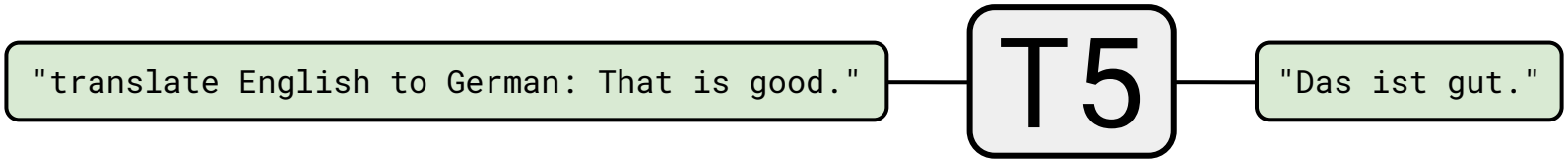
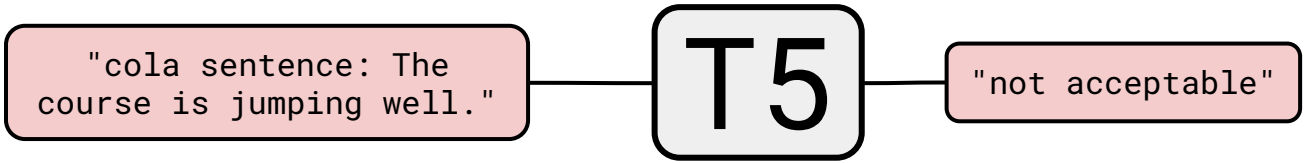
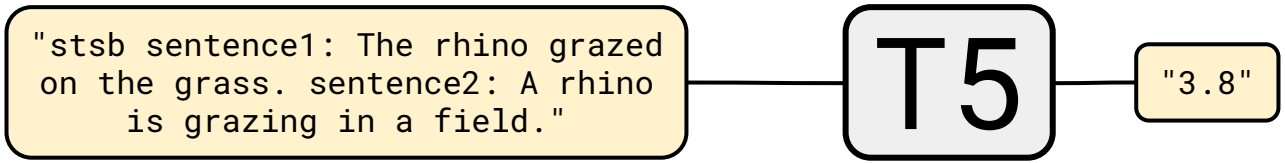
The benefits of unified frameworks for language understanding

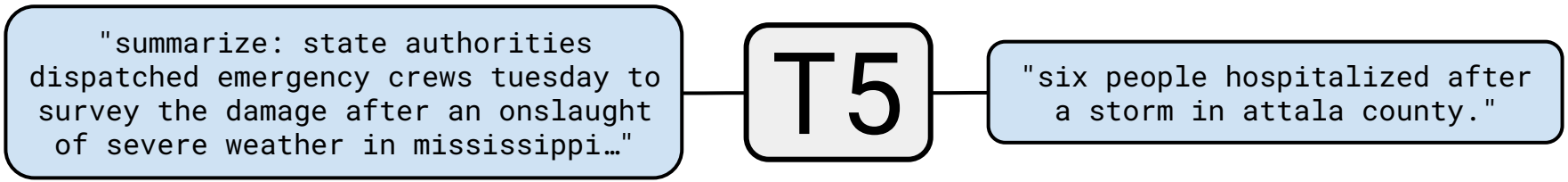
Colin Raffel
UNC Chapel Hill and Hugging Face

