Prompt Tuning

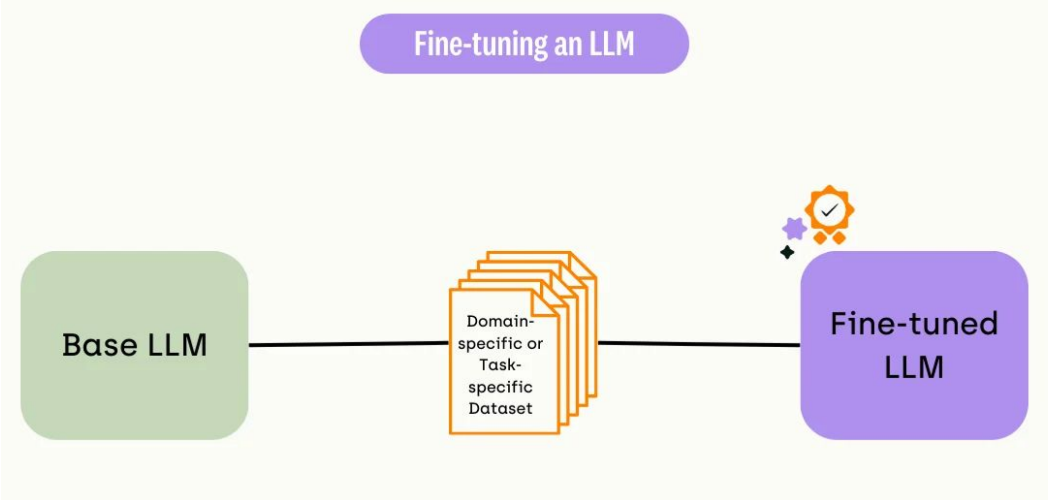
By: Hiva Mohammadzadeh

Outline

- Motivation / Fine Tuning
- Prompt Engineering / “Hard Prompts”
- “Soft Prompts”
- Prefix Tuning
- Prompt Tuning
- Prompt Ensembling
- Chain of Thought Prompting
- Medicine Case Study
- Other Methods / Applications

Fine Tuning / Model Tuning

- **All** the model's parameters are updated and tuned to adapt the model to a specific task
- Use smaller labeled dataset

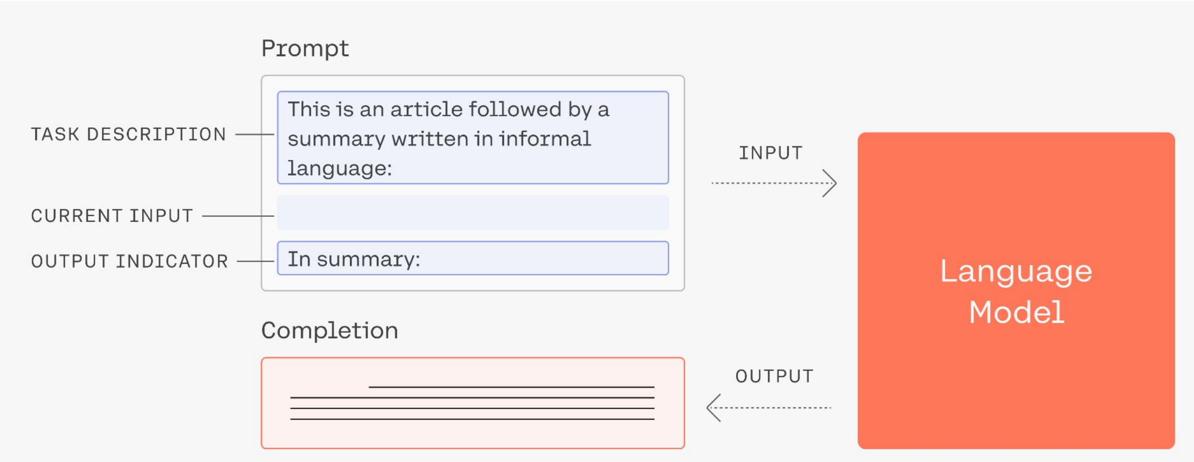


- Problem:
  - Has to be done for each task separately
  - Expensive

https://kili-technology.com/large-language-models-llms/the-ultimate-guide-to-fine-tuning-llms-2023

Prompt Engineering / Hard Prompts

- Develop prompts that guide the LLM to perform specific tasks
- Adds an engineered prompt to the beginning of the input at inference time using the pre trained model
- Discrete input tokens

