Session #6:
In-Context Learning

Thursday, Sept 8

CSCI 601.771: Self-supervised Statistical Models

Johns Hopkins University

Project Proposals

Fri Sept 30	Project proposal submission deadline	
#11 - Tue Oct 4	Memorization and Privacy	Slides Main Reading: Quantifying Memorization Across Neural Language Models Additional Reading(s): 1. Extracting Training Data from Large Language Models 2. Counterfactual Memorization in Neural Language Models 3. Data Contamination: From Memorization to Exploitation 4. Memorization Without Overfitting: Analyzing the Training Dynamics of Large Language Models
#12 - Thu Oct 6	Memorization and Privacy	Slides Main Reading: Deduplicating Training Data Mitigates Privacy Risks in Language Models Additional Reading(s): 1. Differentially Private Fine-tuning of Language Models 2. Can a Model Be Differentially Private and Fair? 3. Large Language Models Can Be Strong Differentially Private Learners 4. What Does it Mean for a Language Model to Preserve Privacy?
#13 - Tue Oct 11	External Speaker: Anjalie Field	
#14 - Thu Oct 13	Project Proposal Presentation	Slides

Project Proposals

Fri Sept 30	Project proposal submission deadline	
#11 - Tue Oct 4	Memorization and Privacy	Slides
#12 - Thu Oct 6	Memorization and Privacy	
#13 - Tue Oct 11	External Speaker: Anjalie Field	
#14 - Thu Oct 13	Project Proposal Presentation	Slides

- **Deadline:** Friday, Sept 30
- **Topic:** open-ended
- **Content:** a single-paragraph description of what you intend to do (experiments, datasets, methods, etc.)
- I will provide feedback on these ideas to help the teams with finding a concrete idea.
- Teamwork is optional but encouraged!

Lightening Proposal Presentations

Fri Sept 30	Project proposal submission deadline	
#11 - Tue Oct 4	Memorization and Privacy	Slides Main Reading: Quantifying Memorization Across Neural Language Models
#12 - Thu Oct 6	Memorization	1. Differentially Private Fine-tuning of Language Models 2. Can a Model Be Differentially Private and Fair? 3. Large Language Models Can Be Strong Differentially Private Learners 4. What Does it Mean for a Language Model to Preserve Privacy?
#13 - Tue Oct 11	External Speaker: Anjalie Field	
#14 - Thu Oct 13	Project Proposal Presentation	Slides

- **When?** Thursday, Oct 13 (the usual class time)
- **What:** each time will present their proposal in a few minutes.

Week's prompt

This would have been a much better paper if _____

Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity

Yao Lu[†] **Max Bartolo**[†] **Alastair Moore**[‡] **Sebastian Riedel**[†] **Pontus Stenetorp**[†]

[†]University College London [‡]Mishcon de Reya LLP

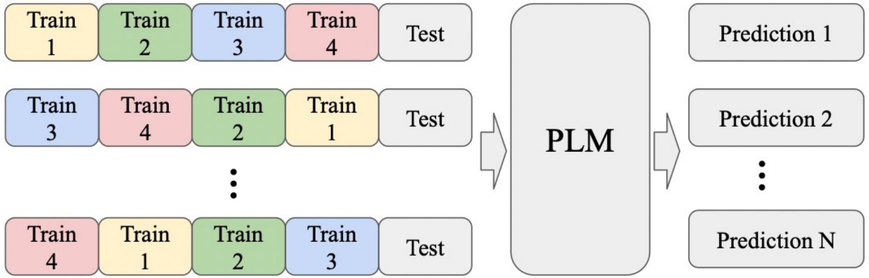
{yao.lu,m.bartolo,s.riedel,p.stenetorp}@cs.ucl.ac.uk

alastair.moore@mishcon.com

✍: Aayush Mishra and Lingfeng Shen

Prompt Order Sensitivity

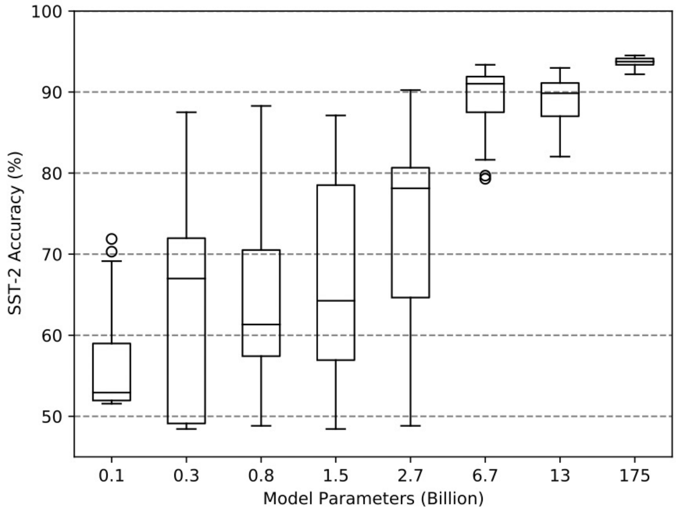
1. Take 4 samples, create all 24 permutations, test prediction performance.

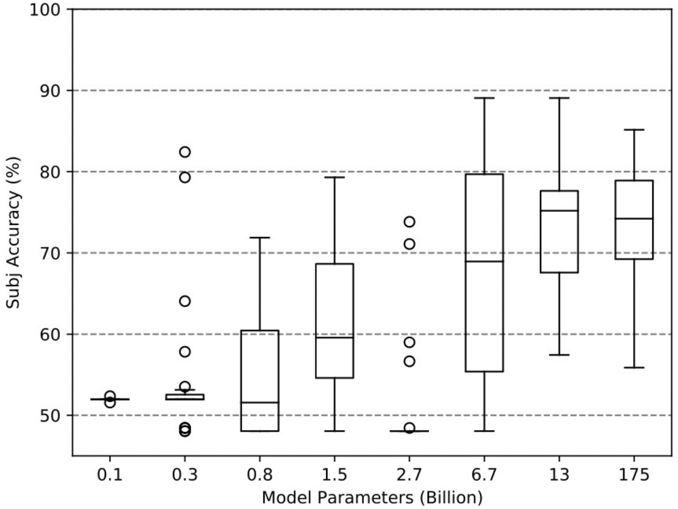


2. Test on variety of tasks (datasets) and models (4 GPT-2, 4 GPT-3 sizes)

✍: Aayush Mishra and Lingfeng Shen

Do Large Language Models really understand prompts well?





Order matters for self-supervised models

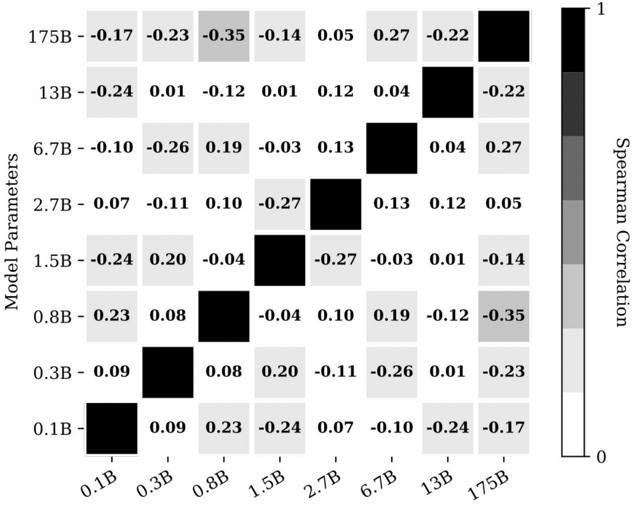
Model size matters, but not always

Order does not matter as much for Supervised models[1]

