Multi-Task Benchmarks

NLP: Fall 2024

Anoop Sarkar

Evaluating language understanding

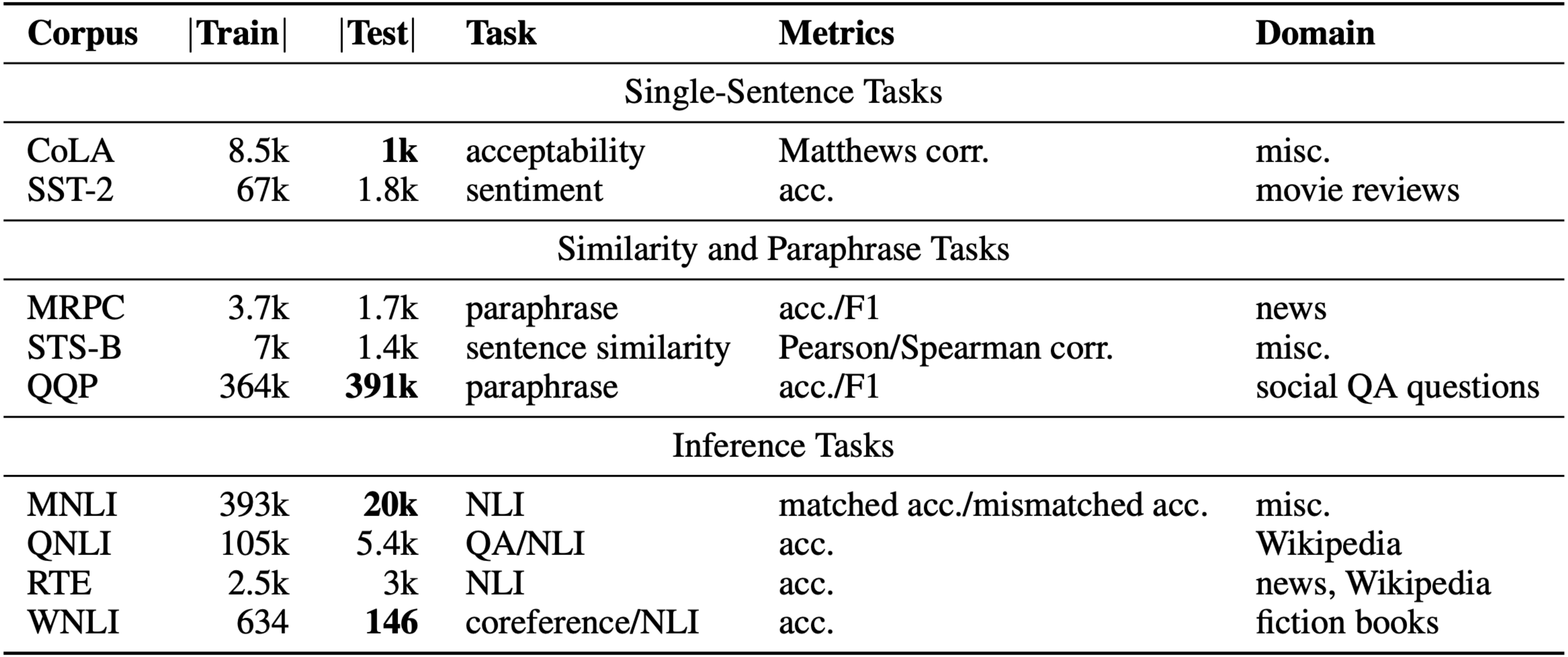
Benchmarks for LLMs

- LLMs can be fine-tuned to many different tasks in NLP
- Evaluation of representation learning (sentential embeddings or QA tasks)
- Online Leaderboard, e.g. decaNLP (2018)
- Combine diverse set of NLU tasks that are quick to fine-tune and test
    - GLUE (ICLR 2019) → SuperGLUE (NeurIPS 2019)
    - SWAG (EMNLP 2018) → HellaSWAG (ACL 2019)
- New benchmarks:
    - MMLU (2020)
    - MTEB (2022)

GLUE

GLUE

https://gluebenchmark.com/



Corpus	|Train|	|Test|	Task	Metrics	Domain
Single-Sentence Tasks					
CoLA	8.5k	**1k**	acceptability	Matthews corr.	misc.
SST-2	67k	1.8k	sentiment	acc.	movie reviews
Similarity and Paraphrase Tasks					
MRPC	3.7k	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.4k	sentence similarity	Pearson/Spearman corr.	misc.
QQP	364k	**391k**	paraphrase	acc./F1	social QA questions
Inference Tasks					
MNLI	393k	**20k**	NLI	matched acc./mismatched acc.	misc.
QNLI	105k	5.4k	QA/NLI	acc.	Wikipedia
RTE	2.5k	3k	NLI	acc.	news, Wikipedia
WNLI	634	**146**	coreference/NLI	acc.	fiction books

Single-sentence tasks

CoLA, Corpus of Linguistic Acceptability

- English acceptability judgments drawn from books and journal articles on linguistic theory.
- Each example is a sequence of words annotated with whether it is a grammatical English sentence. Score is between [-1,1]
    - "Bill seems to be obnoxious, but I don't think that Sam seems."
    - "John is impressed as pompous."
    - "We proved Smith to the authorities to be the thief."
    - "I only eat fish raw fresh."
    - "What did that Bill wore surprise everyone?"

Single-sentence tasks

SST-2, Stanford Sentiment Treebank

- Sentences from movie reviews and human annotations of sentiment
- Task: predict sentiment using two-way class split (positive/negative)
  - "uneasy mishmash of styles and genres ."
  - "it 's also heavy-handed and devotes too much time to bigoted views ."
  - "waydowntown is by no means a perfect film , but its boasts a huge charm factor and smacks of originality ."
  - "a remarkable 179-minute meditation on the nature of revolution ."
  - "starts off with a bang , but then fizzles like a wet stick of dynamite at the very end ."

Similarity and Paraphrase Tasks

MRPC, Microsoft Research Paraphrase Corpus

- Sentence pairs with human annotations whether sentences are semantically equivalent
- Classes are imbalanced (68% positive) so accuracy and F1 score is reported

"PCCW 's chief operating officer , Mike Butcher , and Alex Arena , the chief financial officer , will report directly to Mr So ."	"Current Chief Operating Officer Mike Butcher and Group Chief Financial Officer Alex Arena will report to So ."	1
"The company didn 't detail the costs of the replacement and repairs ."	"But company officials expect the costs of the replacement work to run into the millions of dollars ."	0
"Ballmer has been vocal in the past warning that Linux is a threat to Microsoft ."	"In the memo , Ballmer reiterated the open-source threat to Microsoft ."	0
"Ricky Clemons ' brief , troubled Missouri basketball career is over ."	"Missouri kicked Ricky Clemons off its team , ending his troubled career there ."	1

Similarity and Paraphrase Tasks

QQP, Quora Question Pairs

- Question pairs from community QA website, Quora
- Classes are imbalanced (63% negative) so accuracy and F1 score is reported

"What are the best things to do in Hong Kong?"	"What is the best thing in Hong Kong?"	1
"Which is the best gaming laptop under 40k?"	"Which is the best gaming laptop under 40,000 rs?"	1
"How is vanilla extract made?"	"How do you make sugar cookies without vanilla extract?"	0
"How do you close a Bank of America account?"	"How can one close a bank account online?"	0

Similarity and Paraphrase Tasks

STS-B, Semantic Textual Similarity Benchmark

- Sentence pairs drawn from various sources, human annotated with similarity score from 1 to 5
- Task: predict [1,5]. Report Pearson and Spearman correlation coefficients

"A man with a hard hat is dancing."	"A man wearing a hard hat is dancing."	5
"A young child is riding a horse."	"A child is riding a horse."	4.75
"The girl sang into a microphone."	"The lady sang into the microphone."	2.4
"A man is speaking."	"A man is spitting."	0.636

Inference Tasks

MNLI, Multi-Genre Natural Language Inference Corpus

- Crowd-sourced sentence pairs with entailment annotations: entailment (0), contradiction (2) and neutral (1)
- Task: produce these three labels and you get an accuracy score

"One of our number will carry out your instructions minutely."	"A member of my team will execute your orders with immense precision."	0
"Fun for adults and children."	"Fun for only children."	2
"yeah well you're a student right"	"Well you're a mechanics student right?"	1
"Get individuals to invest their time and the funding will follow."	"If individuals will invest their time, funding will come along, too."	0

Inference Tasks

QNLI, Stanford Question Answering Dataset

- Based on SQuAD. Task converted into sentence pair classification by forming a pair between each question and each sentence in the corresponding context, and filtering out pairs with low lexical overlap between the question and the context sentence.

