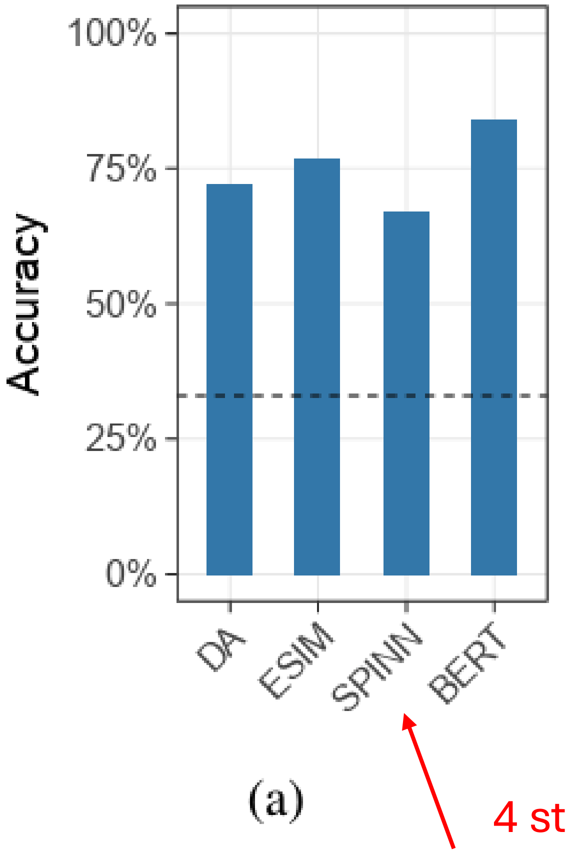
HANS: (Heuristic Analysis for NLI Systems) tests syntactic heuristics in NLI



Heuristic	Definition	Example
Lexical overlap	Assume that a premise entails all hypotheses constructed from words in the premise	**The doctor** was **paid** by **the actor**. $\xrightarrow[\text{WRONG}]{}$ The doctor paid the actor.
Subsequence	Assume that a premise entails all of its contiguous subsequences.	The doctor near **the actor danced**. $\xrightarrow[\text{WRONG}]{}$ The actor danced.
Constituent	Assume that a premise entails all complete subtrees in its parse tree.	If **the artist slept**, the actor ran. $\xrightarrow[\text{WRONG}]{}$ The artist slept.