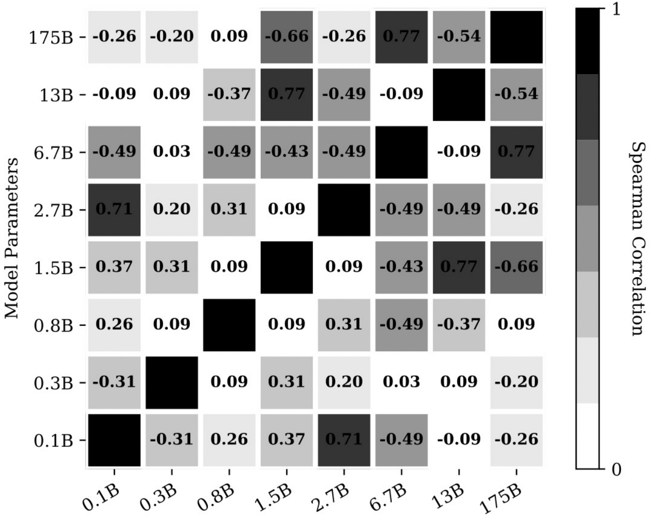
✍: Aayush Mishra and Lingfeng Shen

[1] Gao et. al. Making pre-trained language models better few-shot learners. arXiv preprint arXiv:2012.15723

Prompt Design Study



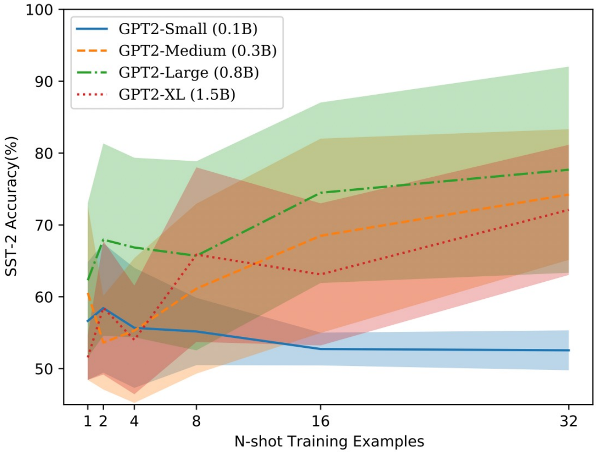
Performance Prompts are not transferable across models

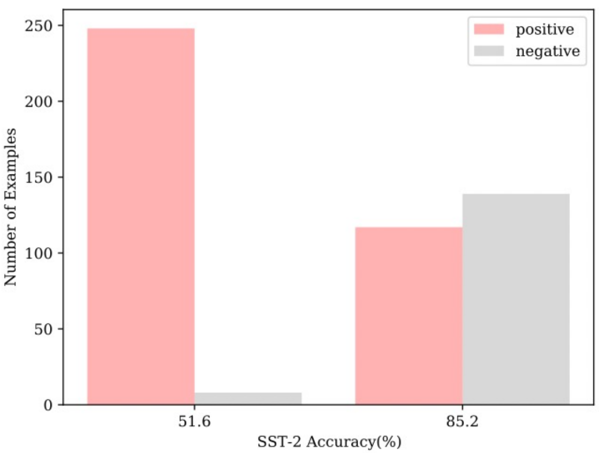


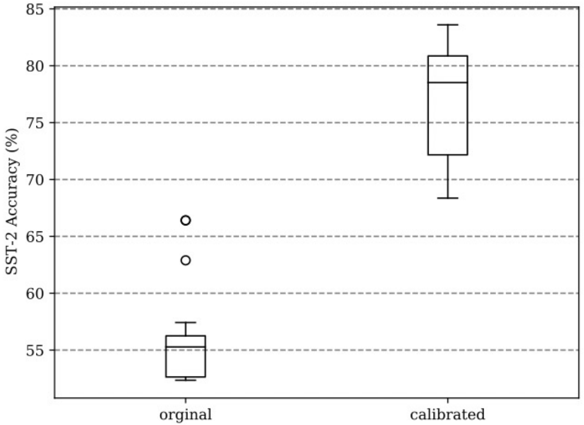
Label ordering does not matter

✍: Aayush Mishra and Lingfeng Shen

Prompt Design Study







Increasing N does not reduce performance variance much

Failing prompts suffer from unbalanced label distribution

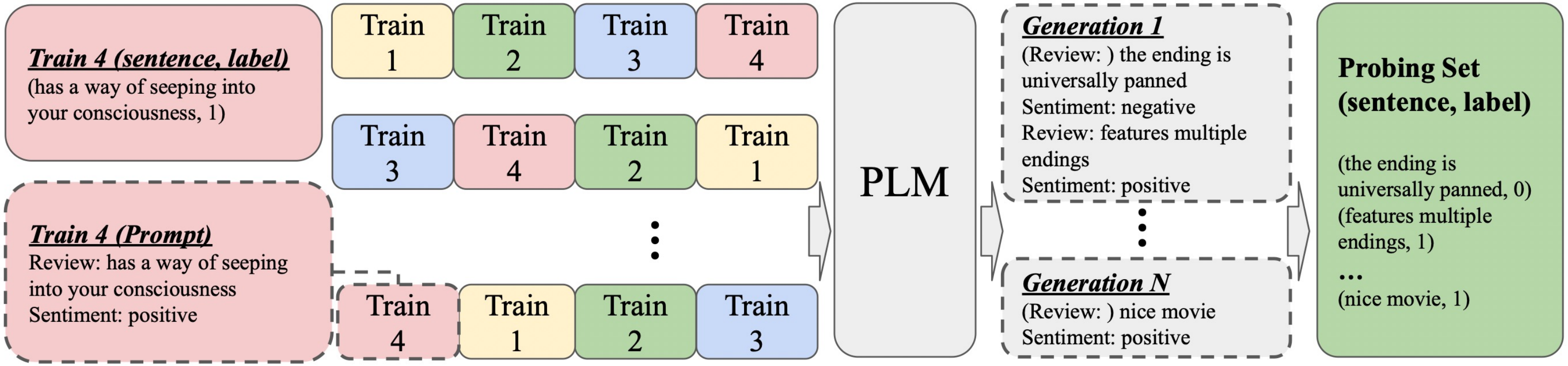
Calibration[1] improves performance but variance stays high

[1] Zhao et. al. Calibrate before use: Improving few-shot performance of language models. arXiv preprint arXiv:2102.09690.

✍: Aayush Mishra and Lingfeng Shen

Prompt Engineering

How to automatically generate a 'probing set' to find performant prompt orderings?



✍: Aayush Mishra and Lingfeng Shen

Probing Metrics

$$p_m^v = \frac{\sum_i \mathbb{1}_{\{\hat{y}_{i,m}=v\}}}{|D|}$$

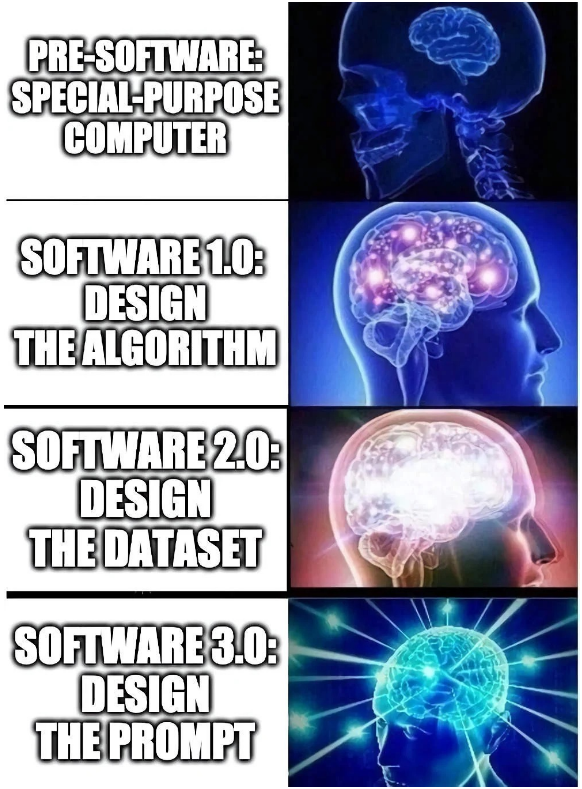
$$p_{i,m}^v = P_{(x_i',y_i')\sim D}(v|c_m \oplus \mathcal{T}(x_i'); \theta), v \in V$$

$$\text{GlobalE}_m = \sum_{v \in V} -p_m^v \log p_m^v$$

For prompts that avoid extremely unbalanced predictions.