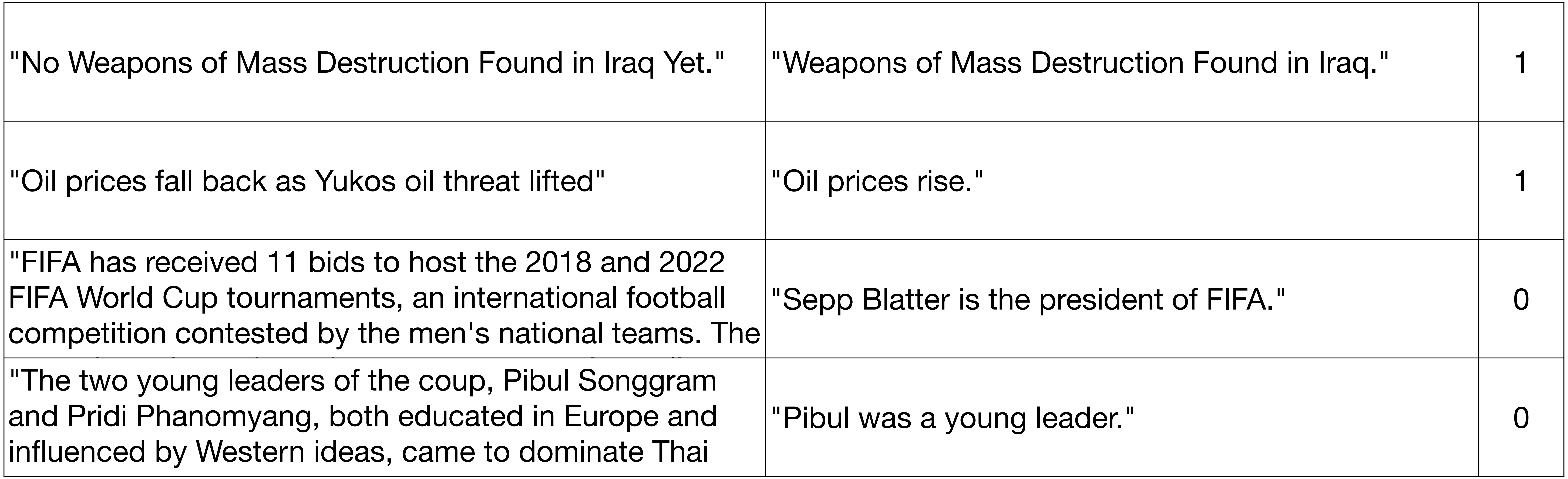
- The task is to predict yes (0) or no (1), the context sentence contains the answer to the question.

"When did the third Digimon series begin?"	"Unlike the two seasons before it and most of the seasons that followed, Digimon Tamers takes a darker and more realistic approach to its story	1
"What is the name of the village 9 miles north of Calafat where the Ottoman forces attacked the Russians?"	"On 31 December 1853, the Ottoman forces at Calafat moved against the Russian force at Chetatea or Cetate, a small village nine miles north of Calafat,	0
"How were the Portuguese expelled from Myanmar?"	"From the 1720s onward, the kingdom was beset with repeated Meithei raids into Upper Myanmar and a nagging rebellion in Lan Na."	1
"Which collection of minor poems are sometimes attributed to Virgil?"	"A number of minor poems, collected in the Appendix Vergiliana, are sometimes attributed to him."	0

Inference Tasks

RTE, Recognizing Textual Entailment datasets

- The task is to predict entailment (0) or not_entailment (1) between pairs of sentences



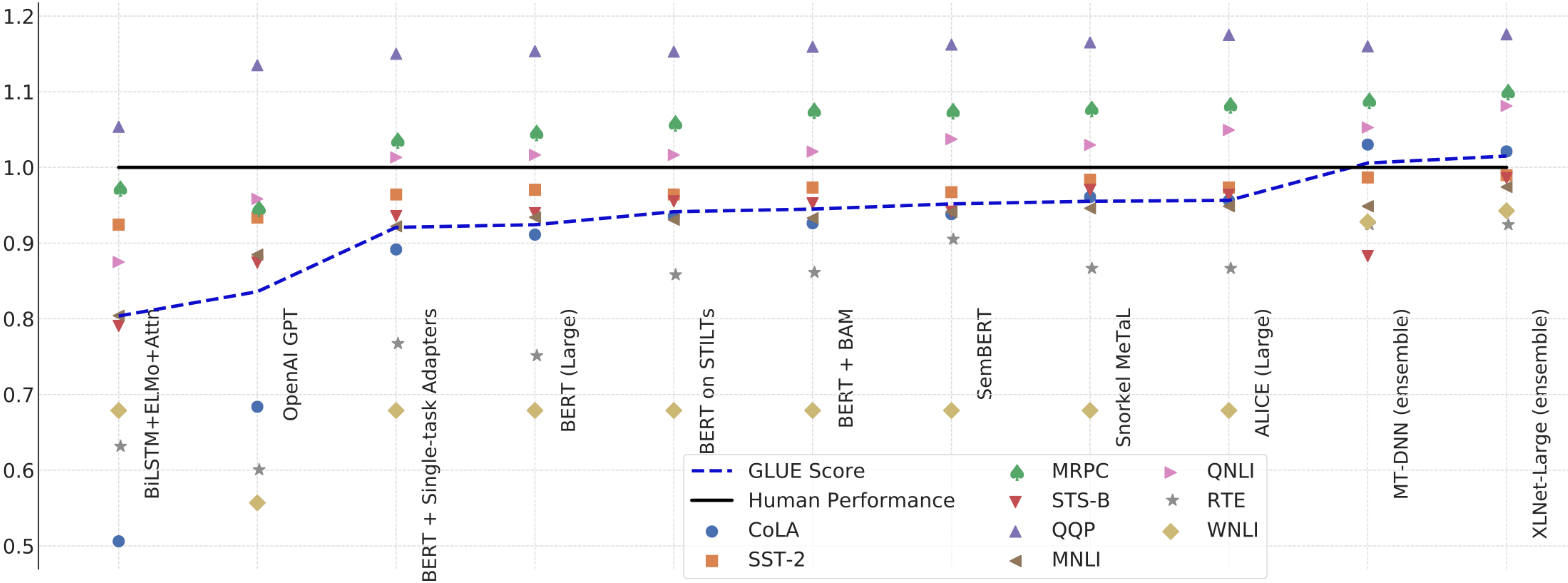
"No Weapons of Mass Destruction Found in Iraq Yet."	"Weapons of Mass Destruction Found in Iraq."	1
"Oil prices fall back as Yukos oil threat lifted"	"Oil prices rise."	1
"FIFA has received 11 bids to host the 2018 and 2022 FIFA World Cup tournaments, an international football competition contested by the men's national teams. The	"Sepp Blatter is the president of FIFA."	0
"The two young leaders of the coup, Pibul Songgram and Pridi Phanomyang, both educated in Europe and influenced by Western ideas, came to dominate Thai	"Pibul was a young leader."	0

Inference Tasks

WNLI, Winograd Schema Challenge

- Reading comprehension task to identify referent of a pronoun using entailment between two sentences (one has pronoun reference explicit)
- Predict 1 (entailment) or 0 (not_entailment)
- Designed to fool simple statistical techniques.
- Test set is imbalanced (65% not entailment) and dev set is adversarial (memorization will hurt performance)

"I stuck a pin through a carrot. When I pulled the pin out, it had a hole."	"The carrot had a hole."	1
"George got free tickets to the play, but he gave them to Eric, because he was particularly eager to see it."	"George was particularly eager to see it."	0

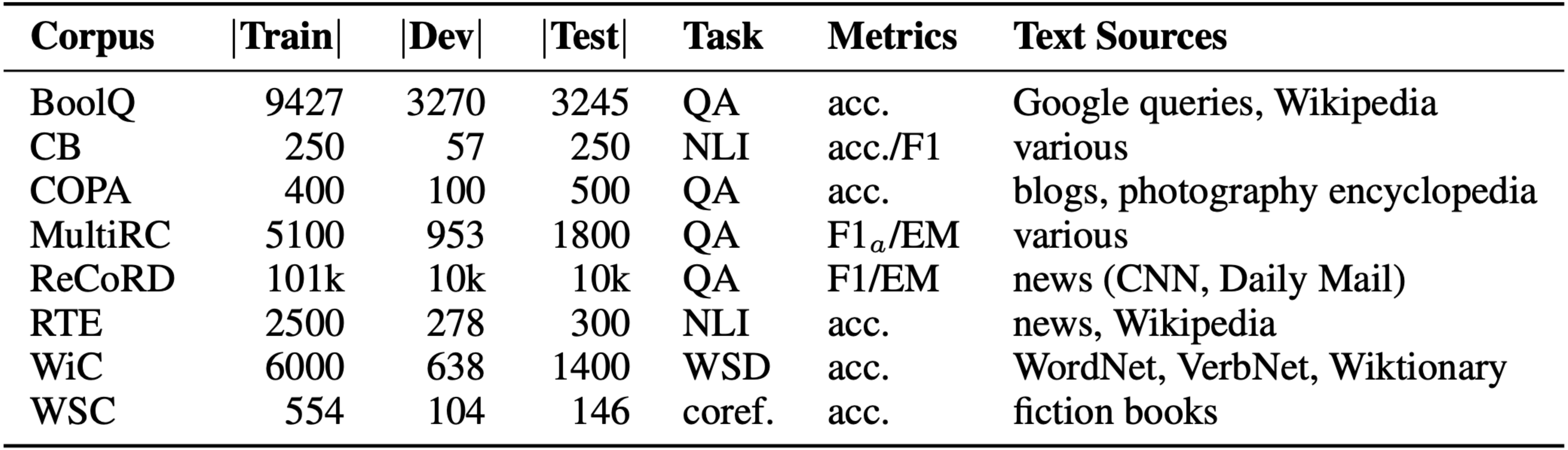

1.2
1.1
1.0
0.9
0.8
0.7
0.6
0.5
BiLSTM+ELMo+Attn
OpenAI GPT
BERT + Single-task Adapters
BERT (Large)
BERT on STILTs
BERT + BAM
SemBERT
Snorkel MeTaL
ALICE (Large)
MT-DNN (ensemble)
XLNet-Large (ensemble)
GLUE Score
Human Performance
CoLA
SST-2
MRPC
STS-B
QQP
MNLI
QNLI
RTE
WNLI