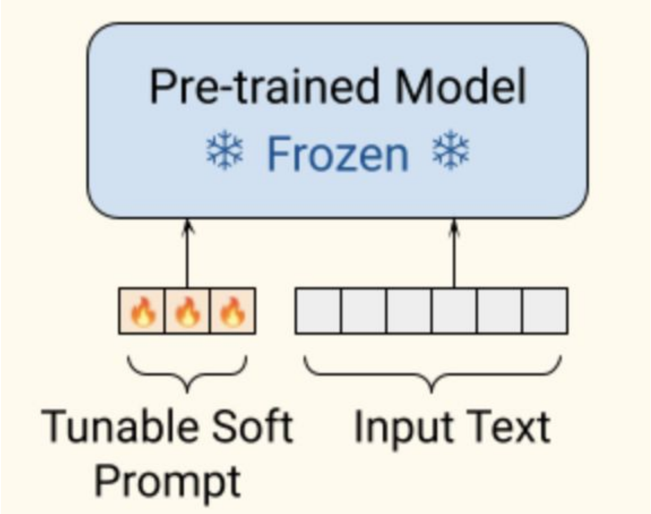


- Problem:
  - Hard to create good human generated prompts
  - Difficult / impossible to know the impact of the prompt

https://docs.cohere.com/docs/prompt-engineering

Soft Prompts

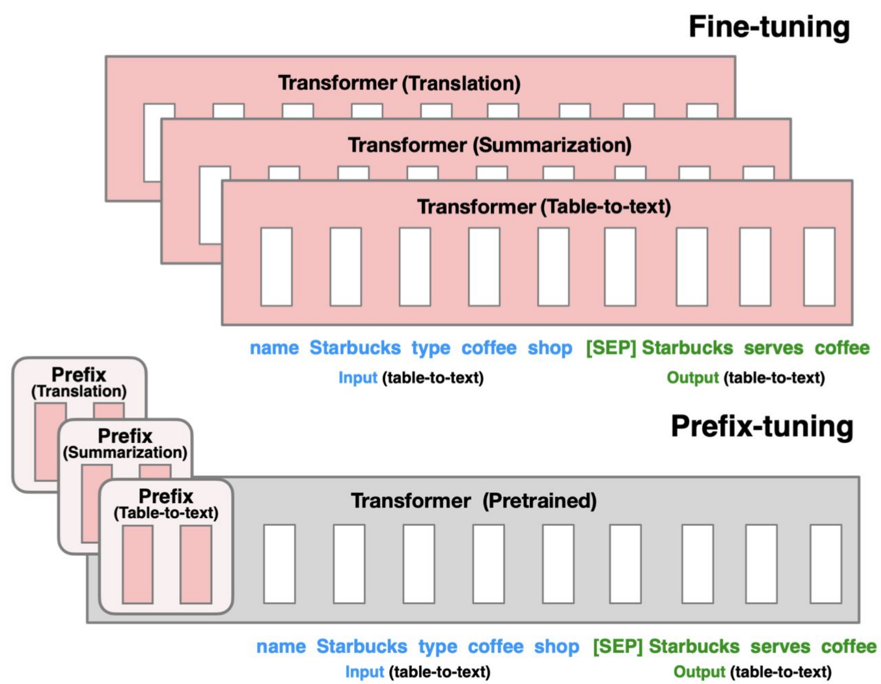
- Embeddings representing the patterns learned by the LLM during training
- Can be high level or task specific
- More effective than hard prompts
- Problem:
  - Soft prompts are not interpretable