$$\text{LocalE}_m = \frac{\sum_i \sum_{v \in V} -p_{i,m}^v \log p_{i,m}^v}{|D|}$$

To penalize overconfident predictions.



https://www.gwern.net/GPT-3

✍: Aayush Mishra and Lingfeng Shen

Experiments

- Models: GPT-2 (with 0.1B, 0.3B, 0.8B, and 1.5B parameters), GPT-3 (with 2.7B, and 175B parameters)
- Benchmarks: Classification dataset : SST-2, SST-5, DBPedia ......
- Baselines:
  - Majority: predict the majority label in the dataset (lower-bound)
  - Oracle: select the top four orderings based on performance on the dev set (upper-bound)

https://www.gwern.net/GPT-3

✍: Aayush Mishra and Lingfeng Shen

	SST-2	SST-5	DBPedia	MR	CR	MPQA	Subj	TREC	AGNews	RTE	CB
Majority	50.9	23.1	9.4	50.0	50.0	50.0	50.0	18.8	25.0	52.7	51.8
Finetuning (Full)	95.0	58.7	99.3	90.8	89.4	87.8	97.0	97.4	94.7	80.9	90.5
GPT-2 0.1B	$58.9_{7.8}$	$29.0_{4.9}$	$44.9_{9.7}$	$58.6_{7.6}$	$58.4_{6.4}$	$68.9_{7.1}$	$52.1_{0.7}$	$\mathbf{49.2}_{4.7}$	$50.8_{11.9}$	$49.7_{2.7}$	$50.1_{1.0}$
LocalE	$\mathbf{65.2}_{3.9}$	$34.4_{3.4}$	$53.3_{4.9}$	$66.0_{6.3}$	$\mathbf{65.0}_{3.4}$	$72.5_{6.0}$	$52.9_{1.3}$	$48.0_{3.9}$	$61.0_{5.9}$	$\mathbf{53.0}_{3.3}$	$49.9_{1.6}$
GlobalE	$63.8_{5.8}$	$\mathbf{35.8}_{2.0}$	$\mathbf{56.1}_{4.3}$	$\mathbf{66.4}_{5.8}$	$64.8_{2.7}$	$\mathbf{73.5}_{4.5}$	$\mathbf{53.0}_{1.3}$	$46.1_{3.7}$	$\mathbf{62.1}_{5.7}$	$\mathbf{53.0}_{3.0}$	$\mathbf{50.3}_{1.6}$
Oracle	$73.5_{1.7}$	$38.2_{4.0}$	$60.5_{4.2}$	$74.3_{4.9}$	$70.8_{4.4}$	$81.3_{2.5}$	$55.2_{1.7}$	$58.1_{4.3}$	$70.3_{2.8}$	$56.8_{2.0}$	$52.1_{1.3}$
GPT-2 0.3B	$61.0_{13.2}$	$25.9_{5.9}$	$51.7_{7.0}$	$54.2_{7.8}$	$56.7_{9.4}$	$54.5_{8.8}$	$54.4_{7.9}$	$52.6_{4.9}$	$47.7_{10.6}$	$48.8_{2.6}$	$50.2_{5.3}$
LocalE	$75.3_{4.6}$	$31.0_{3.4}$	$47.1_{3.7}$	$65.2_{6.6}$	$\mathbf{70.9}_{6.3}$	$67.6_{7.2}$	$\mathbf{66.7}_{9.3}$	$53.0_{3.9}$	$51.2_{7.3}$	$\mathbf{51.8}_{1.0}$	$47.1_{4.2}$
GlobalE	$\mathbf{78.7}_{5.2}$	$\mathbf{31.7}_{5.2}$	$\mathbf{58.3}_{5.4}$	$\mathbf{67.0}_{5.9}$	$70.7_{6.7}$	$\mathbf{68.3}_{6.9}$	$65.8_{10.1}$	$\mathbf{53.3}_{4.6}$	$\mathbf{59.6}_{7.2}$	$51.1_{1.9}$	$\mathbf{50.3}_{3.7}$
Oracle	$85.5_{4.3}$	$40.5_{6.3}$	$65.2_{7.6}$	$74.7_{6.1}$	$80.4_{5.4}$	$77.3_{2.3}$	$79.4_{2.4}$	$63.3_{2.9}$	$68.4_{8.0}$	$53.9_{1.3}$	$62.5_{7.4}$
GPT-2 0.8B	$74.5_{10.3}$	$34.7_{8.2}$	$55.0_{12.5}$	$64.6_{13.1}$	$70.9_{12.7}$	$65.5_{8.7}$	$56.4_{9.1}$	$56.5_{2.7}$	$62.2_{11.6}$	$53.2_{2.0}$	$38.8_{8.5}$
LocalE	$81.1_{5.5}$	$40.3_{4.7}$	$56.7_{7.5}$	$82.6_{4.2}$	$85.4_{3.8}$	$73.6_{4.8}$	$\mathbf{70.4}_{4.2}$	$56.2_{1.7}$	$62.7_{8.1}$	$53.3_{1.6}$	$38.4_{5.2}$
GlobalE	$\mathbf{84.8}_{4.1}$	$\mathbf{46.9}_{1.1}$	$\mathbf{67.7}_{3.6}$	$\mathbf{84.3}_{2.9}$	$\mathbf{86.7}_{2.5}$	$\mathbf{75.8}_{3.1}$	$68.6_{6.5}$	$\mathbf{57.2}_{2.3}$	$\mathbf{70.7}_{3.6}$	$\mathbf{53.5}_{1.5}$	$\mathbf{41.2}_{4.5}$
Oracle	$88.9_{1.8}$	$48.4_{0.7}$	$72.3_{3.3}$	$87.5_{1.1}$	$89.9_{0.9}$	$80.3_{4.9}$	$76.6_{4.1}$	$62.1_{1.5}$	$78.1_{1.3}$	$57.3_{1.0}$	$53.2_{5.3}$
GPT-2 1.5B	$66.8_{10.8}$	$41.7_{6.7}$	$82.6_{2.5}$	$59.1_{11.9}$	$56.9_{9.0}$	$73.9_{8.6}$	$59.7_{10.4}$	$53.1_{3.3}$	$77.6_{7.3}$	$55.0_{1.4}$	$53.8_{4.7}$
LocalE	$76.7_{8.2}$	$\mathbf{45.1}_{3.1}$	$83.8_{1.7}$	$78.1_{5.6}$	$71.8_{8.0}$	$78.5_{3.6}$	$69.7_{5.8}$	$53.6_{3.1}$	$79.3_{3.7}$	$\mathbf{56.8}_{1.1}$	$52.6_{3.9}$
GlobalE	$\mathbf{81.8}_{3.9}$	$43.5_{4.5}$	$\mathbf{83.9}_{1.8}$	$\mathbf{77.9}_{5.7}$	$\mathbf{73.4}_{6.0}$	$\mathbf{81.4}_{2.1}$	$\mathbf{70.9}_{6.0}$	$\mathbf{55.5}_{3.0}$	$\mathbf{83.9}_{1.2}$	$56.3_{1.2}$	$\mathbf{55.1}_{4.6}$
Oracle	$86.1_{1.5}$	$50.9_{1.0}$	$87.3_{1.5}$	$84.0_{2.7}$	$80.3_{3.3}$	$85.1_{1.4}$	$79.9_{5.7}$	$59.0_{2.3}$	$86.1_{0.7}$	$58.2_{0.6}$	$63.9_{4.3}$
GPT-3 2.7B	$78.0_{10.7}$	$35.3_{6.9}$	$81.1_{1.8}$	$68.0_{12.9}$	$76.8_{11.7}$	$66.5_{10.3}$	$49.1_{2.9}$	$55.3_{4.4}$	$72.9_{4.8}$	$48.6_{1.9}$	$50.4_{0.7}$
LocalE	$\mathbf{81.0}_{6.0}$	$42.3_{4.7}$	$80.3_{1.7}$	$75.6_{4.1}$	$79.0_{5.5}$	$72.5_{5.8}$	$54.2_{4.2}$	$54.0_{2.6}$	$72.3_{4.6}$	$50.4_{1.9}$	$50.5_{0.8}$
GlobalE	$80.2_{4.2}$	$\mathbf{43.2}_{4.3}$	$\mathbf{81.2}_{0.9}$	$\mathbf{76.1}_{3.8}$	$\mathbf{80.3}_{3.4}$	$\mathbf{73.0}_{4.3}$	$\mathbf{54.3}_{4.0}$	$\mathbf{56.7}_{2.0}$	$\mathbf{78.1}_{1.9}$	$\mathbf{51.3}_{1.8}$	$\mathbf{51.2}_{0.8}$
Oracle	$89.8_{0.7}$	$48.0_{1.1}$	$85.4_{1.6}$	$87.4_{0.9}$	$90.1_{0.7}$	$80.9_{1.4}$	$60.3_{10.3}$	$62.8_{4.2}$	$81.3_{2.9}$	$53.4_{3.1}$	$52.5_{1.4}$
GPT-3 175B	$\mathbf{93.9}_{0.6}$	$54.4_{2.5}$	$95.4_{0.9}$	$\mathbf{94.6}_{0.7}$	$91.0_{1.0}$	$83.2_{1.5}$	$71.2_{7.3}$	$72.1_{2.7}$	$85.1_{1.7}$	$70.8_{2.8}$	$75.1_{5.1}$
LocalE	$93.8_{0.5}$	$\mathbf{56.0}_{1.7}$	$95.5_{0.9}$	$94.5_{0.7}$	$91.3_{0.5}$	$\mathbf{83.3}_{1.7}$	$75.0_{4.6}$	$71.8_{3.2}$	$\mathbf{85.9}_{0.7}$	$\mathbf{71.9}_{1.4}$	$74.6_{4.2}$
GlobalE	$\mathbf{93.9}_{0.6}$	$53.2_{2.1}$	$\mathbf{95.7}_{0.7}$	$\mathbf{94.6}_{0.2}$	$\mathbf{91.7}_{0.4}$	$82.0_{0.8}$	$\mathbf{76.3}_{3.5}$	$\mathbf{73.6}_{2.5}$	$85.7_{1.0}$	$71.8_{1.9}$	$\mathbf{79.9}_{3.3}$
Oracle	$94.7_{0.2}$	58.2	$96.7_{0.2}$	$95.5_{0.2}$	$92.6_{0.4}$	$85.5_{0.8}$	$81.1_{4.9}$	$77.0_{1.2}$	$87.7_{0.6}$	$74.7_{0.4}$	$83.0_{0.9}$