SuperGLUE

SuperGLUE

https://super.gluebenchmark.com/

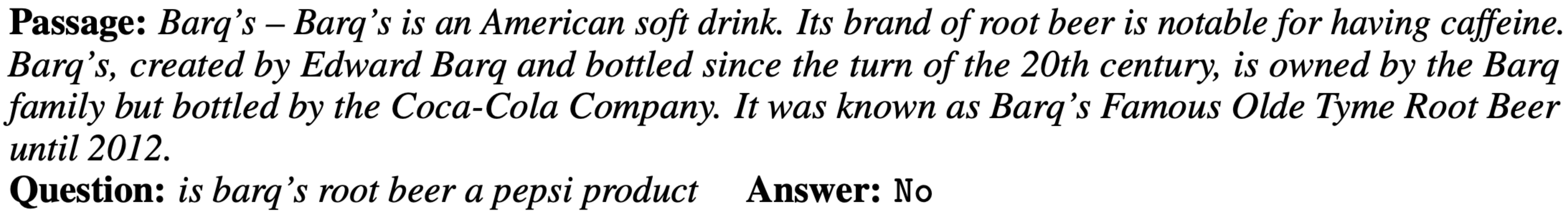
- GLUE was too easy for LLMs after July 2019
- SuperGLUE is a more rigorous test of NLU



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	$\mathrm{F1}_a$/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

BoolQ

Boolean Questions, QA task with yes/no answers

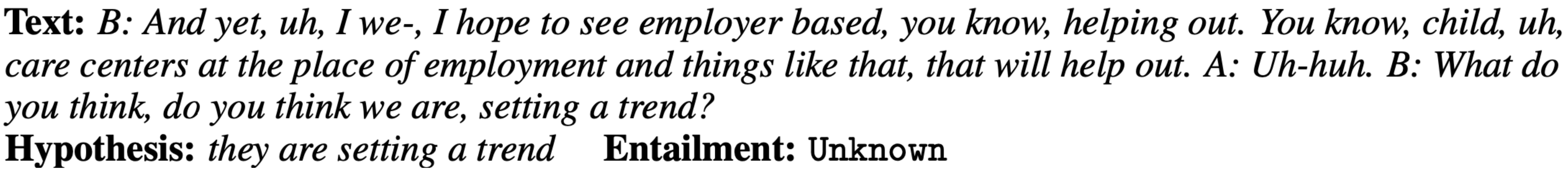


Passage: *Barq's – Barq's is an American soft drink. Its brand of root beer is notable for having caffeine. Barq's, created by Edward Barq and bottled since the turn of the 20th century, is owned by the Barq family but bottled by the Coca-Cola Company. It was known as Barq's Famous Olde Tyme Root Beer until 2012.*

Question: *is barq's root beer a pepsi product* **Answer:** `No`

CB

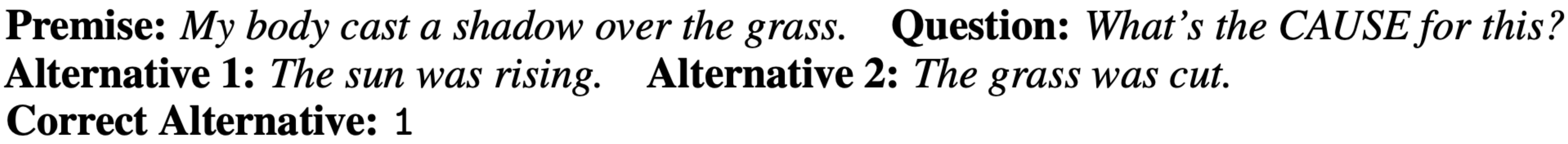
CommitmentBank, is author committed to truth of embedded clause?



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`

COPA

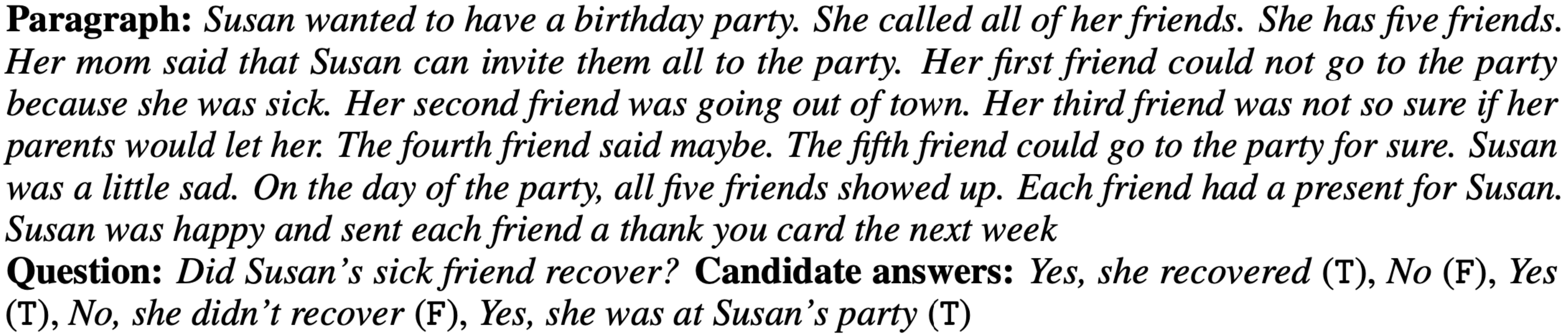
Choice of Plausible Alternatives, causal reasoning task



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

MultiRC

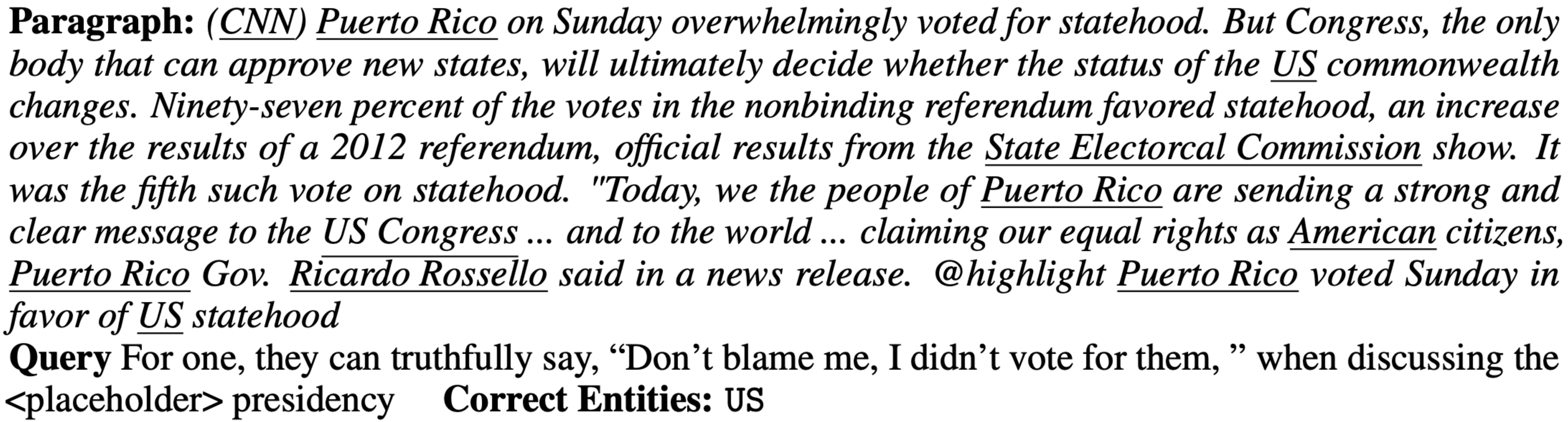
Multi-Sentence Reading Comprehension



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`)

ReCoRD

Reading Comprehension with Commonsense Reasoning Dataset

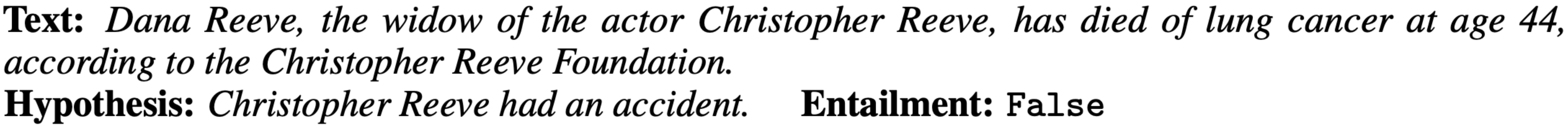


Paragraph: *(CNN) Puerto Rico on Sunday overwhelmingly voted for statehood. But Congress, the only body that can approve new states, will ultimately decide whether the status of the US commonwealth changes. Ninety-seven percent of the votes in the nonbinding referendum favored statehood, an increase over the results of a 2012 referendum, official results from the State Electorcal Commission show. It was the fifth such vote on statehood. "Today, we the people of Puerto Rico are sending a strong and clear message to the US Congress ... and to the world ... claiming our equal rights as American citizens, Puerto Rico Gov. Ricardo Rossello said in a news release. @highlight Puerto Rico voted Sunday in favor of US statehood*