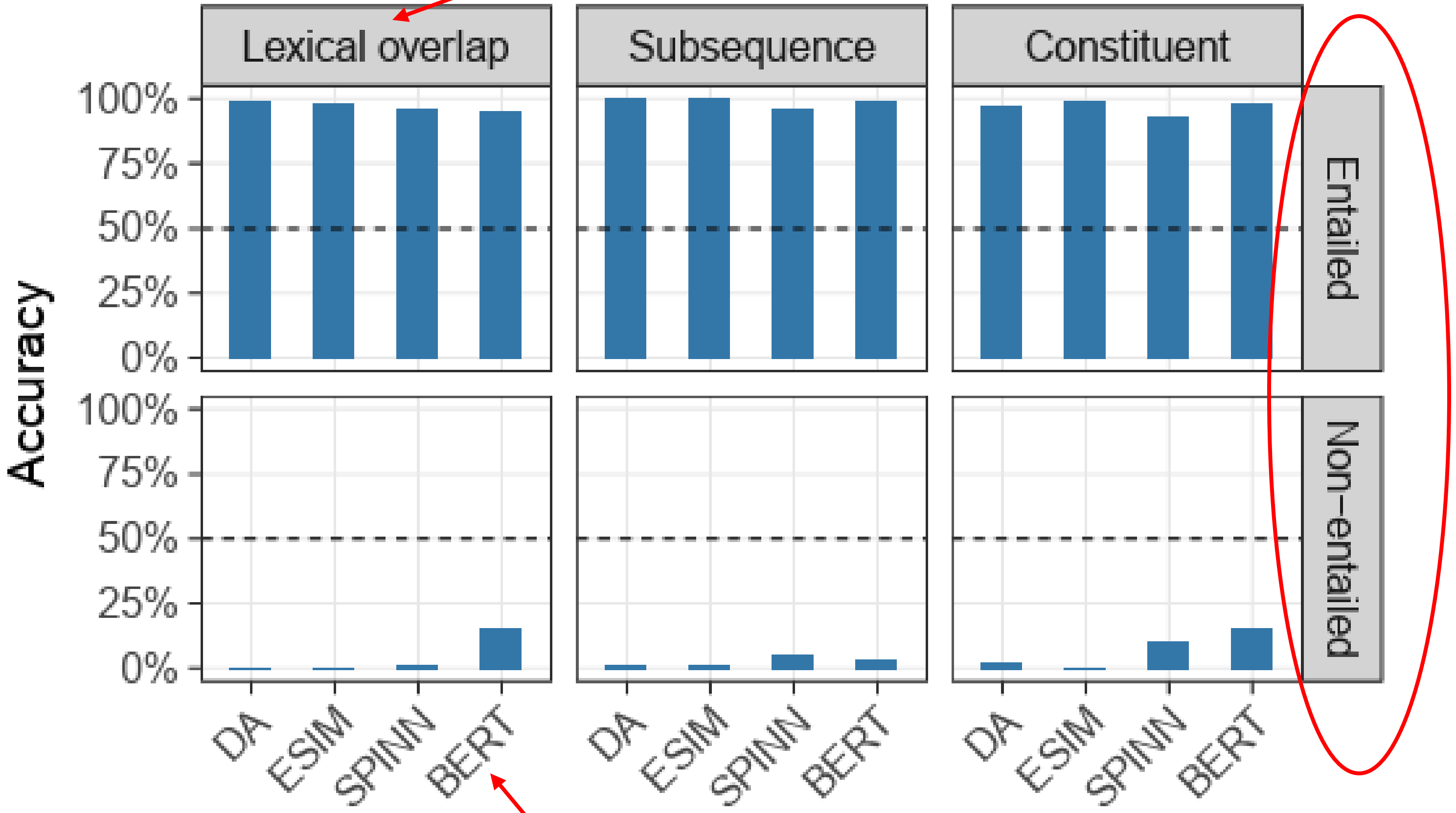
Table 1: The heuristics targeted by the HANS dataset, along with examples of incorrect entailment predictions that these heuristics would lead to.

HANS model analysis in natural language inference



4 strong Natural language inference models

Accuracy is high when simple heuristics work!



And when using simple heuristics doesn’t work, accuracy is low… → do these models really understand language then?

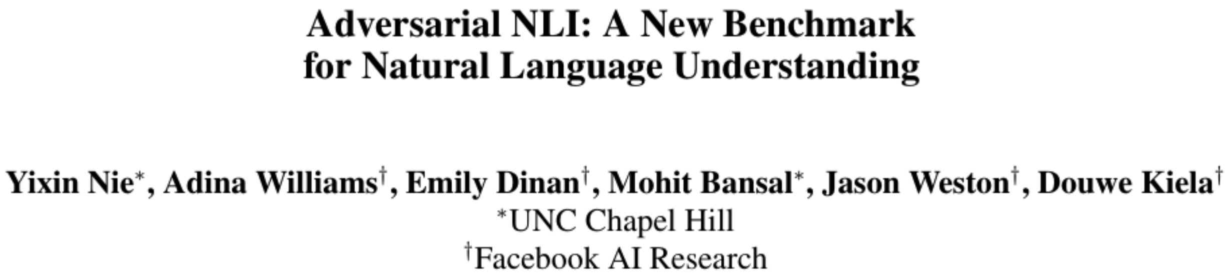
Fitting the dataset vs learning the task

- Imagine if your model can see an image of your calendar and answer questions about your schedule. Do you think it should also be able to understand an image of a floorplan and answer any related questions?
- Domain matters!
  - But a model should do well on reasonable out-of-domain test instances
  - Is the model learning the dataset or the task?
- Food for thought:

If your training data cover a huge variety of topics and knowledge, do you think:
since the LM has gained the language and reasoning capability from this, will it do well on anything, even knowledge it didn’t learn from training?