✍: Aayush Mishra and Lingfeng Shen

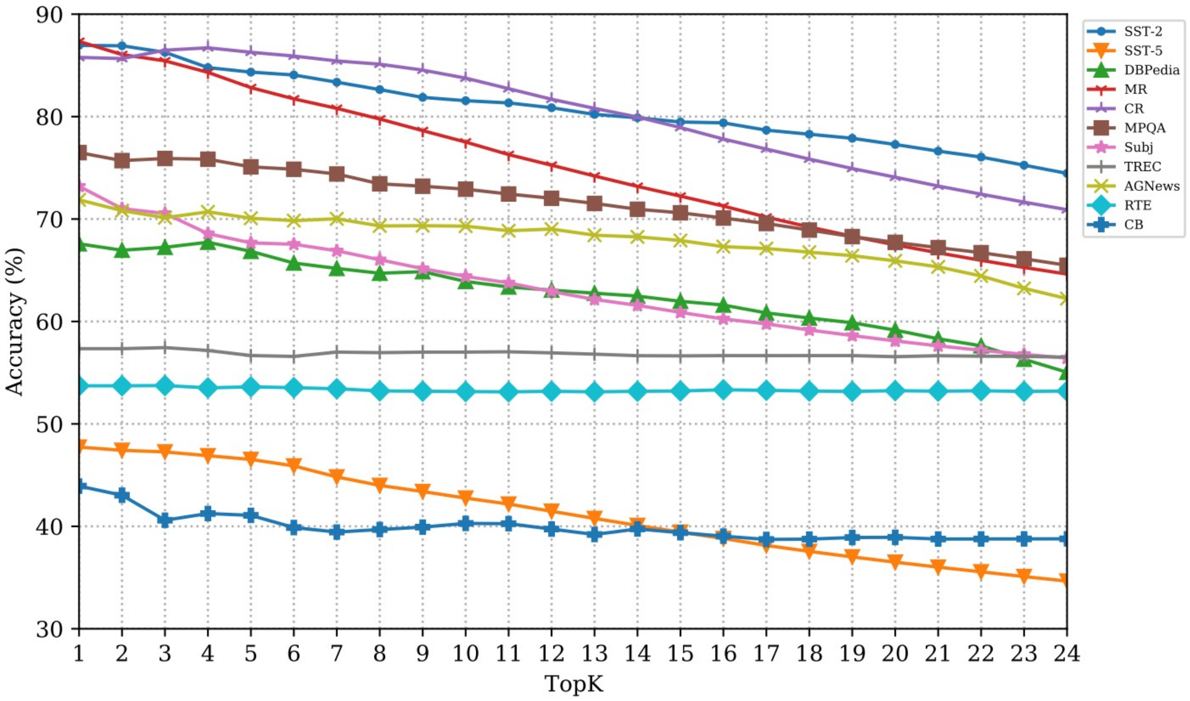
Results

- Entropy-based probing is effective for performant prompt selection regardless of model size
  - GlobalE achieves, on average, a 13% relative improvement across the eleven different sentence classification tasks in comparison to without probing.
  - LocalE provides results slightly inferior to GlobalE, with an average 9.6% relative improvement over the baseline model.

✍: Aayush Mishra and Lingfeng Shen

Results

- Ranking using Entropy-based probing is robust



✍: Aayush Mishra and Lingfeng Shen

Results

- Entropy-based probing is effective across templates

	Template 1	Template 2	Template 3	Template 4
GPT-2 0.1B	$58.9_{7.8}$	$57.5_{6.8}$	$58.1_{7.4}$	$56.6_{6.6}$
LocalE	$\mathbf{65.2}_{3.9}$	$\mathbf{60.7}_{4.6}$	$\mathbf{65.4}_{4.8}$	$61.0_{4.7}$
GlobalE	$63.8_{5.8}$	$59.0_{2.9}$	$64.3_{4.8}$	$\mathbf{63.5}_{4.8}$
GPT-2 0.3B	$61.0_{13.2}$	$63.9_{11.3}$	$68.3_{11.8}$	$59.2_{6.4}$
LocalE	$75.3_{4.6}$	$70.0_{7.2}$	$80.2_{4.2}$	$62.2_{3.4}$
GlobalE	$\mathbf{78.7}_{5.2}$	$\mathbf{73.3}_{4.5}$	$\mathbf{81.3}_{4.1}$	$\mathbf{62.8}_{4.3}$
GPT-2 0.8B	$74.5_{10.3}$	$66.6_{10.6}$	$70.3_{10.5}$	$63.7_{8.9}$
LocalE	$81.1_{5.5}$	$80.0_{5.6}$	$73.7_{6.2}$	$\mathbf{71.3}_{4.5}$
GlobalE	$\mathbf{84.8}_{4.1}$	$\mathbf{80.9}_{3.6}$	$\mathbf{79.8}_{3.9}$	$70.7_{5.3}$
GPT-2 1.5B	$66.8_{10.8}$	$80.4_{7.6}$	$54.5_{7.9}$	$69.1_{10.5}$
LocalE	$76.7_{8.2}$	$83.1_{3.6}$	$66.9_{7.5}$	$72.7_{5.5}$
GlobalE	$\mathbf{81.8}_{3.9}$	$\mathbf{83.4}_{3.2}$	$\mathbf{67.2}_{6.1}$	$\mathbf{74.2}_{5.3}$

Overview

- Pointing out a counter-intuitive phenomenon
  - few-shot prompts suffer from order sensitivity

✍: Aayush Mishra and Lingfeng Shen

Overview

- Pointing out a counter-intuitive phenomenon
  - few-shot prompts suffer from order sensitivity

- Conduct comprehensive empirical analyses from different aspects
  - tasks, model sizes, prompt templates, samples, and number of training samples.