Query For one, they can truthfully say, “Don’t blame me, I didn’t vote for them, ” when discussing the <placeholder> presidency **Correct Entities:** `US`

RTE

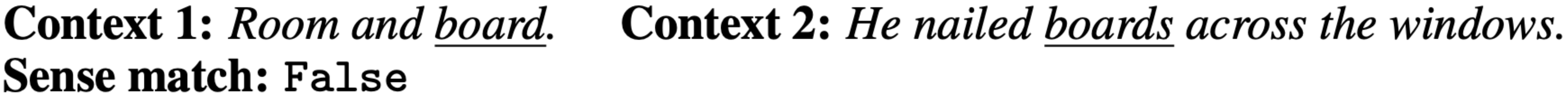
Recognizing Textual Entailment (from GLUE)



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.* **Entailment:** `False`

WiC

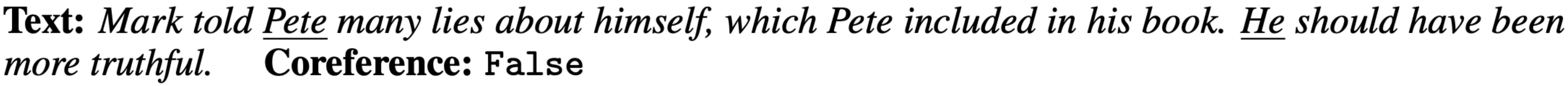
Word in Context, word-sense dataset



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

WSC

Winograd Schema Challenge (from GLUE)



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

Swag and HellaSwag

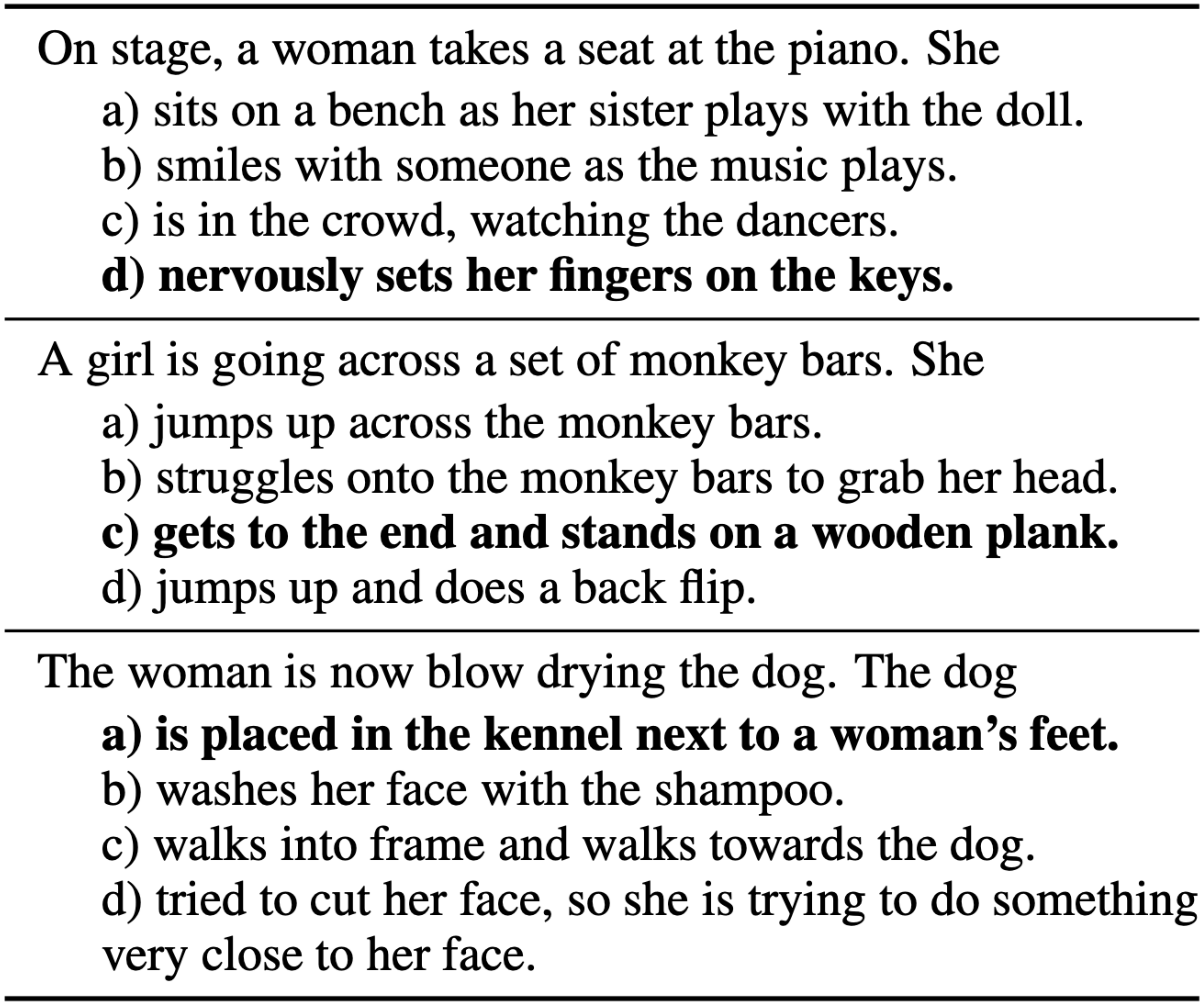
Swag

https://rowanzellers.com/swag/

Large-scale adversarial dataset for grounded commonsense inference

- Given a partial description: "she opened the hood of the car"
- Humans can reason: "then, she examined the engine"
- Dataset constructed using Adversarial Filtering to identify the answers that are meaning-based rather than based on vocabulary overlap
- "on stage, a woman takes a seat at the piano". What next?

Swag



On stage, a woman takes a seat at the piano. She

a) sits on a bench as her sister plays with the doll.
b) smiles with someone as the music plays.
c) is in the crowd, watching the dancers.
d) nervously sets her fingers on the keys.

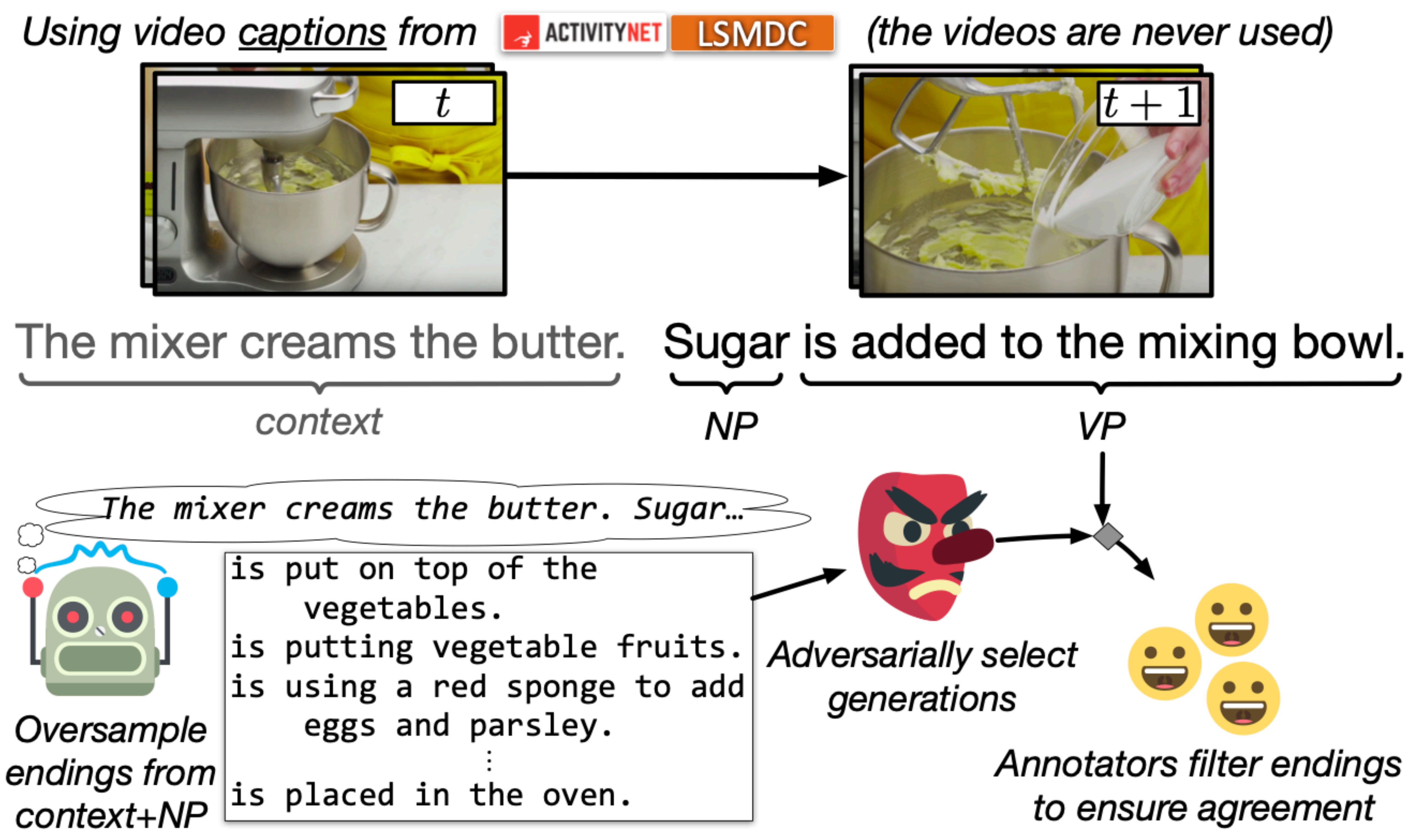
A girl is going across a set of monkey bars. She

a) jumps up across the monkey bars.
b) struggles onto the monkey bars to grab her head.
c) gets to the end and stands on a wooden plank.
d) jumps up and does a back flip.