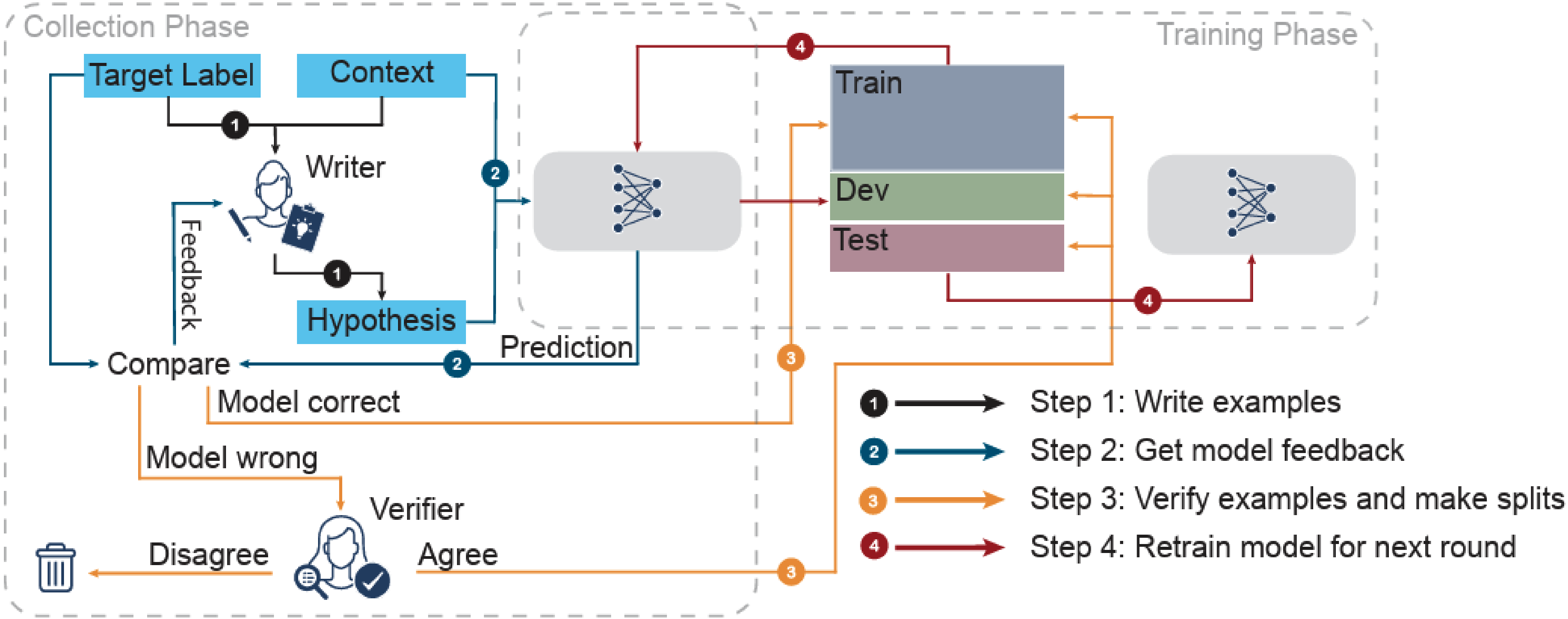
Adversarial benchmarking

- Start with a strong baseline model good at natural language inference
- Show the model to crowdworkers
  - Crowdworkers will interact with the model via a special interface
  - Workers will see what the model got wrong,
  - And workers will write new tricky examples
- Add these tricky examples to training set
- Retrain and repeat the process



Adversarial NLI: A New Benchmark for Natural Language Understanding

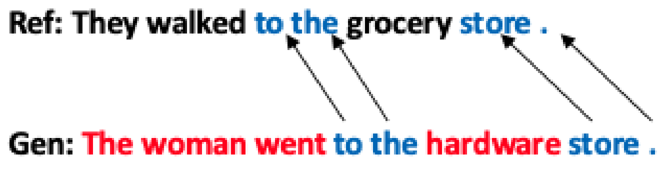
Yixin Nie[*], Adina Williams[†], Emily Dinan[†], Mohit Bansal[*], Jason Weston[†], Douwe Kiela[†]
[*]UNC Chapel Hill
[†]Facebook AI Research


Collection Phase
Training Phase
Target Label
Context
1
Writer
2
Feedback
1
Hypothesis
Prediction
Compare
2
Model correct
Model wrong
Verifier
Disagree
Agree
3
3
4
Train
Dev
Test
4
1 Step 1: Write examples
2 Step 2: Get model feedback
3 Step 3: Verify examples and make splits
4 Step 4: Retrain model for next round