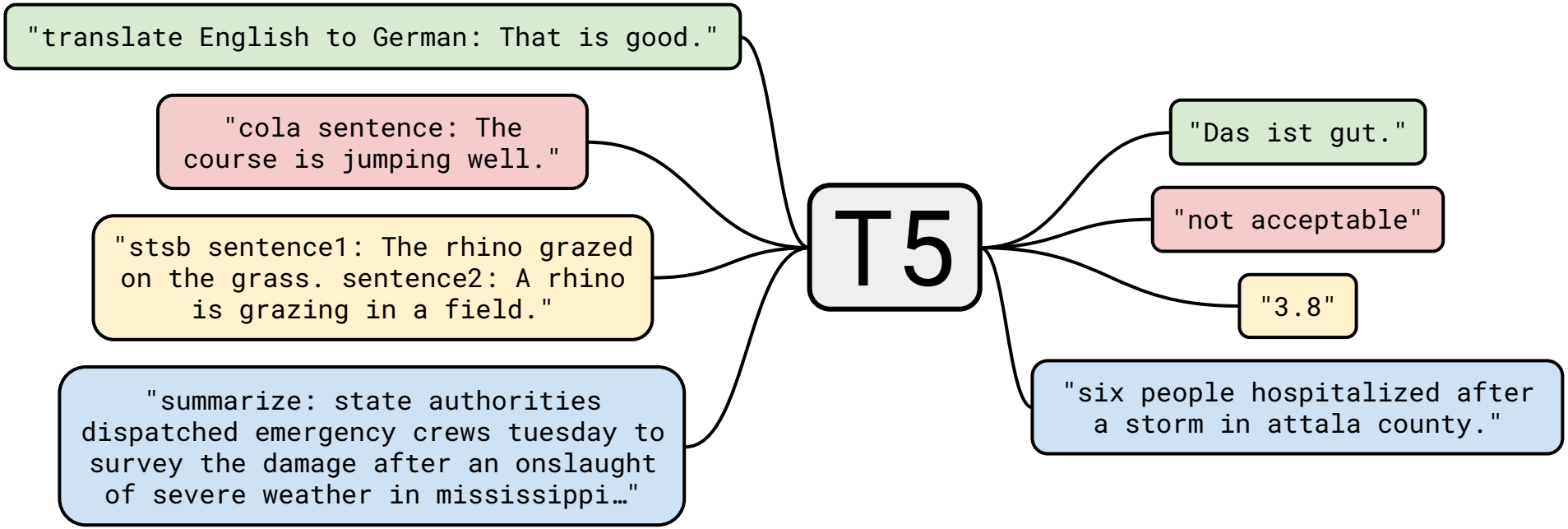
*T*ext-*t*o-*T*ext
*T*ransfer
*T*ransformer


"translate English to German: That is good."
T5
"Das ist gut."


"cola sentence: The course is jumping well."
T5
"not acceptable"


"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."
T5
"3.8"


"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi…"
T5
"six people hospitalized after a storm in attala county."


"translate English to German: That is good."
"cola sentence: The course is jumping well."
"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."
"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi…"
T5
"Das ist gut."
"not acceptable"
"3.8"
"six people hospitalized after a storm in attala county."

How to Fine-Tune BERT for Text Classification?

Chi Sun, Xipeng Qiu*, Yige Xu, Xuanjing Huang
Shanghai Key Laboratory of Intelligent Information Processing, Fudan University
School of Computer Science, Fudan University
825 Zhangheng Road, Shanghai, China
{sunc17,xpqiu,ygxu18,xjhuang}@fudan.edu.cn

On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines

Marius Mosbach
Spoken Language Systems (LSV)
Saarland Informatics Campus, Saarland University
mmosbach@lsv.uni-saarland.de

Maksym Andriushchenko
Theory of Machine Learning Lab
École polytechnique fédérale de Lausanne
maksym.andriushchenko@epfl.ch

Dietrich Klakow
Spoken Language Systems (LSV)
Saarland Informatics Campus, Saarland University
dietrich.klakow@lsv.uni-saarland.de

Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping

Jesse Dodge[1,2] **Gabriel Ilharco**[3] **Roy Schwartz**[2,3] **Ali Farhadi**[2,3,4] **Hannaneh Hajishirzi**[2,3] **Noah Smith**[2,3]

Revisiting Few-sample BERT Fine-tuning

Tianyi Zhang*[△,§] **Felix Wu***[†] **Arzoo Katiyar**[△,◇] **Kilian Q. Weinberger**[†,‡] **Yoav Artzi**[†,‡]
[†]ASAPP Inc. [§]Stanford University [◇]Penn State University [‡]Cornell University
tz58@stanford.edu {fwu, kweinberger, yoav}@asapp.com arzoo@psu.edu