The woman is now blow drying the dog. The dog

a) is placed in the kennel next to a woman’s feet.
b) washes her face with the shampoo.
c) walks into frame and walks towards the dog.
d) tried to cut her face, so she is trying to do something very close to her face.

Swag



Using video captions from ACTIVITYNET LSMDC (the videos are never used)
t
$t+1$
The mixer creams the butter.
Sugar is added to the mixing bowl.
context
NP
VP
Oversample endings from context+NP
Adversarially select generations
Annotators filter endings to ensure agreement

The mixer creams the butter. Sugar…

```
is put on top of the
    vegetables.
is putting vegetable fruits.
is using a red sponge to add
    eggs and parsley.
           ⋮
is placed in the oven.
```

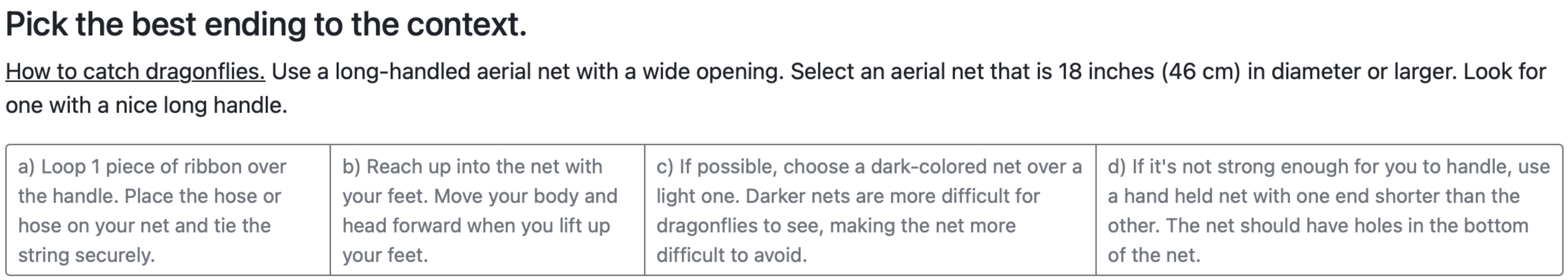
Figure 1: Overview of the data collection process. For a pair of sequential video captions, the second caption is split into noun and verb phrases. A language model generates many negative endings, of which a difficult subset are human-annotated.

Swag

- Soon after release BERT models fine-tuned on Swag obtain 86% accuracy.
- DeBERTa-large gets 94.12% accuracy.

HellaSwag https://rowanzellers.com/hellaswag/

Can a Machine *Really* Finish Your Sentence?

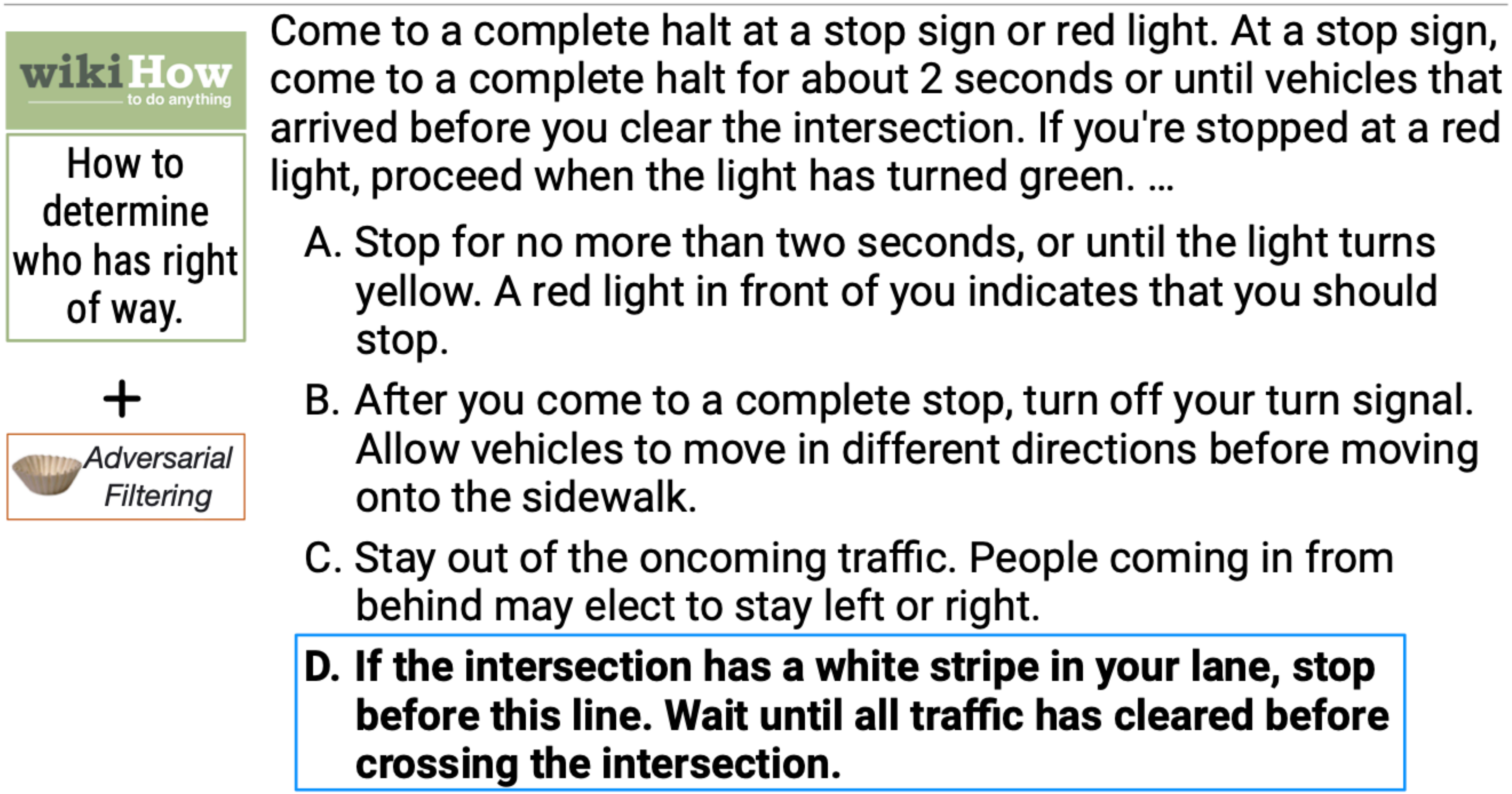


Pick the best ending to the context.

How to catch dragonflies. Use a long-handled aerial net with a wide opening. Select an aerial net that is 18 inches (46 cm) in diameter or larger. Look for one with a nice long handle.

a) Loop 1 piece of ribbon over the handle. Place the hose or hose on your net and tie the string securely.	b) Reach up into the net with your feet. Move your body and head forward when you lift up your feet.	c) If possible, choose a dark-colored net over a light one. Darker nets are more difficult for dragonflies to see, making the net more difficult to avoid.	d) If it's not strong enough for you to handle, use a hand held net with one end shorter than the other. The net should have holes in the bottom of the net.

HellaSwag



wikiHow
to do anything

How to determine who has right of way.

+

Adversarial Filtering

Come to a complete halt at a stop sign or red light. At a stop sign, come to a complete halt for about 2 seconds or until vehicles that arrived before you clear the intersection. If you're stopped at a red light, proceed when the light has turned green. …

A. Stop for no more than two seconds, or until the light turns yellow. A red light in front of you indicates that you should stop.

B. After you come to a complete stop, turn off your turn signal. Allow vehicles to move in different directions before moving onto the sidewalk.

C. Stay out of the oncoming traffic. People coming in from behind may elect to stay left or right.

D. If the intersection has a white stripe in your lane, stop before this line. Wait until all traffic has cleared before crossing the intersection.