✍: Aayush Mishra and Lingfeng Shen

Overview

- Pointing out a counter-intuitive phenomenon
  - few-shot prompts suffer from order sensitivity
- Conduct comprehensive empirical analyses from different aspects
  - tasks, model sizes, prompt templates, samples, and number of training samples.
- Propose an effective method to tackle the problem
  - a probing method that construct an artificial development set.

Future directions

- Linguistic perspective
  - Are there any linguistic commonalities in these good orders?
  - How do these good orders arise?
  - Does it correlate with some linguistic distributions in the pre-trained corpus?

Future directions

- Linguistic perspective
  - Are there any linguistic commonalities in these good orders?
  - How do these good orders arise?
  - Does it correlate with some linguistic distributions in the pre-trained corpus?

- Mathematical perspective
  - Does the uncertainty issue really come from biased/over-confident predictions?
  - Where does the uncertainty come from? (Error of estimated distribution towards ground-truth)
  - PAC-bayes or something?

✍: Aayush Mishra and Lingfeng Shen

Cons - Vicky

1. Unexplored theoretical grounding | Lack of transferability
    a. Prompt ordering affects performance greatly but is not transferable
        i. Why does ordering matter? Why is it not transferable? Is this similar to brute-force
    b. Probing metrics: Each motivation explained, but does not explain why only these two / how these two compare, and reason about their differing performances
2. Ablations not fully covered
    a. Argument on template invariance: Singled out sentiment analysis that inherently has limited template formats
    b. Lack of coverage on the 11 tasks evaluated: Pointed out sentence-pair tasks, but what about others? Complete breakdown beneficial
    c. Argument on probing to be better than train-devel split: Is it really better than original data, or is the split unfair? (Train set cut to half, expected drop)
3. General comments
    a. Figure captions can be improved
        i. Fig 1: Lack of description on variation within single sample run
        ii. Fig 3: Insufficient description on variance shade
        iii. Fig 4/5: Insufficient description on correlation value (small = worse)
    b. Introduce some context earlier for better grounding
        i. Reason for choosing 4-shot (limited by window size)
        ii. Each sample run is averaged across 5 subsets, each with 24 permutations

Pros - Aowei

- Figures clearly showed that:
  - models different order of the sample would obtain different performance;
  - For the same dataset, different models require different prompt ordering to reach a high performance.
- Novel strategy for obtaining different prompt set
  - Achievement : Automatically selecting prompt set. Without using development set
  - Generating from train set (Limited develop set in few shot).
  - Global Entropy. Avoid unbalancing,non-performant prompts. Local Entropy. Find the prompt with high ability in differentiating classes.
  - Considered the situation that not rely on generation samples. Also did experiment with gathering prompt set from the development dataset. Also obtain better performance comparing to Baseline.

Summary: Few-Shot Prompt Order Sensitivity

When primed with a handful of prompts in few-shot learning, changing the order of prompts provided can cause performance to improve from random (50%) to state-of-the-art (90%)[1]. This is present across tasks, model sizes, number of prompts, and prompt templates. To optimize this order sensitivity, the paper presents a novel probing method that generates an artificial development set from the language model via sampling of existing data. Entropy statistics is run on this development set to identify the best order permutations, leading to an average of 13% improvement across eleven text classification tasks.

1. 6.7B-parameter GPT-3 for Subject Classification Task

🔎 **Vicky Zeng and Aowei Ding**

Review: Summary

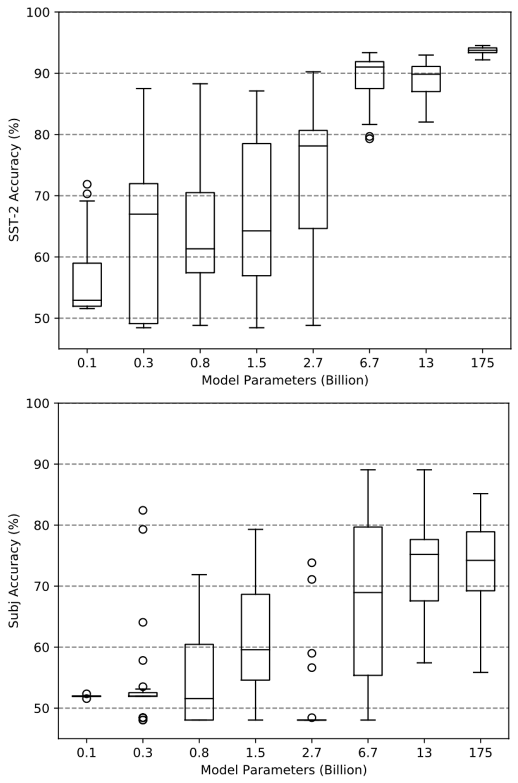
When primed with a handful of samples in few-shot learning, changing the order of samples provided can cause performance to improve from random (50%) to state-of-the-art (90%)[1]. This is present across tasks, model sizes, number of samples, and prompt templates. To optimize this order sensitivity, the paper presents a novel probing method that generates an artificial development set from the language model via sampling of existing data. Entropy statistics is run to identify the best order permutations, leading to an average of 13% improvement across eleven text classification tasks.

1. 6.7B-parameter GPT-3 for Subject Classification Task

- 🔎 **Vicky Zeng and Aowei Ding**

Strength

- Figures clearly showed that:
    - models different order of the sample would obtain different performance;



- 🔎**Vicky Zeng and Aowei Ding**

Strength

- For the same dataset, different models require different prompt ordering to reach a high performance.

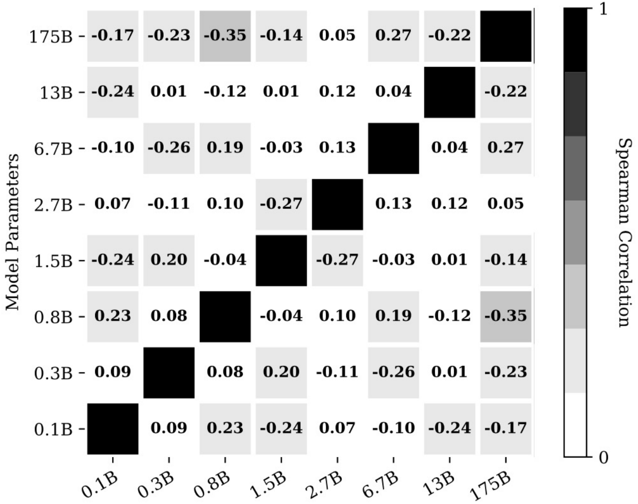


Figure 4: Training sample permutation performance correlation across different models.

pairwise Spearman's rank correlation coefficient between the scores.