Evaluation for text generation

- More difficult because it could be open-ended
- Sometimes we have example or expected output
  - Generated output doesn’t have to look identical, but how to make sure the meaning is the same (or equally good)
- This is where human evaluation is really helpful!
  - But could be expensive and laborious

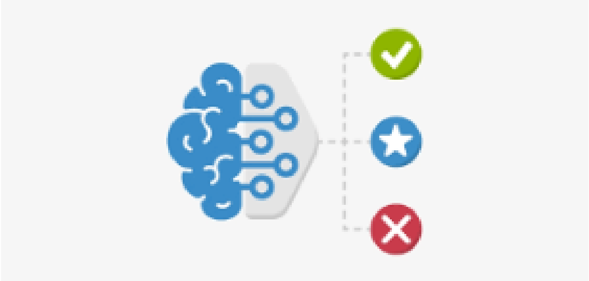
Evaluation for text generation



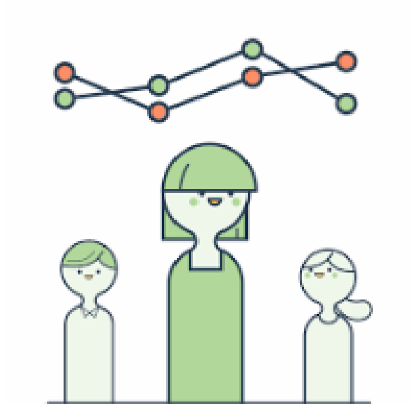
Ref: They walked to the grocery store .

Gen: The woman went to the hardware store .

Content Overlap Metrics



Model-based Metrics



Human Evaluations

Evaluation for text generation: n-gram overlap

- Compute a score of the lexical similarity between the generated output and the gold-standard
  - E.g. how many words in the generated translation is overlapping with the reference translation?
- Simplest, fast, and efficient
- N-gram overlap metrics: BLEU, ROUGE, METEOR, etc.
- Sometimes an equally good answer can have very little or no overlapping words with your ground truth.
- Vice versa, you can overlap a lot of words, but the meaning is different.

Tricky evaluation for text generation

Question: What to say to a friend who just broke their arm?
Ground truth: Man! That's gnarly! Get well soon!

Generated output 1: I am so sorry to hear that! (1 overlapping word)

Generated output 2: So sorry to hear that! Wish you a speedy recovery! (no overlapping word)

Tricky evaluation for text generation

Question: How do you feel about this horror movie?
Ground truth: It’s terrifying! I like it!

Generated output 1: It’s terrifying! I don’t like it!

Completely different sentiment though mostly overlapping!

Evaluation for text generation: n-gram overlap

- If you are working tasks that are more open-ended or allow creative generated output, this metric may not work:

Ok but not ideal: machine translation – good translations can be similar

Worse: summarization – might have similarity due to reference text

Much worse: dialogue – very open-ended, conversation can turn in many different valid ways

Much much worse: story generation – could be creative and sometimes there shouldn't be a ground truth

Model-based evaluation: semantic similarity

- Usually use pre-trained models to encode the input
    - The embeddings will capture the underlying semantics rather than surface-level word patterns
- There are many similarity metrics you can then use on these embeddings
    - E.g. cosine similarity

- For example, two paraphrases of each other can have completely different set of words, but their embedding distance will be close.