HellaSwag

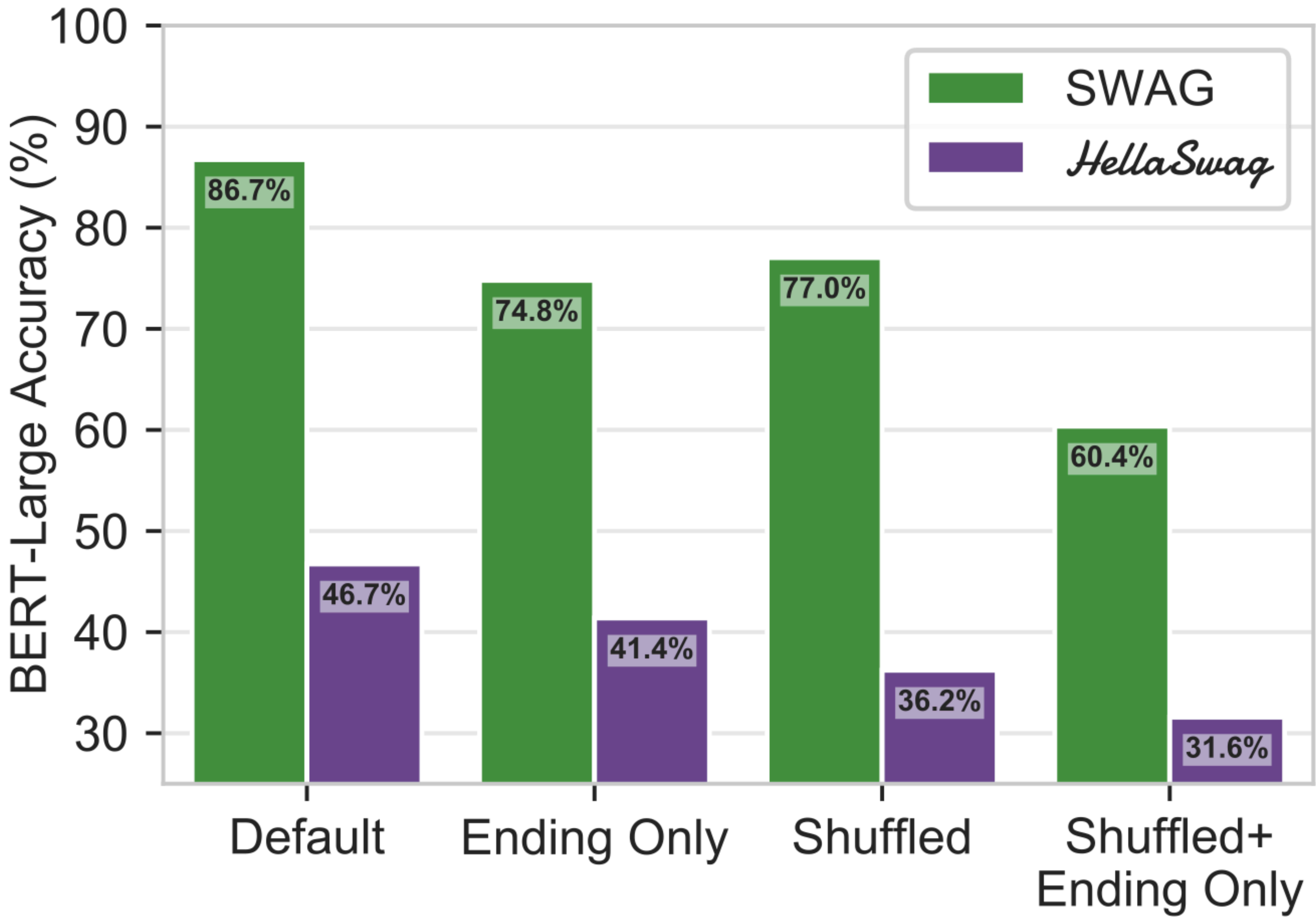


Figure 4: BERT validation accuracy when trained and evaluated under several versions of SWAG, with the new dataset *HellaSwag* as comparison. We compare:

`Ending Only` No context is provided; just the endings.

`Shuffled` Endings that are indidivually tokenized, shuffled, and then detokenized.

`Shuffled+ Ending Only` No context is provided *and* each ending is shuffled.

HellaSwag

- GPT 3 gets 85.5% accuracy (10-shot learning)
- GPT 4 gets 95.3% accuracy (10-shot learning)

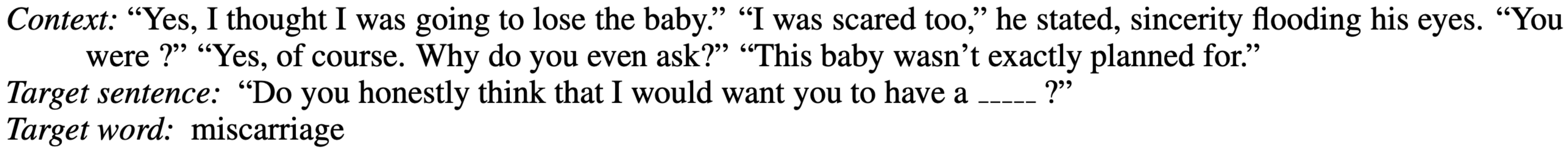
Lambada

Lambada https://zenodo.org/record/2630551

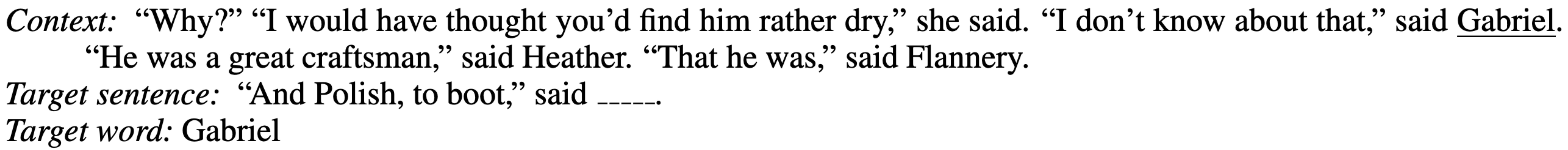
Language modeling broadened to account for discourse aspects

- Collection of narrative passages. Predict the last word.
- Humans can guess the word if they read the whole passage but not otherwise
- The last sentence that contains the word to be predicted is insufficient for prediction
- Training data is 10K passages from BookCorpus.
- Long range dependency evaluated as (last) word prediction

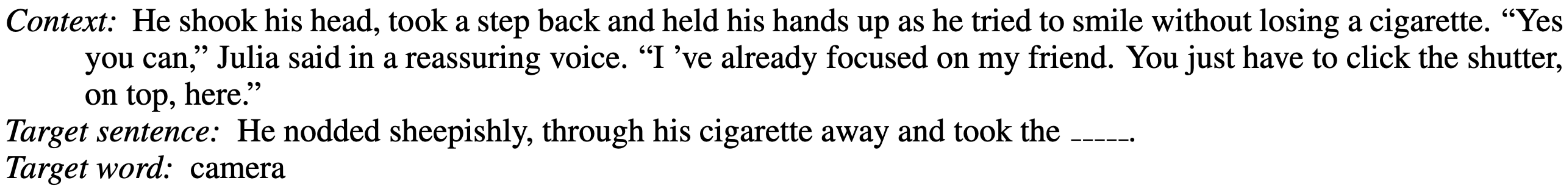
Lambada



Context: “Yes, I thought I was going to lose the baby.” “I was scared too,” he stated, sincerity flooding his eyes. “You were ?” “Yes, of course. Why do you even ask?” “This baby wasn’t exactly planned for.”
Target sentence: “Do you honestly think that I would want you to have a _____ ?”
Target word: miscarriage



Context: “Why?” “I would have thought you’d find him rather dry,” she said. “I don’t know about that,” said Gabriel. “He was a great craftsman,” said Heather. “That he was,” said Flannery.
Target sentence: “And Polish, to boot,” said _____.
Target word: Gabriel



Context: He shook his head, took a step back and held his hands up as he tried to smile without losing a cigarette. “Yes you can,” Julia said in a reassuring voice. “I ’ve already focused on my friend. You just have to click the shutter, on top, here.”
Target sentence: He nodded sheepishly, through his cigarette away and took the _____.
Target word: camera