- 🔎Vicky Zeng and Aowei Ding

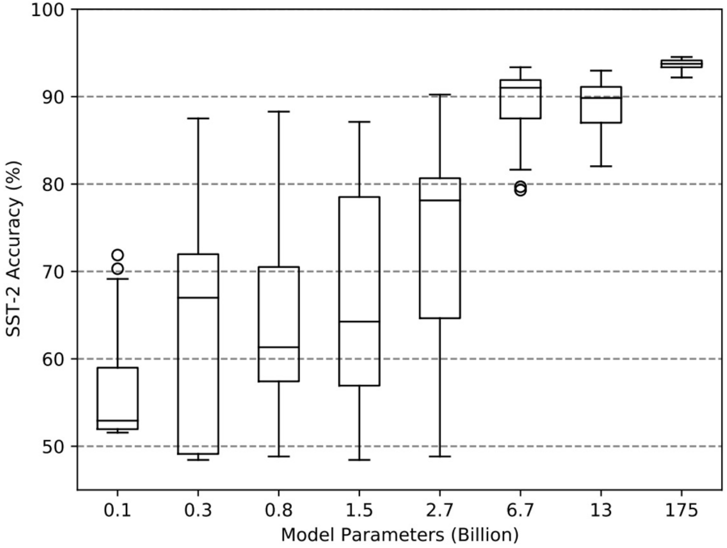
Strength

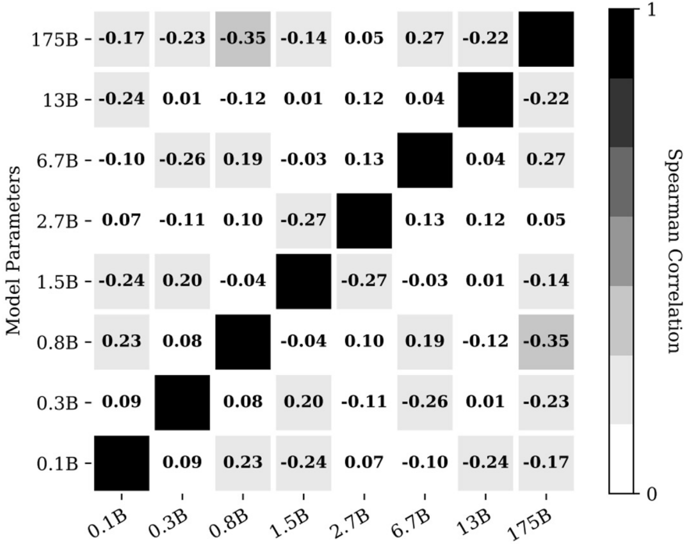
- Novel strategy for obtaining different prompt set
  - Generating from train set (Limited development set in few shot).
  - Global Entropy. Avoid unbalancing, non-performant prompts. Local Entropy. Find the prompt with high ability in differentiating classes.
  - Considered the situation that not relying on generation samples(sufficient development dataset). Also did experiment with gathering prompt set from the development dataset. Also obtain better performance comparing to Baseline.

	GPT-2 0.1B	GPT-2 0.3B	GPT-2 0.8B	GPT-2 1.5B
Baseline	$58.9_{7.8}$	$61.0_{13.2}$	$74.5_{10.3}$	$66.8_{10.8}$
LocalE	$\mathbf{65.2}_{3.9}$	$75.3_{4.6}$	$81.1_{5.5}$	$76.7_{8.2}$
GlobalE	$63.8_{5.8}$	$\mathbf{78.7}_{5.2}$	$\mathbf{84.8}_{4.1}$	$\mathbf{81.8}_{3.9}$
Split Training Set	$62.8_{5.3}$	$64.2_{6.1}$	$75.1_{6.8}$	$71.4_{7.8}$

- 🔎**Vicky Zeng and Aowei Ding**

Why? - Theoretical Grounding, Transferability



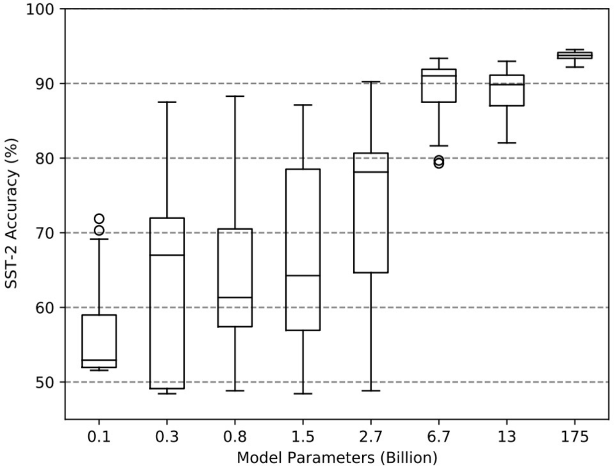


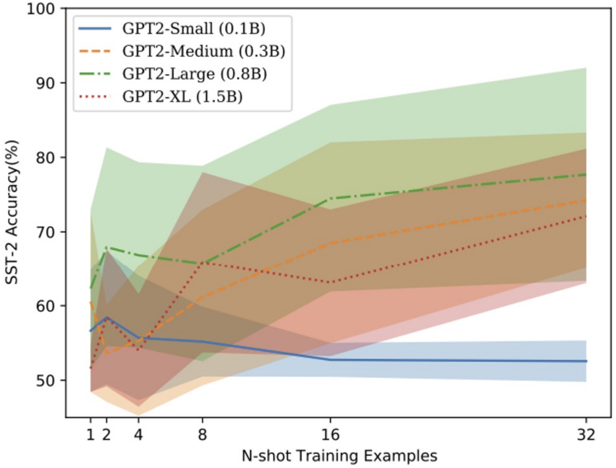
Why: Why is there order sensitivity / performance difference so great?

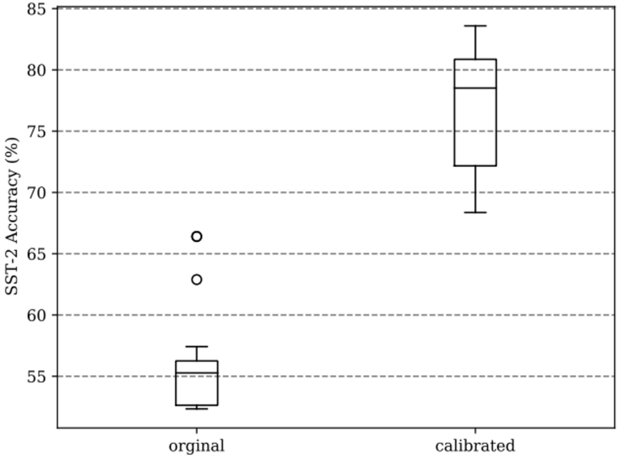
Why: Why are optimal prompt permutations not transferable across models?

🔎 **Vicky Zeng and Aowei Ding**

Better ifs - Ablations on Templates, Tasks, Train-Devel Split







GOOD

🔎 **Vicky Zeng and Aowei Ding**

Better ifs - Ablations on Templates, Tasks, Train-Devel Split

ID	Template	Label Mapping
1	Review: {Sentence} Sentiment: {Label}	positive/negative
2	Input: {Sentence} Prediction: {Label}	positive/negative
3	Review: {Sentence} Sentiment: {Label}	good/bad
4	{Sentence} It was {Label}	good/bad

	Template 1	Template 2	Template 3	Template 4
GPT-2 0.1B	$58.9_{7.8}$	$57.5_{6.8}$	$58.1_{7.4}$	$56.6_{6.6}$
LocalE	$\mathbf{65.2}_{3.9}$	$\mathbf{60.7}_{4.6}$	$\mathbf{65.4}_{4.8}$	$61.0_{4.7}$
GlobalE	$63.8_{5.8}$	$59.0_{2.9}$	$64.3_{4.8}$	$\mathbf{63.5}_{4.8}$
GPT-2 0.3B	$61.0_{13.2}$	$63.9_{11.3}$	$68.3_{11.8}$	$59.2_{6.4}$
LocalE	$75.3_{4.6}$	$70.0_{7.2}$	$80.2_{4.2}$	$62.2_{3.4}$
GlobalE	$\mathbf{78.7}_{5.2}$	$\mathbf{73.3}_{4.5}$	$\mathbf{81.3}_{4.1}$	$\mathbf{62.8}_{4.3}$
GPT-2 0.8B	$74.5_{10.3}$	$66.6_{10.6}$	$70.3_{10.5}$	$63.7_{8.9}$
LocalE	$81.1_{5.5}$	$80.0_{5.6}$	$73.7_{6.2}$	$\mathbf{71.3}_{4.5}$
GlobalE	$\mathbf{84.8}_{4.1}$	$\mathbf{80.9}_{3.6}$	$\mathbf{79.8}_{3.9}$	$70.7_{5.3}$
GPT-2 1.5B	$66.8_{10.8}$	$80.4_{7.6}$	$54.5_{7.9}$	$69.1_{10.5}$
LocalE	$76.7_{8.2}$	$83.1_{3.6}$	$66.9_{7.5}$	$72.7_{5.5}$
GlobalE	$\mathbf{81.8}_{3.9}$	$\mathbf{83.4}_{3.2}$	$\mathbf{67.2}_{6.1}$	$\mathbf{74.2}_{5.3}$