The Natural Language Decathlon: Multitask Learning as Question Answering

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, Richard Socher
Salesforce Research
{bmccann,nkeskar,cxiong,rsocher}@salesforce.com

Examples

Question	Context	Answer
What is a major importance of Southern California in relation to California and the US?	...Southern California is a major economic center for the state of California and the US....	major economic center
What is the translation from English to German?	Most of the planet is ocean water.	Der Großteil der Erde ist Meerwasser
What is the summary?	Harry Potter star Daniel Radcliffe gains access to a reported £320 million fortune...	Harry Potter star Daniel Radcliffe gets £320M fortune...
Hypothesis: Product and geography are what make cream skimming work. Entailment, neutral, or contradiction?	Premise: Conceptually cream skimming has two basic dimensions – product and geography.	Entailment
Is this sentence positive or negative?	A stirring, funny and finally transporting re-imagining of Beauty and the Beast and 1930s horror film.	positive

Question	Context	Answer
What has something experienced?	Areas of the Baltic that have experienced eutrophication.	eutrophication
Who is the illustrator of Cycle of the Werewolf?	Cycle of the Werewolf is a short novel by Stephen King, featuring illustrations by comic book artist Bernie Wrightson.	Bernie Wrightson
What is the change in dialogue state?	Are there any Eritrean restaurants in town?	food: Eritrean
What is the translation from English to SQL?	The table has column names... Tell me what the notes are for South Australia	SELECT notes from table WHERE 'Current Slogan' = 'South Australia'
Who had given help? Susan or Joan?	Joan made sure to thank Susan for all the help she had given.	Susan