GPT-2 gets 63.24%

GPT-3 gets 86.24% few-shot

MMLU

https://arxiv.org/abs/2009.03300

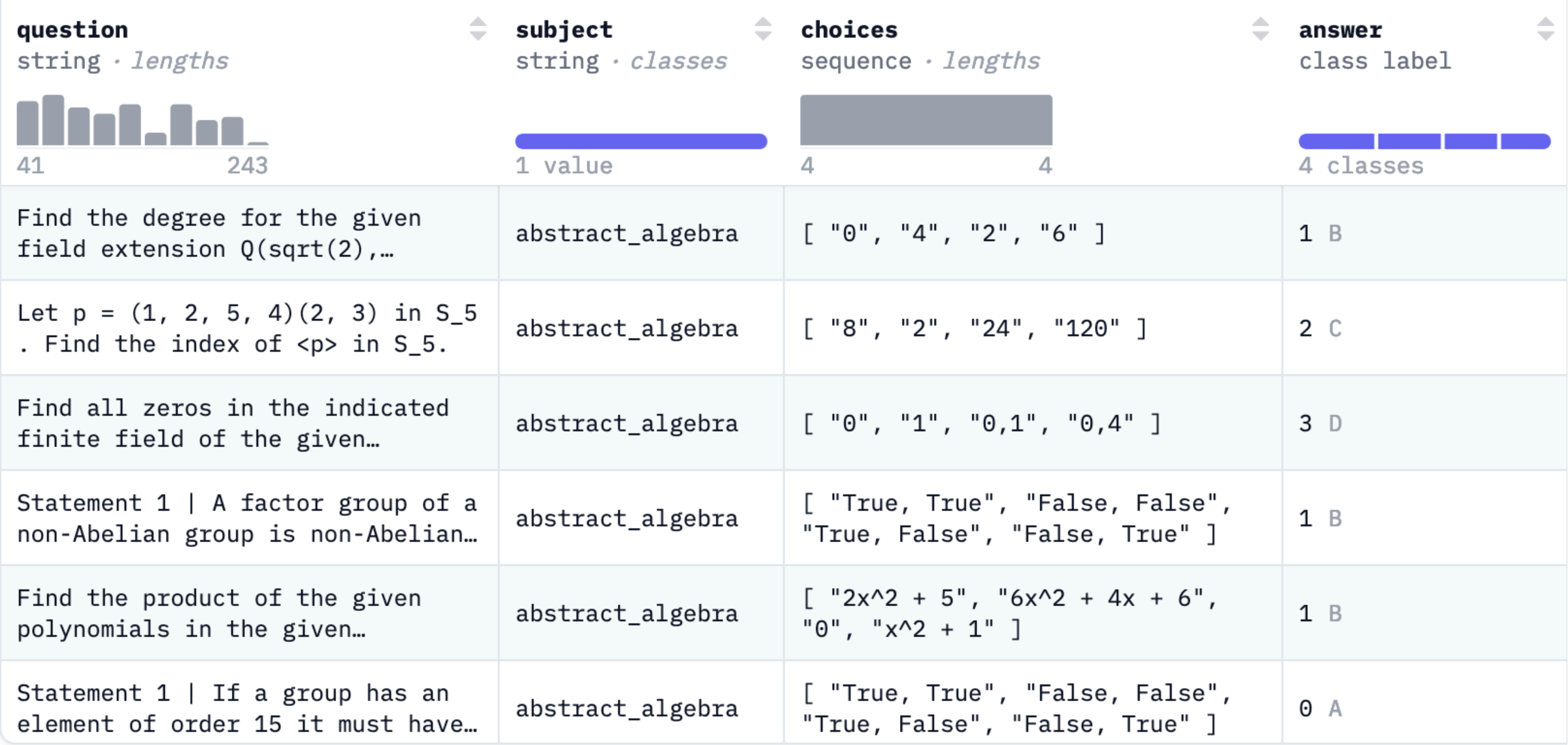
MMLU

Measuring Massive Multitask Language Understanding

- This is a massive multitask test consisting of multiple-choice questions from various branches of knowledge.
- The test spans subjects in the humanities, social sciences, hard sciences and covers 57 tasks including elementary mathematics, US history, computer science, law, and more.
- To attain high accuracy on this test, models must possess extensive world knowledge and problem solving ability.

MMLU

Measuring Massive Multitask Language Understanding



question string · *lengths* 41 243	subject string · *classes* 1 value	choices sequence · *lengths* 4 4	answer class label 4 classes
Find the degree for the given field extension Q(sqrt(2),…	abstract_algebra	["0", "4", "2", "6"]	1 B
Let p = (1, 2, 5, 4)(2, 3) in S_5 . Find the index of <p> in S_5.	abstract_algebra	["8", "2", "24", "120"]	2 C
Find all zeros in the indicated finite field of the given…	abstract_algebra	["0", "1", "0,1", "0,4"]	3 D
Statement 1 | A factor group of a non-Abelian group is non-Abelian…	abstract_algebra	["True, True", "False, False", "True, False", "False, True"]	1 B
Find the product of the given polynomials in the given…	abstract_algebra	["2x^2 + 5", "6x^2 + 4x + 6", "0", "x^2 + 1"]	1 B
Statement 1 | If a group has an element of order 15 it must have…	abstract_algebra	["True, True", "False, False", "True, False", "False, True"]	0 A

MMLU

Measuring Massive Multitask Language Understanding

- Abstract Algebra
- Anatomy
- Astronomy
- Business Ethics
- Clinical Knowledge
- College
    - Biology
    - Chemistry
    - Comp Sci
    - Mathematics
    - Medicine
    - Physics
- Computer Security
- Conceptual Physics
- Econometrics
- Electrical Engineering
- High School
    - Biology
    - Chemistry
    - Comp Sci
    - European History
    - Geography
    - Gov't and Politics
    - Macroeconomics
    - Mathematics
    - Microeconomics
    - Physics
    - Psychology
    - Statistics
    - US History
    - World History
- Elementary Mathematics
- Formal Logic
- Global Facts
- Human Aging
- Human Sexuality
- International Law
- Jurisprudence
- Logical Fallacies
- Machine Learning
- Management
- Marketing
- Medical Genetics
- Miscellaneous
- Moral Disputes
- Moral Scenarios
- Nutrition
- Philosophy
- Prehistory
- Professional
    - Accounting
    - Law
    - Medicine
    - Psychology
- Public Relations
- Security Studies
- Sociology
- US Foreign Policy
- Virology
- World Religions

MTEB

https://arxiv.org/abs/2210.07316

MTEB

Massive Text Embedding Benchmark

- MTEB spans 8 embedding tasks covering a total of 58 datasets and 112 languages.
- The benchmark comprises 8 different tasks, with up to 15 datasets each.
- Of the 58 total datasets in MTEB, 10 are multilingual, covering 112 different languages
- Sentence-level and paragraph-level datasets are included to contrast performance on short and long texts. (tasks are split into S2S; P2P; S2P)
- New datasets for existing tasks can be benchmarked in MTEB via a single file that specifies the task and a Hugging Face dataset name

MTEB

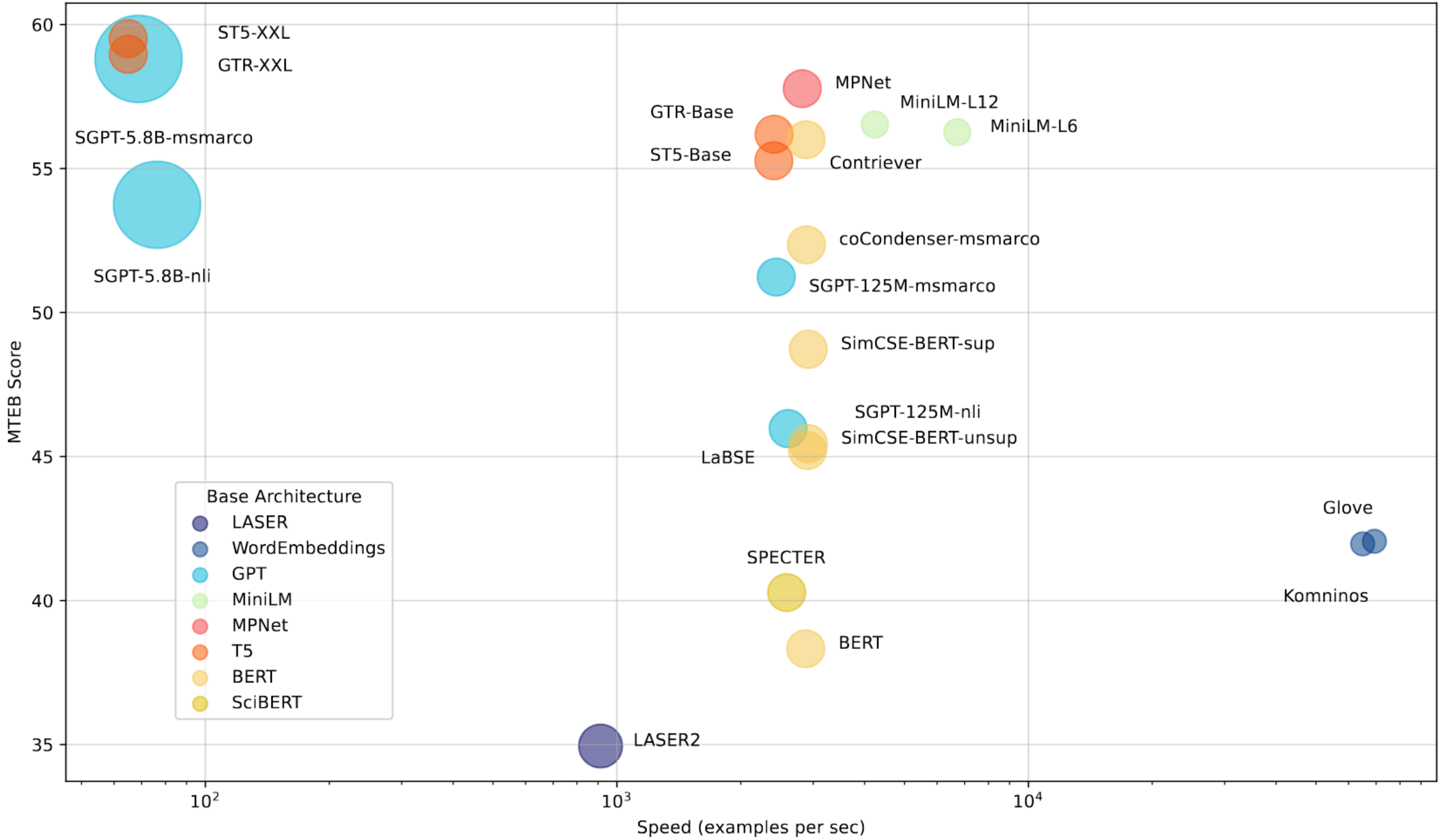
Massive Text Embedding Benchmark

- Tasks
  - **Bitext mining**: Inputs are two sets of sentences from two different languages. For each sentence in the first set, the best match in the second set needs to be found.
  - **Classification**: A train and test set are embedded with the provided model. The train set embeddings are used to train a logistic regression classifier
  - **Clustering**: Given a set of sentences or paragraphs, the goal is to group them into meaningful clusters.
  - **Pair Classification**: A pair of text inputs is provided and a label needs to be assigned.
  - **Reranking**: Inputs are a query and a list of relevant and irrelevant reference texts. The aim is to rank the results according to their relevance to the query