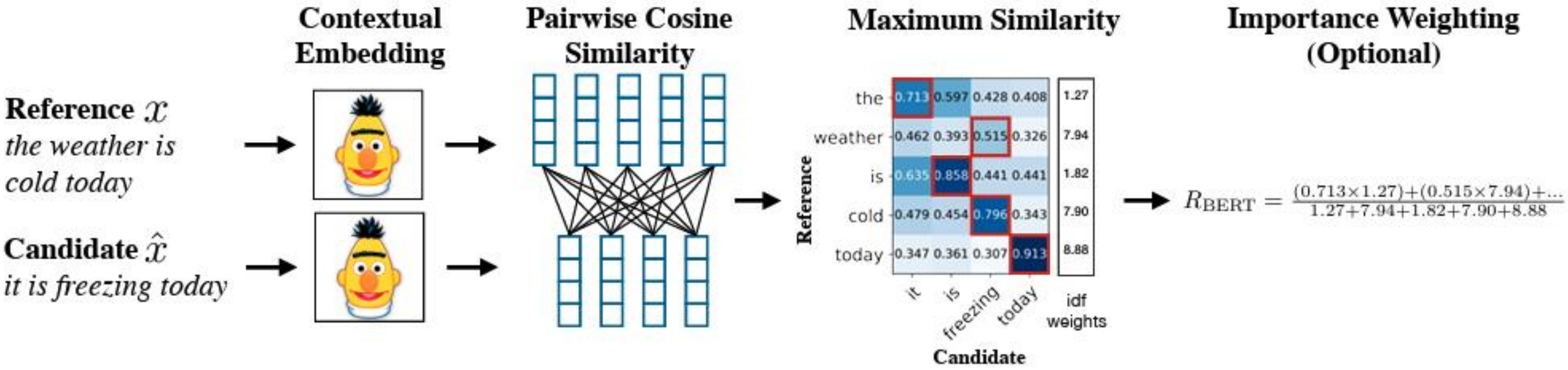
Model-based evaluation: semantic similarity

- Example metrics:
  - Word distance:
    - YISI (Lo, 2019): aligns words between system and reference texts using multilingual embeddings
    - Word mover’s distance (Kusner et al., 2015; Zhao et al., 2019): measures minimal distance that word embeddings must travel to transform one text to another

Model-based evaluation: semantic similarity

- Example metrics:
  - Beyond words:
    - BERTScore (Zhang et al, 2020): using contextualized BERT embeddings to compute precision, recall, F1 between candidate and reference sentence at token level
    - BLEURT (Sellam et al., 2020): a learned eval metric that fine-tunes BERT on human ratings to predict the quality of text based on semantic similarity and fluency
    - COMET (Rei et al., 2020): for translation, also trained on human eval data to estimate translation quality using source, hypothesis, and reference translation embeddings
    - MAUVE (Pillutla et al., 2022): measure the divergence between the distribution of human text and generated text in a quantized embedding space