https://blog.research.google/2022/02/guiding-frozen-language-models-with.html

Prefix Tuning

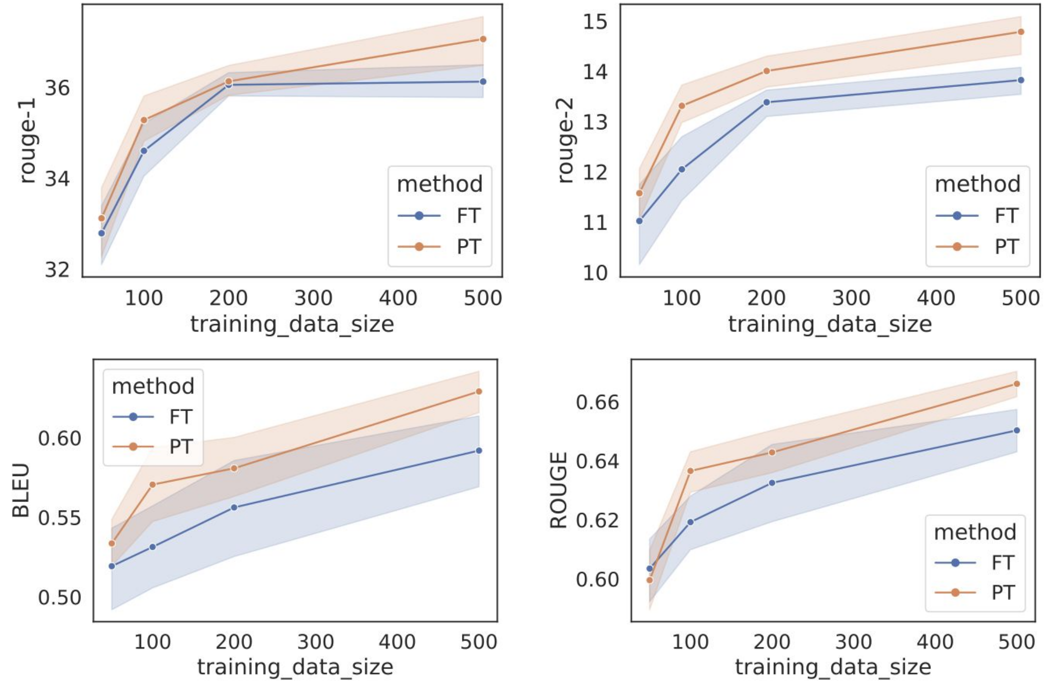
- Adds task-specific token vectors to the input for the K/V blocks that can be trained and updated
- Prefix parameters are inserted in **all** of the layers of the model and optimized by a separate feed-forward network (FFN)



Xiang Lisa Li and Percy Liang. 2021. Prefix- tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*

Prefix Tuning Cont.

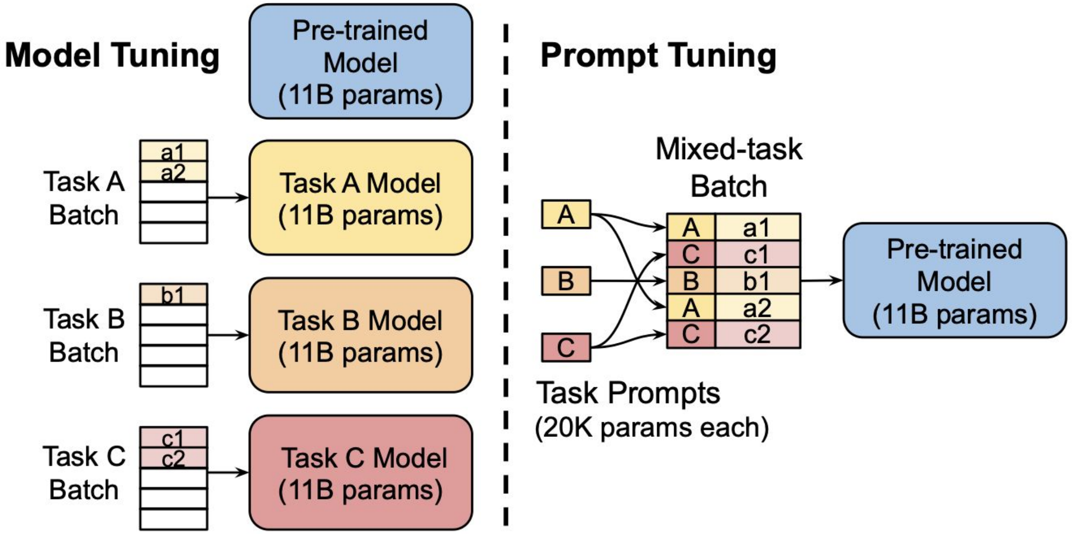
- Pros:
  - Easy to batch requests
  - Improves out-of-domain performance



Xiang Lisa Li and Percy Liang. 2021. Prefix- tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*