MTEB

Massive Text Embedding Benchmark

- Tasks
  - **Retrieval**: Each dataset consists of a corpus, queries and a mapping for each query to relevant documents from the corpus. The aim is to find these relevant documents.
  - **Semantic Textual Similarity** (STS): Given a sentence pair the aim is to determine their similarity. Labels are continuous scores with higher numbers indicating more similar sentences
  - **Summarization**: A set of human-written and machine-generated summaries are provided. The aim is to score the machine summaries.



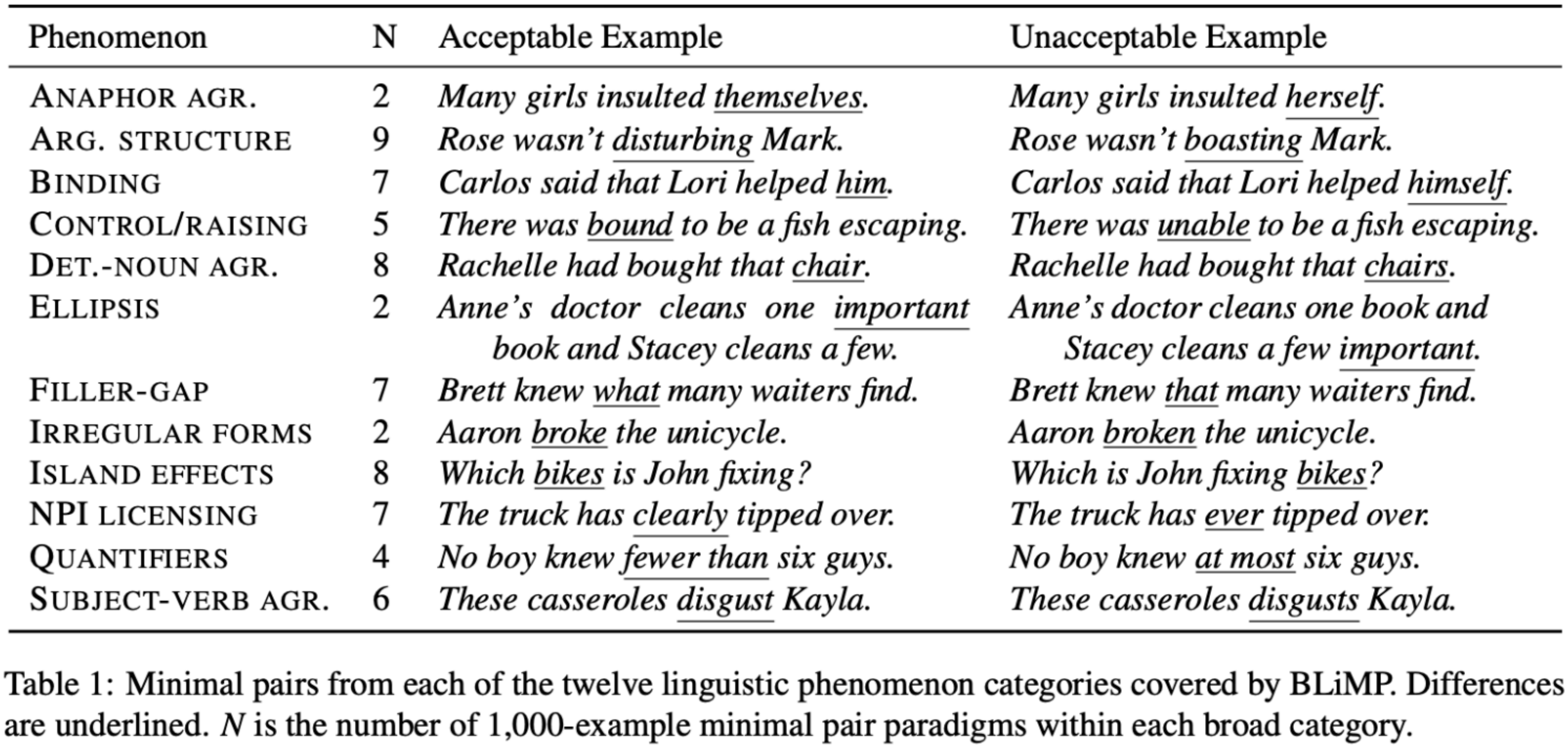
ST5-XXL
GTR-XXL
SGPT-5.8B-msmarco
SGPT-5.8B-nli
MPNet
MiniLM-L12
GTR-Base
MiniLM-L6
ST5-Base
Contriever
coCondenser-msmarco
SGPT-125M-msmarco
SimCSE-BERT-sup
SGPT-125M-nli
SimCSE-BERT-unsup
LaBSE
Glove
Komninos
SPECTER
BERT
LASER2
Base Architecture
LASER
WordEmbeddings
GPT
MiniLM
MPNet
T5
BERT
SciBERT
MTEB Score
60
55
50
45
40
35
10^2
10^3
10^4
Speed (examples per sec)

Other Datasets

Other Datasets

- MMLU - Multiple choice questions in 57 subjects (professional & academic)
- MTEB - Evaluation of sentence embeddings (58 datasets; 112 languages)
- ARC - AI2 Reasoning Challenge, grade school multiple choice science quiz
- WinoGrande - bigger version of the Winograd Schema Challenge from GLUE
- DROP - Reading comprehension & arithmetic
- GSM-8K - Grade school math questions
- StoryCloze - similar to Lambada with longer context
- BLiMP - Benchmark of minimal linguistic pairs; new version of CoLA
- Some shared tasks combine different benchmarks to make bigger ones

BLiMP



Phenomenon	N	Acceptable Example	Unacceptable Example
ANAPHOR AGR.	2	*Many girls insulted themselves.*	*Many girls insulted herself.*
ARG. STRUCTURE	9	*Rose wasn't disturbing Mark.*	*Rose wasn't boasting Mark.*
BINDING	7	*Carlos said that Lori helped him.*	*Carlos said that Lori helped himself.*
CONTROL/RAISING	5	*There was bound to be a fish escaping.*	*There was unable to be a fish escaping.*
DET.-NOUN AGR.	8	*Rachelle had bought that chair.*	*Rachelle had bought that chairs.*
ELLIPSIS	2	*Anne's doctor cleans one important book and Stacey cleans a few.*	*Anne's doctor cleans one book and Stacey cleans a few important.*
FILLER-GAP	7	*Brett knew what many waiters find.*	*Brett knew that many waiters find.*
IRREGULAR FORMS	2	*Aaron broke the unicycle.*	*Aaron broken the unicycle.*
ISLAND EFFECTS	8	*Which bikes is John fixing?*	*Which is John fixing bikes?*
NPI LICENSING	7	*The truck has clearly tipped over.*	*The truck has ever tipped over.*
QUANTIFIERS	4	*No boy knew fewer than six guys.*	*No boy knew at most six guys.*
SUBJECT-VERB AGR.	6	*These casseroles disgust Kayla.*	*These casseroles disgusts Kayla.*

Table 1: Minimal pairs from each of the twelve linguistic phenomenon categories covered by BLiMP. Differences are underlined. *N* is the number of 1,000-example minimal pair paradigms within each broad category.

Question Answering

- Natural Questions
- Web Questions
- TriviaQA
- COQA - Conversational Question Answering Challenge (https://stanfordnlp.github.io/coqa/)

Summarization

- CNN-Daily Mail
- BookSum
- NYT dataset
- DUC dataset (from NIST)

Natural Language Generation

- E2E
- WebNLG
- DART

Translation

- FLORES-200
- WMT 2014
- IWSLT shared tasks (translated TED talks)
- OPUS-MT (subtitles)
- LORELEI (low-resource languages)

Structured Prediction

- CoNLL-2000 to CoNLL 2018 shared tasks
- Entity linking to Wikipedia: AIDA-CoNLL
- Entity linking to other ontologies like UMLS (biomedical): MedMentions
- Dependency tree and Phrase Structure Parsing (Penn Treebank; PropBank)
- Morphological analysis of word constructions from morphemes (Morphophon)
- Wikipron (word pronunciation modelling for 165 languages)