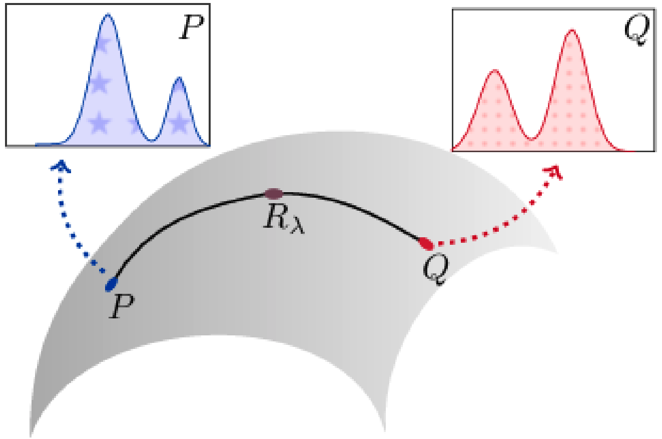
BERTScore

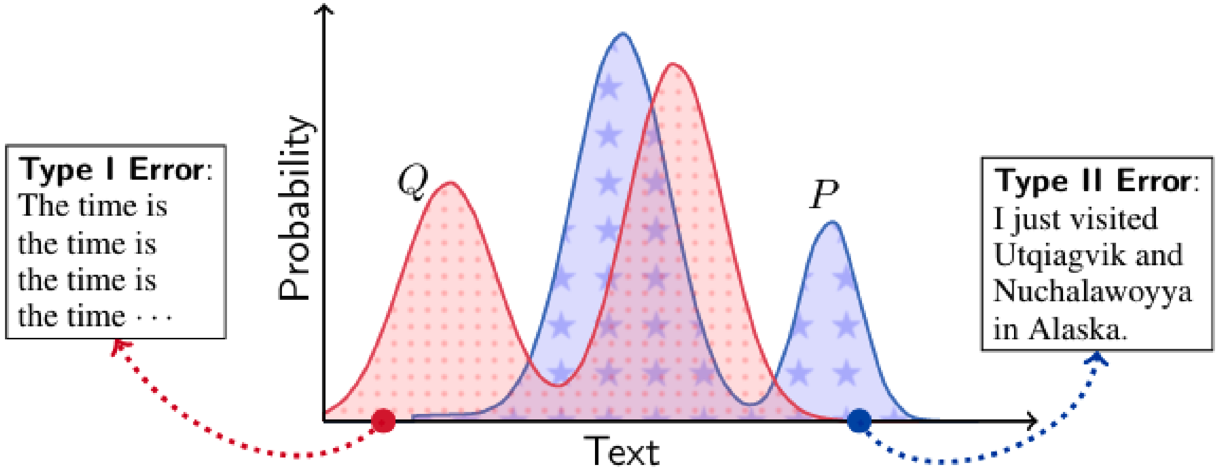


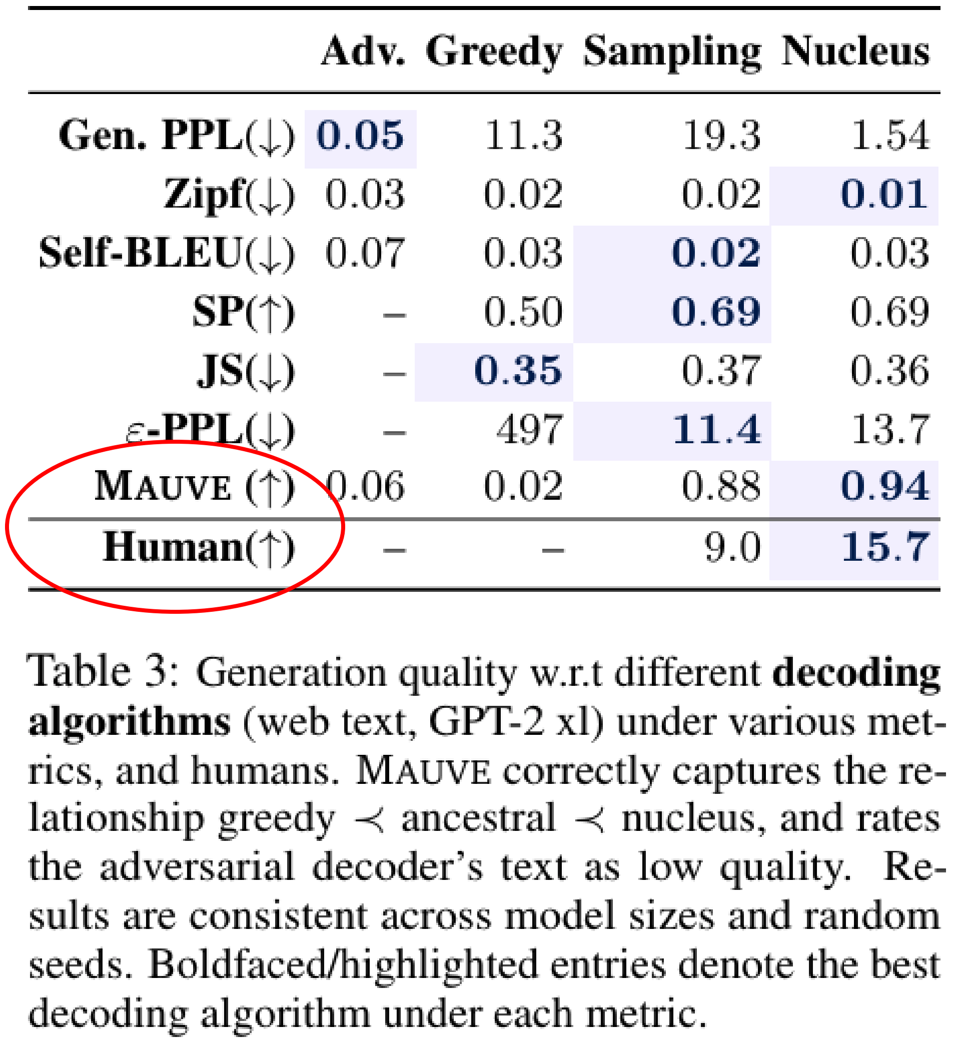
	it	is	freezing	today	idf weights
the	0.713	0.597	0.428	0.408	1.27
weather	0.462	0.393	0.515	0.326	7.94
is	0.635	0.858	0.441	0.441	1.82
cold	0.479	0.454	0.796	0.343	7.90
today	0.347	0.361	0.307	0.913	8.88

$$R_{\mathrm{BERT}} = \frac{(0.713\times1.27)+(0.515\times7.94)+\ldots}{1.27+7.94+1.82+7.90+8.88}$$