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  cheese =>                           ← prompt
```

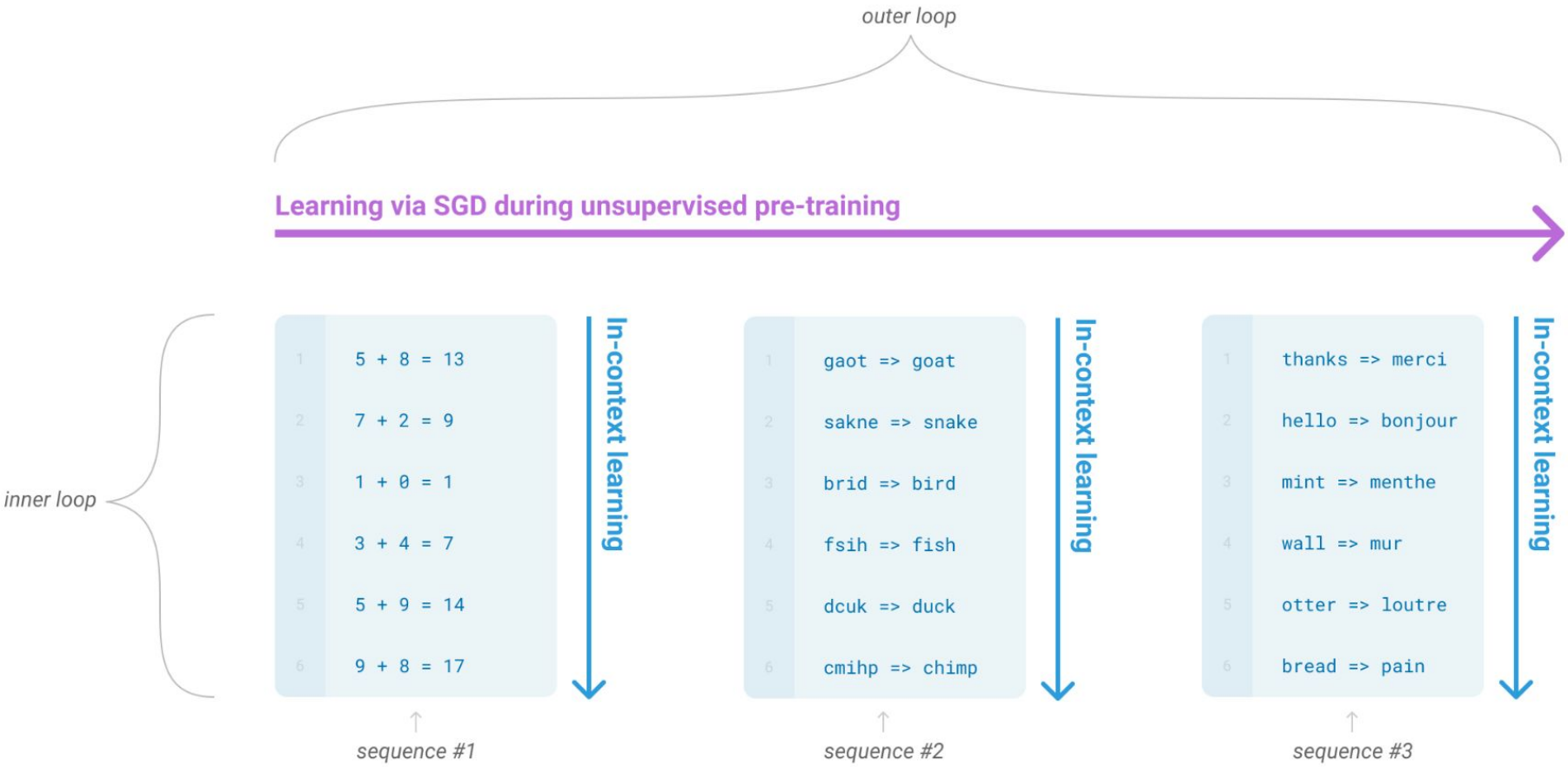
One-shot

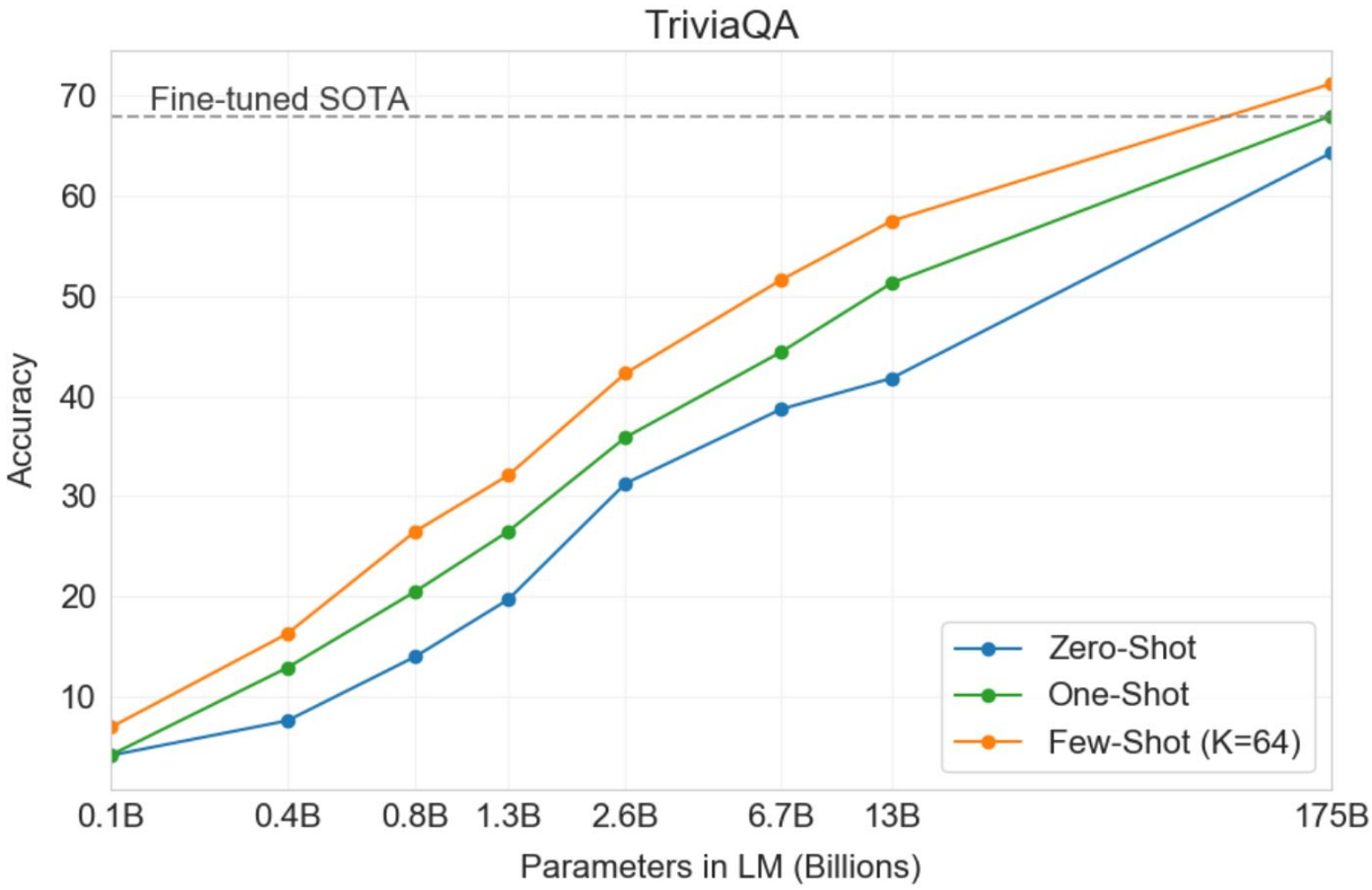
In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← example
3  cheese =>                           ← prompt
```

Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.


outer loop
Learning via SGD during unsupervised pre-training
inner loop
1 5 + 8 = 13
2 7 + 2 = 9
3 1 + 0 = 1
4 3 + 4 = 7
5 5 + 9 = 14
6 9 + 8 = 17
In-context learning
sequence #1
1 gaot => goat
2 sakne => snake
3 brid => bird
4 fsih => fish
5 dcuk => duck
6 cmihp => chimp
In-context learning
sequence #2
1 thanks => merci
2 hello => bonjour
3 mint => menthe
4 wall => mur
5 otter => loutre
6 bread => pain
In-context learning
sequence #3


TriviaQA
Fine-tuned SOTA
70
60
50
40
30
20
10
Accuracy
Zero-Shot
One-Shot
Few-Shot (K=64)
0.1B
0.4B
0.8B
1.3B
2.6B
6.7B
13B
175B
Parameters in LM (Billions)

1. In what year was the first-ever Wimbledon Championship held? **Answer: 1877.**

2. Hg is the chemical symbol of which element? **Answer: Mercury.**

3. Which email service is owned by Microsoft? **Answer: Hotmail.**

4. Which country produces the most coffee in the world? **Answer: Brazil.**

5. In which city was Jim Morrison buried? **Answer: Paris.**

6. Which song by Luis Fonsi and Daddy Yankee has the most views (of all time) on YouTube? **Answer: "Despacito."**

7. What was the first state? **Answer: Delaware.**

8. What is the capital city of Spain? **Answer: Madrid.**

9. What is the painting "La Gioconda" more usually known as? **Answer: The Mona Lisa.**

from https://www.scarymommy.com/best-trivia-questions-answers/

A Mathematical Exploration of Why Language Models Help Solve Downstream Tasks

Nikunj Saunshi, Sadhika Malladi & Sanjeev Arora
Princeton University
{nsaunshi,smalladi,arora}@cs.princeton.edu

Theorem 4.1. *Let $\{\boldsymbol{p}_{\cdot|s}\}$ be a language model that is ϵ-optimal, i.e. $\ell_{xent}(\{\boldsymbol{p}_{\cdot|s}\}) - \ell^*_{xent} \leq \epsilon$, for some $\epsilon > 0$. For a classification task $\mathcal{T}$ that is (τ, B)-natural, we have*

$$\ell_{\mathcal{T}}\left(\{\boldsymbol{p}_{\cdot|s}\}\right) \leq \tau + \sqrt{2B^2\epsilon\left(\gamma(p_{\mathcal{T}})\right)^{-1}}$$