Better if: Greater template variation

🔎 **Vicky Zeng and Aowei Ding**

Better ifs - Ablations on Templates, Tasks, Train-Devel Split

Dataset	Prompt	Label Mapping
SST-2	Review: contains no wit , only labored gags Sentiment: negative	positive/negative
SST-5	Review: apparently reassembled from the cutting-room floor of any given daytime soap . Sentiment: terrible	terrible/bad/okay/good/great
MR	Review: lame sweet home leaves no southern stereotype unturned . Sentiment: negative	negative/positive
CR	Review: bluetooth does not work on this phone . Sentiment: negative	negative/positive
MPQA	Review: dangerous situation Sentiment: negative	negative/positive
Subj	Input: too slow , too boring , and occasionally annoying . Type: subjective	subjective/objective
TREC	Question: When did the neanderthal man live ? Type: number	description/entity/expression/ human/location/number
AGNews	input: Wall St. Bears Claw Back Into the Black (Reuters). type: business	world/sports/business/technology
DBPedia	input: CMC Aviation is a charter airline based in Nairobi Kenya. type: company	company/school/artist/athlete/politics/ transportation/building/nature/village/ animal/plant/album/film/book
CB	premise: It was a complex language. Not written down but handed down. One might say it was peeled down. hypothesis: the language was peeled down prediction: true	true/false/neither
RTE	premise: No Weapons of Mass Destruction Found in Iraq Yet. hypothesis: Weapons of Mass Destruction Found in Iraq. prediction: False	True/False

Better if:

Greater variety of tasks

Quantitative breakdown of existing tasks

Vicky Zeng and Aowei Ding

Better ifs - Ablations on Templates, Tasks, Train-Devel Split

	GPT-2 0.1B	GPT-2 0.3B	GPT-2 0.8B	GPT-2 1.5B
Baseline	$58.9_{7.8}$	$61.0_{13.2}$	$74.5_{10.3}$	$66.8_{10.8}$
LocalE	$\mathbf{65.2}_{3.9}$	$75.3_{4.6}$	$81.1_{5.5}$	$76.7_{8.2}$
GlobalE	$63.8_{5.8}$	$\mathbf{78.7}_{5.2}$	$\mathbf{84.8}_{4.1}$	$\mathbf{81.8}_{3.9}$
Split Training Set	$62.8_{5.3}$	$64.2_{6.1}$	$75.1_{6.8}$	$71.4_{7.8}$

Better if: More generous train-devel split

🔎 **Vicky Zeng and Aowei Ding**

General Improvements - Figure Captions, Reproducibility

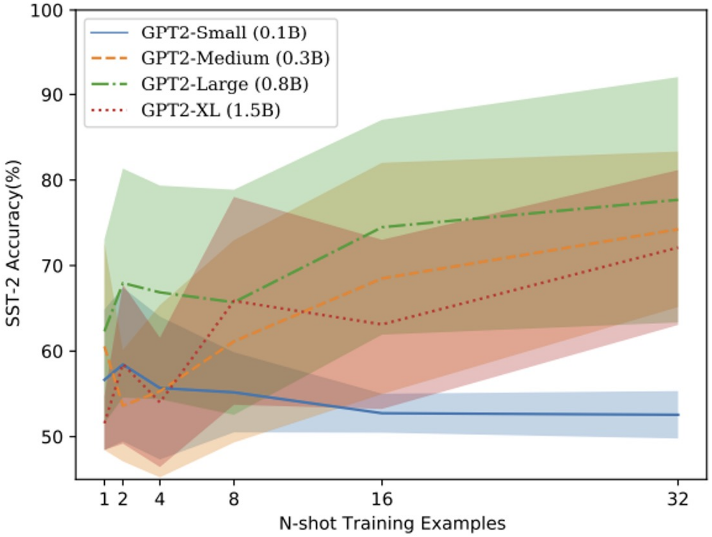


Figure 3: Order sensitivity using different numbers of training samples.

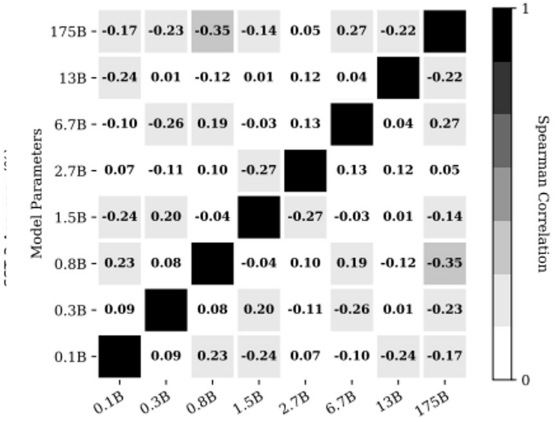


Figure 4: Training sample permutation performance correlation across different models.

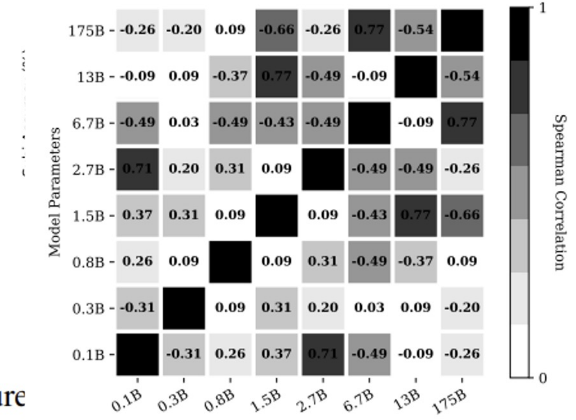
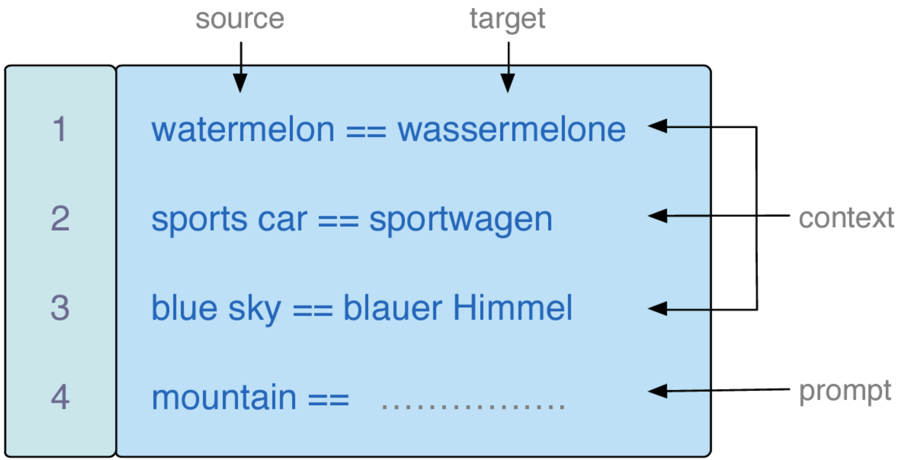


Figure sam-
ple or odels
(GPT- ts.

Figure 5: Training label pattern permutation performance correlation across different models.