MAUVE (Pillutla et al., 2022)




Type I Error: The time is the time is the time is the time · · ·
Probability
Q
P
Type II Error: I just visited Utqiagvik and Nuchalawoyya in Alaska.
Text



	Adv.	Greedy	Sampling	Nucleus
Gen. PPL(↓)	**0.05**	11.3	19.3	1.54
Zipf(↓)	0.03	0.02	0.02	**0.01**
Self-BLEU(↓)	0.07	0.03	**0.02**	0.03
SP(↑)	–	0.50	**0.69**	0.69
JS(↓)	–	**0.35**	0.37	0.36
ε-PPL(↓)	–	497	**11.4**	13.7
MAUVE (↑)	0.06	0.02	0.88	**0.94**
Human(↑)	–	–	9.0	**15.7**

Table 3: Generation quality w.r.t different **decoding algorithms** (web text, GPT-2 xl) under various metrics, and humans. MAUVE correctly captures the relationship greedy ≺ ancestral ≺ nucleus, and rates the adversarial decoder's text as low quality. Results are consistent across model sizes and random seeds. Boldfaced/highlighted entries denote the best decoding algorithm under each metric.

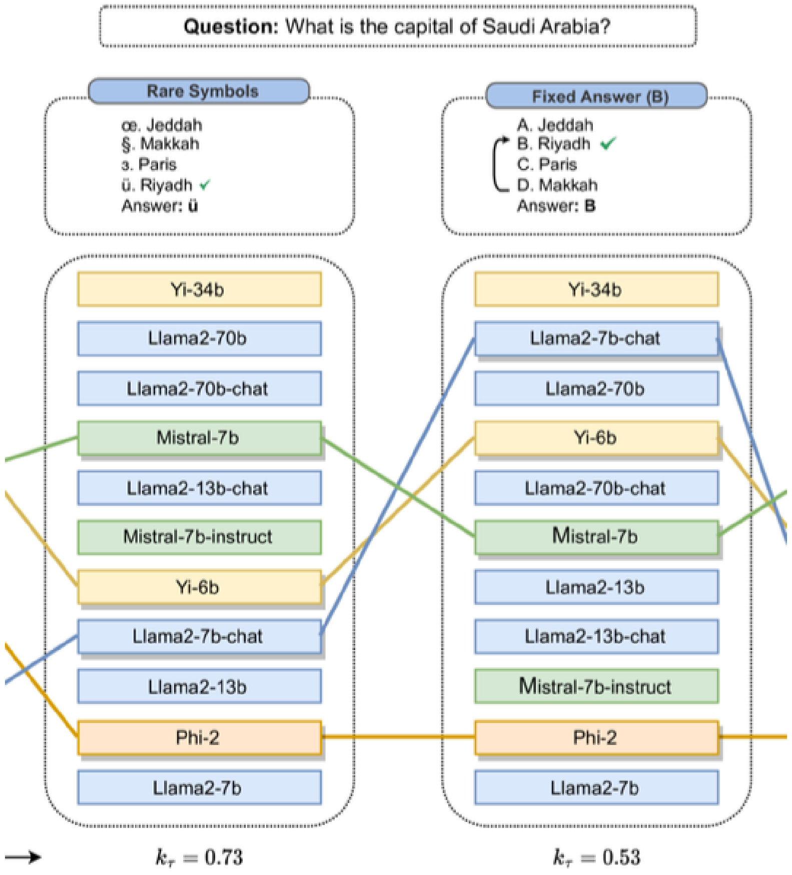
Model-based evaluation: Using LLM as evaluator

- To lower the cost, prompting LLMs to evaluate your model output
  - Depends on the model and task, but could have high correlation with human
- Potential issues:
  - Self-bias: LLMs may prefer LLM-generated text
  - Different prompts may result in different output
- Examples: AlpacaEval, “GPT-as-judge”