Prompt Tuning

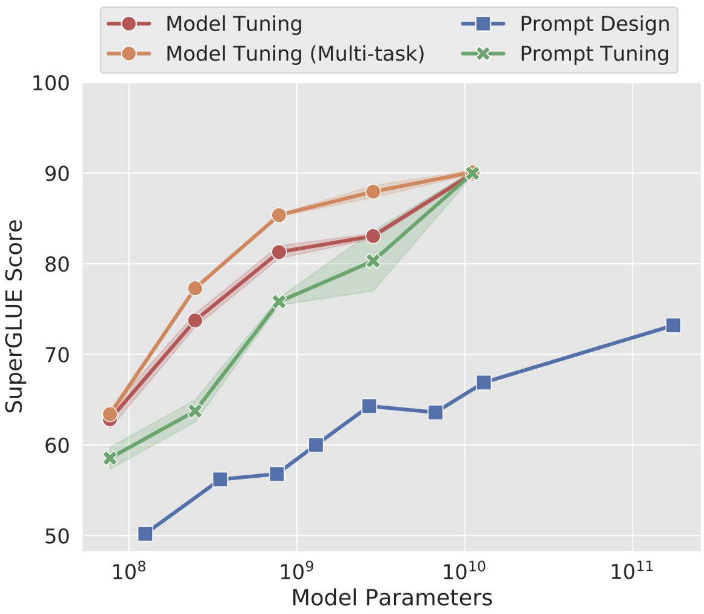
- Adds “soft prompt” into the model's embedding layer only and tuned in back propagation
- More efficient and better results as models grow larger and model parameters scale
- Used for
  - Multi-task Learning
  - Continual Learning



Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Prompt Tuning Cont.

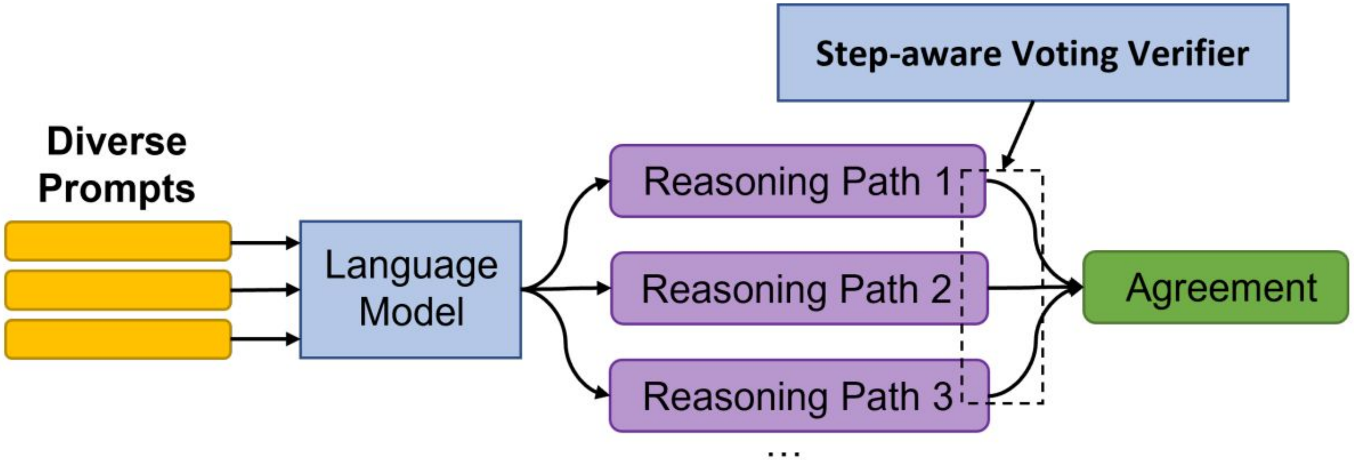
- Pros:
  - Can tune a large model with a small number of parameters
  - Don't need the large labeled datasets
  - Faster and more efficient than fine tuning while achieving same accuracy
  - Universally effective across model scales and NLU tasks.
- Cons:
  - Lack of interpretability of Soft prompts



Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Prompt Ensembling

- A set of few shot prompts that together comprise a “boosted prompt ensemble” using a small dataset that are meant to solve the same problem
- Use when need to guarantee quality of data and output
- The combined predictions do better than the predictions of a single prompt



Pitis, S.; Zhang, M. R.; Wang, A.; and Ba, J. 2023. Boosted Prompt Ensembles for Large Language Models. *arXiv preprint arXiv:2304.05970.*

Prompt Ensembling Cont.