UNIFIEDQA: Crossing Format Boundaries with a Single QA System

Daniel Khashabi[1] **Sewon Min**[2] **Tushar Khot**[1] **Ashish Sabharwal**[1]
Oyvind Tafjord[1] **Peter Clark**[1] **Hannaneh Hajishirzi**[1,2]

[1]Allen Institute for AI, Seattle, U.S.A.
[2]University of Washington, Seattle, U.S.A.

Extractive [SQuAD]

Question: At what speed did the turbine operate?
Context: (Nikola_Tesla) On his 50th birthday in 1906, Tesla demonstrated his 200 horsepower (150 kilowatts) 16,000 rpm bladeless turbine. ...
Gold answer: 16,000 rpm

Abstractive [NarrativeQA]

Question: What does a drink from narcissus's spring cause the drinker to do?
Context: Mercury has awakened Echo, who weeps for Narcissus, and states that a drink from Narcissus's spring causes the drinkers to "Grow dotingly enamored of themselves." ...
Gold answer: fall in love with themselves

Multiple-Choice [ARC-challenge]

Question: What does photosynthesis produce that helps plants grow?
Candidate Answers: (A) water (B) oxygen (C) protein (D) sugar
Gold answer: sugar

Yes/No [BoolQ]

Question: Was America the first country to have a president?
Context: (President) The first usage of the word president to denote the highest official in a government was during the Commonwealth of England ...
Gold answer: no

Measuring Massive Multitask Language Understanding

Dan Hendrycks
UC Berkeley

Collin Burns
Columbia University

Steven Basart
UChicago

Andy Zou
UC Berkeley

Mantas Mazeika
UIUC

Dawn Song
UC Berkeley

Jacob Steinhardt
UC Berkeley

Few Shot Prompt and Predicted Answer

The following are multiple choice questions about high school mathematics.

How many numbers are in the list 25, 26, ..., 100?
(A) 75 (B) 76 (C) 22 (D) 23
Answer: B

Compute $i + i^2 + i^3 + \cdots + i^{258} + i^{259}$.
(A) -1 (B) 1 (C) i (D) $-i$
Answer: A

If 4 daps = 7 yaps, and 5 yaps = 3 baps, how many daps equal 42 baps?
(A) 28 (B) 21 (C) 40 (D) 30
Answer: C

Model	Humanities	Social Science	STEM	Other	Average
Random Baseline	25.0	25.0	25.0	25.0	25.0
RoBERTa	27.9	28.8	27.0	27.7	27.9
ALBERT	27.2	25.7	27.7	27.9	27.1
GPT-2	32.8	33.3	30.2	33.1	32.4
UnifiedQA	45.6	56.6	40.2	54.6	48.9
GPT-3 Small (few-shot)	24.4	30.9	26.0	24.1	25.9
GPT-3 Medium (few-shot)	26.1	21.6	25.6	25.5	24.9
GPT-3 Large (few-shot)	27.1	25.6	24.3	26.5	26.0
GPT-3 X-Large (few-shot)	40.8	50.4	36.7	48.8	43.9