🔎 **Vicky Zeng and Aowei Ding**

Archaeologist-Previous Work



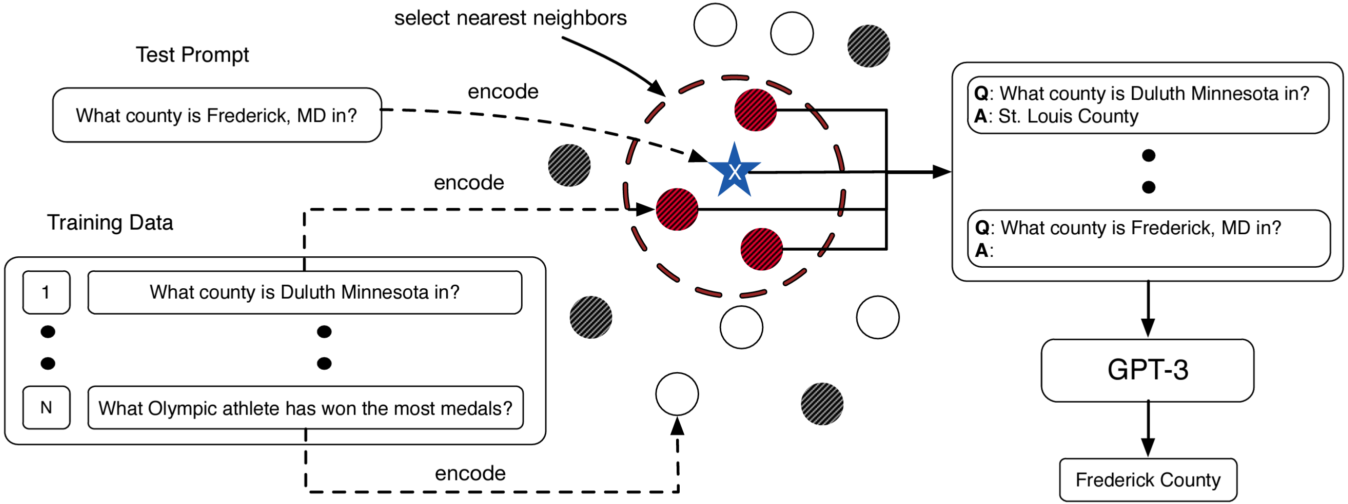
Question: Where do the examples in context come from?

- Training Data?
- Which examples to pick?
- What's the influence of example selection?

Archaeologist-Previous Work

Paper: What Makes Good In-Context Examples for GPT-3? Liu at el. (2021)

Overview: Non-parametric selection approach to retrieve in-context examples according to their semantic similarity(Euclidean Distance) to the test example.

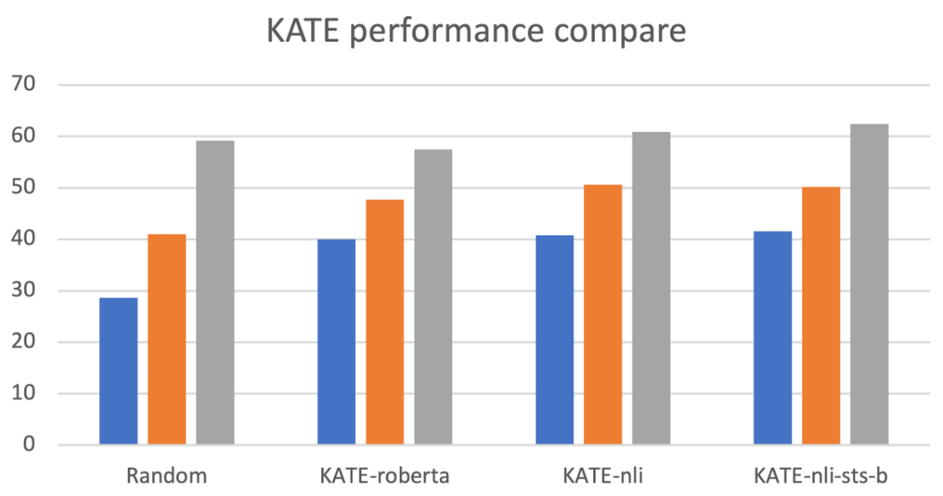


Archaeologist-Previous Work

Impact of In-Context Examples:

Method	Closest	Farthest
Accuracy	**46.0**	31.0

Experiment Results:



Calibrate Before Use:
Improving Few-Shot Performance of Language Models

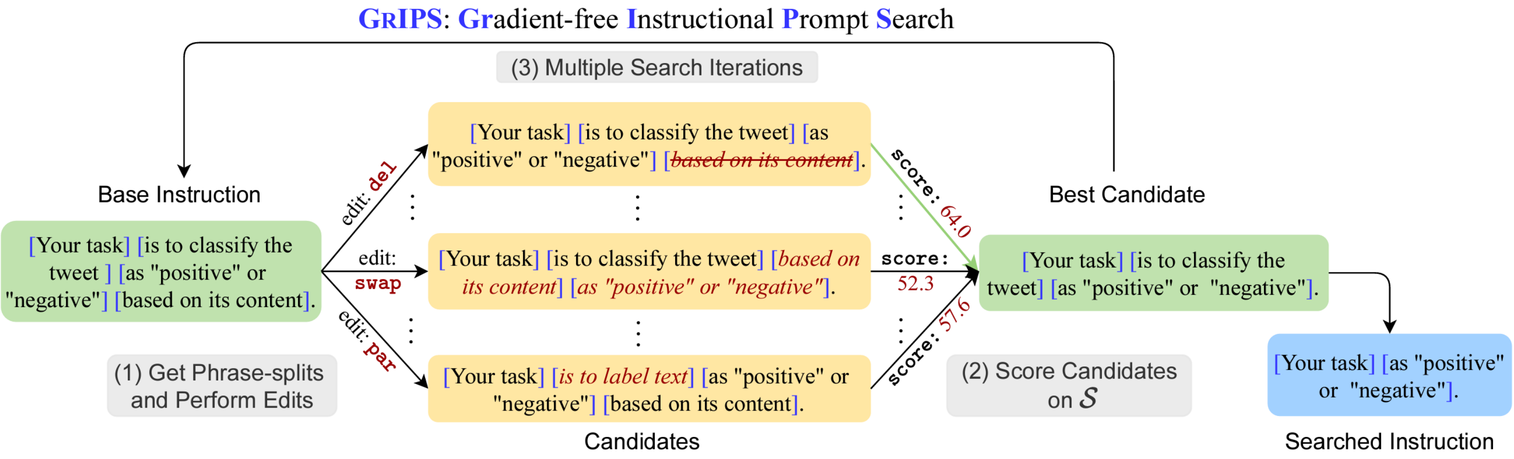
Tony Z. Zhao [*1] **Eric Wallace** [*1] **Shi Feng** [2] **Dan Klein** [1] **Sameer Singh** [3]

Archaeologist-Subsequent Works

Question:

- How about other format of prompts? Instructions, Examples, Discrete Templates?
- Is the current Instruction the best?
- Is Model's instruction aligned with Human Cognition?

Subsequent Works: GRIPS: Gradient-free, Edit-based Instruction Search for Prompting Large Language Models.



: Boyuan Zheng

Archaeologist-Subsequent Works

Results:

- GRIPS works for GPT-2 XL, InstructGPT, in both Instruction-Only and Instruction+Example prompts
- GRIPS > Manual Rewritting and Examples-Only Search
- Semantics: Semantically incorrehent instructions still works

Wrap Ups: Optimization regarding all kinds of prompts still has research space

- Narrow-Down Searching Space?
- Better Scoring?
- Efficient Sampling?

Visionary : Phenomenon

- reveal that this sequence-dependent instability is common in a variety of tasks and does not vary with model size and annotated sample size.
- the fluctuation range is huge and there is no regularity.
- In addition, the authors find the invariance from the changes, and find the rule of label distribution of prediction results caused by different prompt orders.
- Accordingly, a PROMPT screening method based on entropy is proposed, and the effect is verified.

Visionary : Phenomenon

- Do these good sequences have anything in common in linguistics?
- How do these good orders come about?
- Will it be associated with some language distributions in the pre-trained corpus?
- Why are some validation set data sensitive to the PROMPT sample order?
- What data is sensitive and what data is not?
- Do sensitive and insensitive data have any linguistic characteristics?

- 🔭: Tianqi Shang and Zhiqing Zhong

Visionary - Prompt in Industry

Prompt can be useful in...

- Few-shot/zero-shot scenario: reduce data labeling cost.

  Masked Language Model head -> fewer samples

- Parameter-efficient scenario: provide a better application mode for the deployment and service of hyperscale models.

  Fixed weights of the pre-trained model

  fine-tune prompt with a small number of parameters

Visionary - Prompt in Industry

However,…

- Fine tuning is still needed in vertical areas
- The model effect depends on the selection of prompts
- …