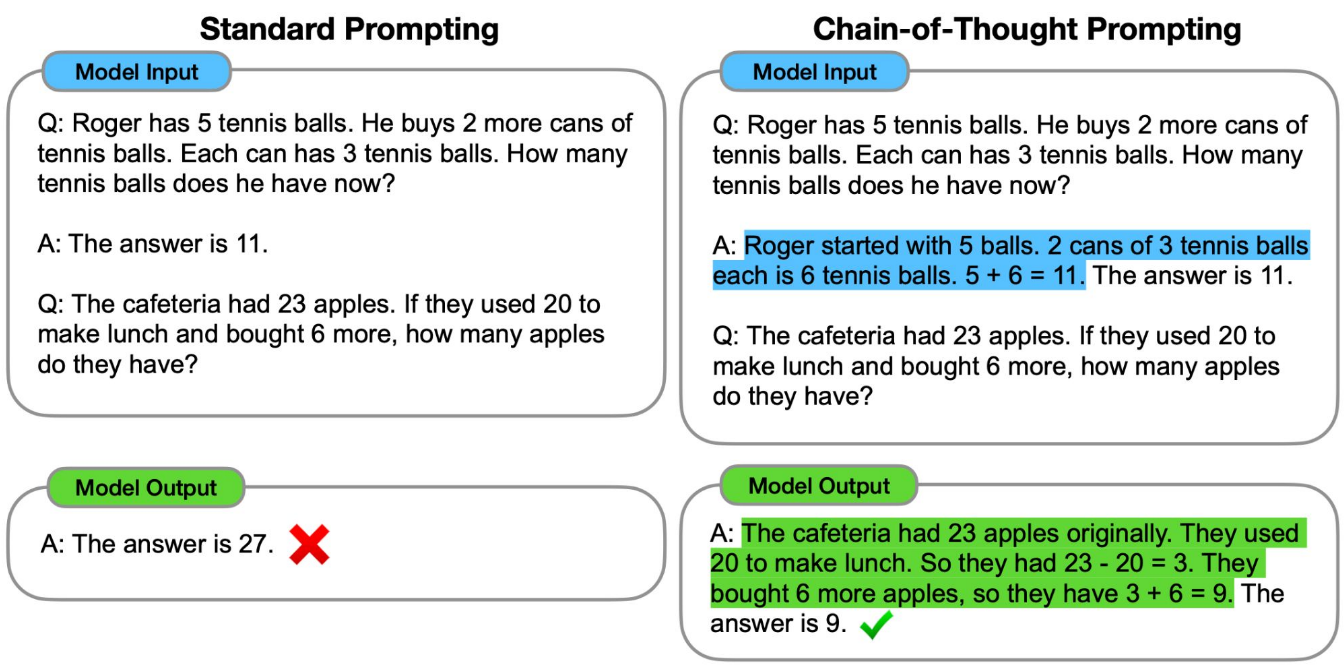
- Methods:
  - Boosting
    - Combines predictions sequentially with each model trying to correct the errors of the previous model to decrease bias
  - Bagging
    - Combines predictions in parallel of different models on different subsets of the training data to decrease variance
- Pros:
  - Helps with tackling hallucination and instability of LLMs

Datasets	SNLI	MNLI	QNLI	RTE	Ethos	Liar	ArSarcasm
Single Prompt	0.587	0.660	0.660	0.720	0.833	0.535	0.511
Single Prompt (CoT)	0.575	0.685	0.660	0.731	0.804	0.549	0.525
Synonym Ensemble	0.580	0.746	0.720	0.659	0.812	0.572	0.569
PromptBoosting	0.619	0.574	0.631	0.673	-	-	-
APO	-	-	-	-	0.964	0.663	0.873
APO*	-	-	-	-	0.947	0.658	0.639
Ours	**0.647**	**0.767**	**0.793**	**0.753**	**0.963**	**0.744**	**0.739**

Chenrui Zhang, Lin Liu, Jinpeng Wang, Chuyuan Wang, Xiao Sun, Hongyu Wang, and Mingchen Cai. Prefer: Prompt ensemble learning via feedback-reflect-refine. *arXiv preprint arXiv:2308.12033*, 2023.

Chain of Thought Prompting

- Prompting LLMs with intermediary reasoning steps
- Coherent series of intermediate reasoning steps that lead to the final answer for a problem.