Muppet: Massive Multi-task Representations with Pre-Finetuning

Armen Aghajanyan
Facebook
armenag@fb.com

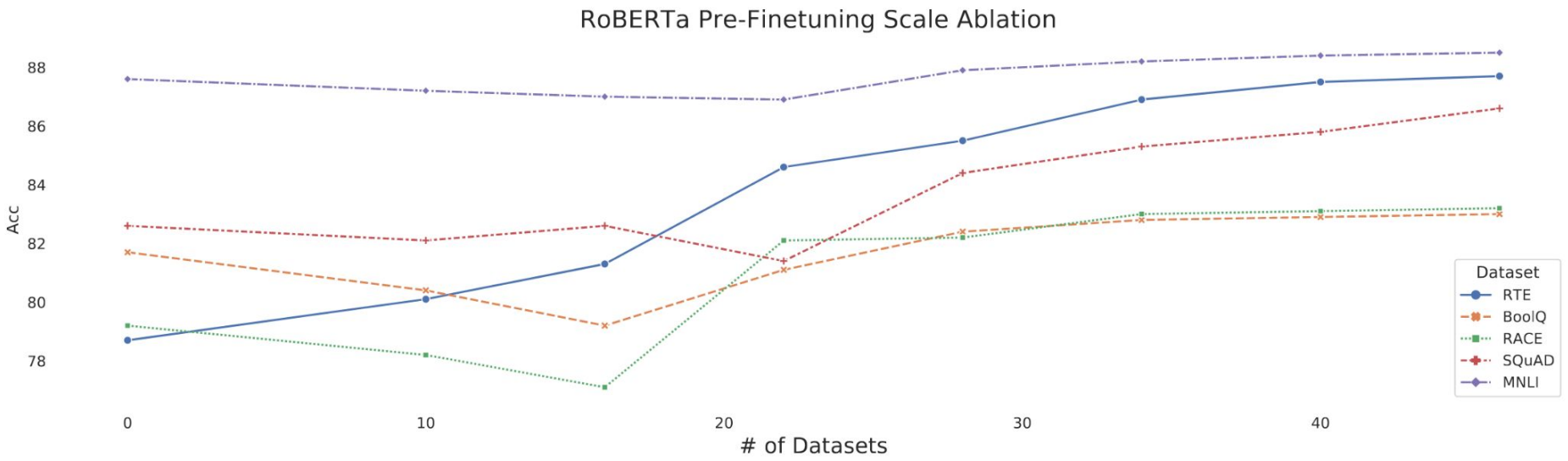
Anchit Gupta
Facebook
anchit@fb.com

Akshat Shrivastava
Facebook
akshats@fb.com

Xilun Chen
Facebook
xilun@fb.com

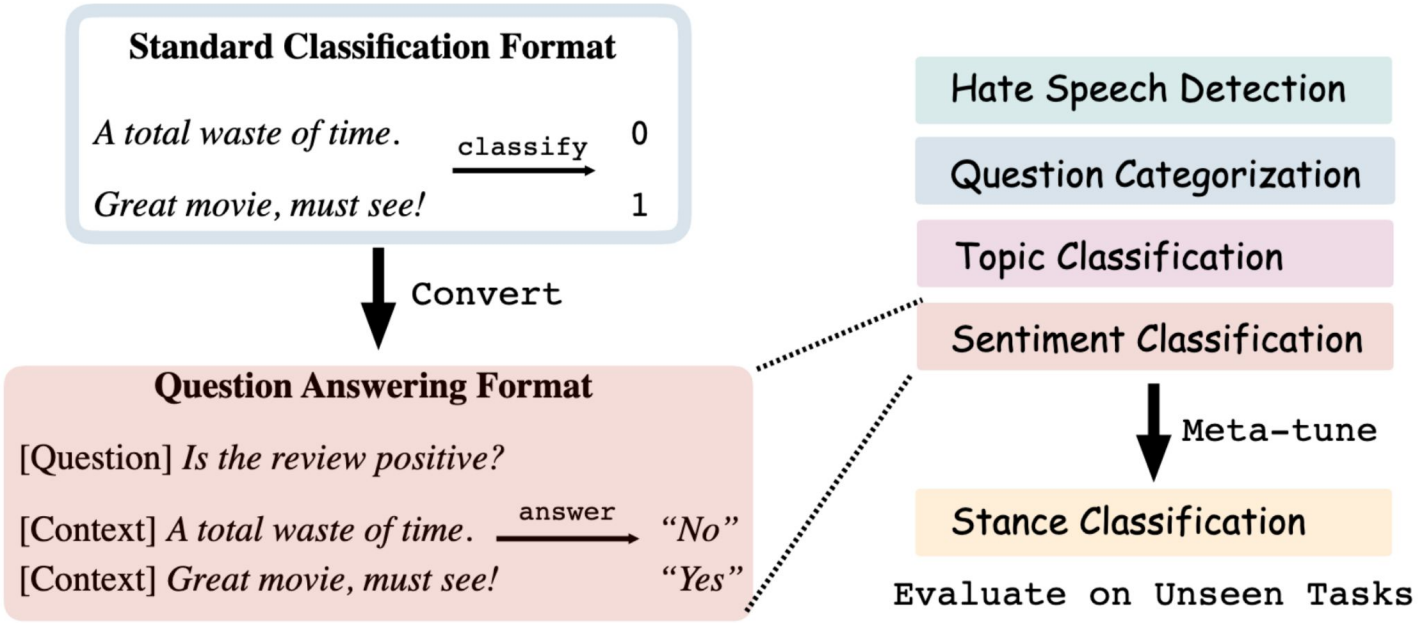
Luke Zettlemoyer
Facebook
lsz@fb.com

Sonal Gupta
Facebook
sonalgupta@fb.com



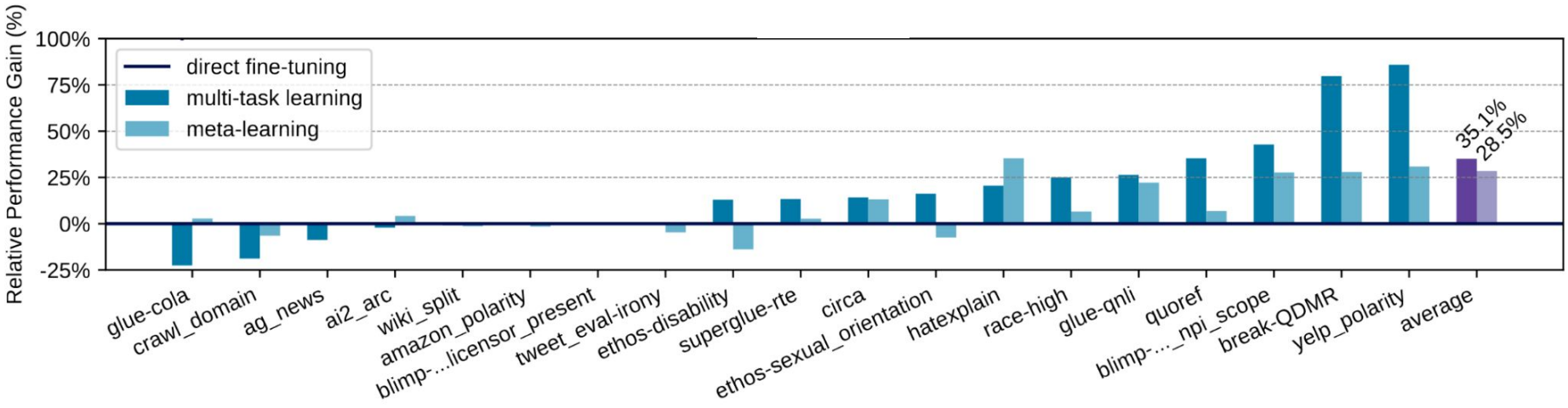
Meta-tuning Language Models to Answer Prompts Better

Ruiqi Zhong **Kristy Lee*** **Zheng Zhang*** **Dan Klein**
Computer Science Division, University of California, Berkeley
{ruiqi-zhong, kristylee, zhengzhang1216, klein}@berkeley.edu



CROSSFIT: A Few-shot Learning Challenge for Cross-task Generalization in NLP

Qinyuan Ye **Bill Yuchen Lin** **Xiang Ren**
University of Southern California
{qinyuany, yuchen.lin, xiangren}@usc.edu



Can we obtain a model that can readily perform tons of NLP tasks at human-level performance without further fine-tuning?

Maybe we should try training a **giant model** in a **massively multi-task setup** with an **extremely diverse** set of task formats.

https://bigscience.huggingface.co/