Finally, human evaluation

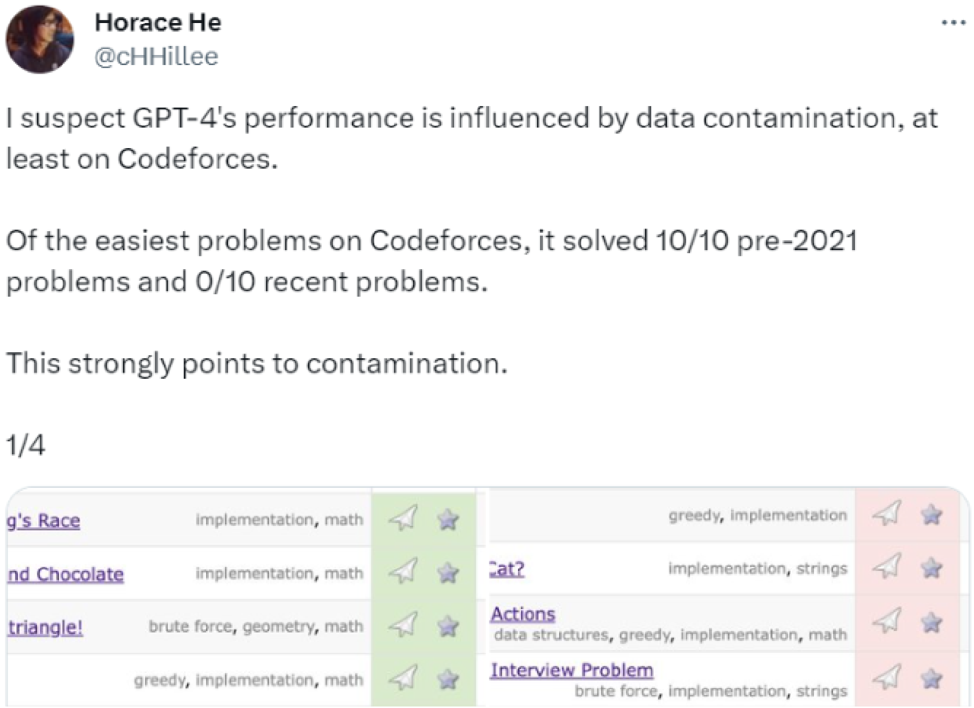
- Gold-standard in developing new automatic metrics
  - New metrics must correlate well with human eval
- Format:
  - Crowdsource: Amazon Mechanical Turk (AMT), Prolific
  - Interview: less common, need IRB approval
- Annotation format:
  - You can either define some categories/dimensions, then ask human to rate on some scale, or make it a multiple choice, or free response for a specific question (then process the response later)
- Downside:
  - Handling noise, ensuring quality is difficult
  - How to handle disagreement among crowdworkers?
  - Cost is high
  - Irreproducible

Evaluation contamination in closed models



[Alzahrani et al 2024]

Consistency



Horace He
@cHHillee

I suspect GPT-4's performance is influenced by data contamination, at least on Codeforces.

Of the easiest problems on Codeforces, it solved 10/10 pre-2021 problems and 0/10 recent problems.

This strongly points to contamination.

1/4

g's Race	implementation, math		greedy, implementation
nd Chocolate	implementation, math	:at?	implementation, strings
triangle!	brute force, geometry, math	Actions	data structures, greedy, implementation, math
	greedy, implementation, math	Interview Problem	brute force, implementation, strings

Contamination

How to design an experiment

General comments and tips for your data and experimental design submission

- Evaluation is a big part of experimental design
    - If you can track progress, it’s difficult to know when to stop for a lot of tasks involves training/finetuning
- How many variables are there?
    - How many you can control?
    - How many latent variables?
    - How many are dependent on other variables?
    - You should design experiment in a way that you know what causes the improvement or change
        - This is crucial for evaluation and analysis

- What are the constraints of your experiment?
  - Is your result applicable to test data from another domain?
  - Is your experiment results just by chance?
    - Is it statistically significant?
  - What did you not consider or what did you filtered out from the data?
- When designing your experiment, it’s important to constantly ask yourself: what is the goal of my experiment? Why should we care about this?

- Just because someone public a paper on something, doesn’t mean there’s no room for improvement, or that’s the best way to conduct an experiment on that topic/task
    - You should always be critical about what you read, and that’s just being a good scientist/researcher
- Finally, always consider what your evaluation metric method does not capture?
    - And what can other people or future work do about this?