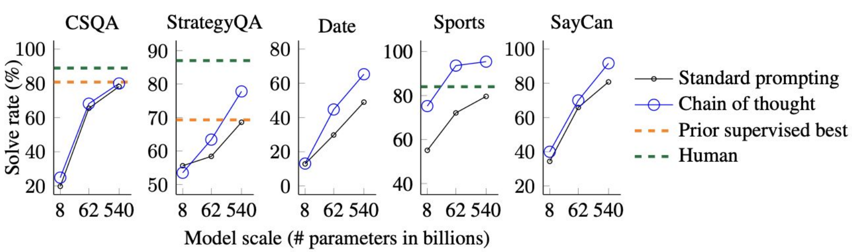


Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

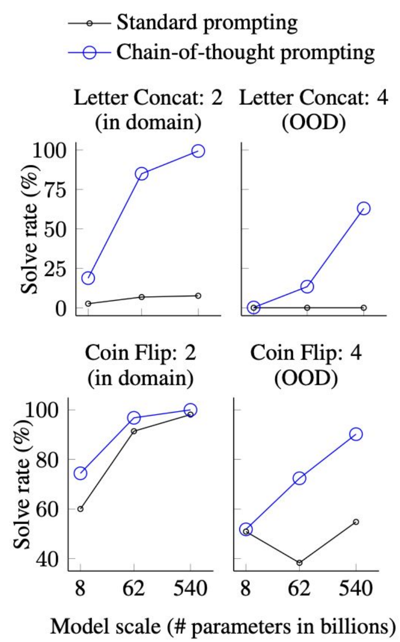
Chain of Thought Prompting Cont.

- Pros:
  - Better performance in reasoning tasks
  - Better interpretability

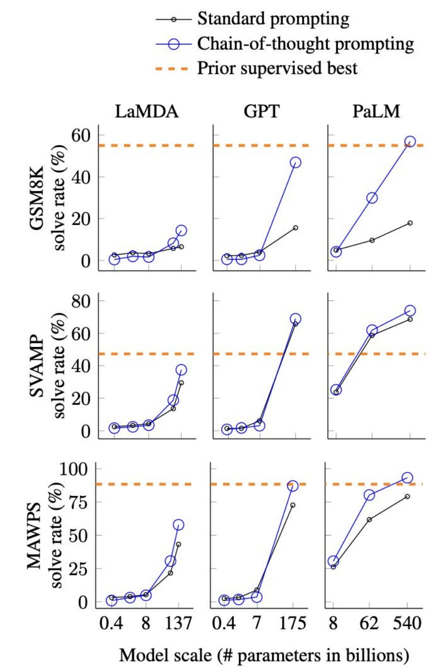
Symbolic Reasoning



Commonsense Reasoning



Arithmetic



Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

Case Study in Medicine

- Goal: Boost performance using prompting techniques

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*, 2023.

Case Study in Medicine

- Goal: Boost performance using prompt engineering
- **Dynamic Few Shot**
  - Use K-NN clustering in the embedding space to find the five best few shot examples
  - Leverages the training data like fine tuning

Inference Time:
Compute the embedding v_Q for the test question Q.
Select the 5 most similar examples $\{(v_{Q_i}, C_{Q_i}, A_{Q_i})\}_{i=1}^{5}$ from the preprocessed training data using KNN, with the distance function as the cosine similarity: $\mathrm{dist}(v_q, v_Q) = 1 - \frac{\langle v_q, v_Q \rangle}{\|v_q\| \|v_Q\|}$.
Format the 5 examples as context $\mathcal{C}$ for the LLM.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*, 2023.

Case Study in Medicine

- Goal: Boost performance using prompt engineering
- **Dynamic Few Shot**
  - Use K-NN clustering in the embedding space to find the five best few shot examples
  - Leverages the training data like fine tuning

Inference Time:
Compute the embedding v_Q for the test question Q.
Select the 5 most similar examples $\{(v_{Q_i}, C_{Q_i}, A_{Q_i})\}_{i=1}^{5}$ from the preprocessed training data using KNN, with the distance function as the cosine similarity: $\text{dist}(v_q, v_Q) = 1 - \frac{\langle v_q, v_Q \rangle}{\|v_q\| \|v_Q\|}$.
Format the 5 examples as context $\mathcal{C}$ for the LLM.

- **Self-Generated Chain of Thought**
  - Ask GPT4 to do the COT

Self-generated Chain-of-thought Template

```
## Question: {{question}}
{{answer_choices}}
## Answer
```
model generated chain of thought explanation
Therefore, the answer is [*final model answer (e.g. A,B,C,D)*]

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*, 2023.

Case Study in Medicine Cont.

- **Choice Shuffling Ensemble**
  - Get rid of position bias in multiple choice questions
  - Shuffle the relative order of the answer choices before generating each reasoning path.
  - Select the most consistent answer.
  - 5 API Calls for COT

for 5 times **do**
  Shuffle the answer choices of the test question.
  Generate a chain-of-thought C_q^k and an answer A_q^k with the LLM and context $\mathcal{C}$.
end for
Compute the majority vote of the generated answers $\{A_q^k\}_{k=1}^K$:

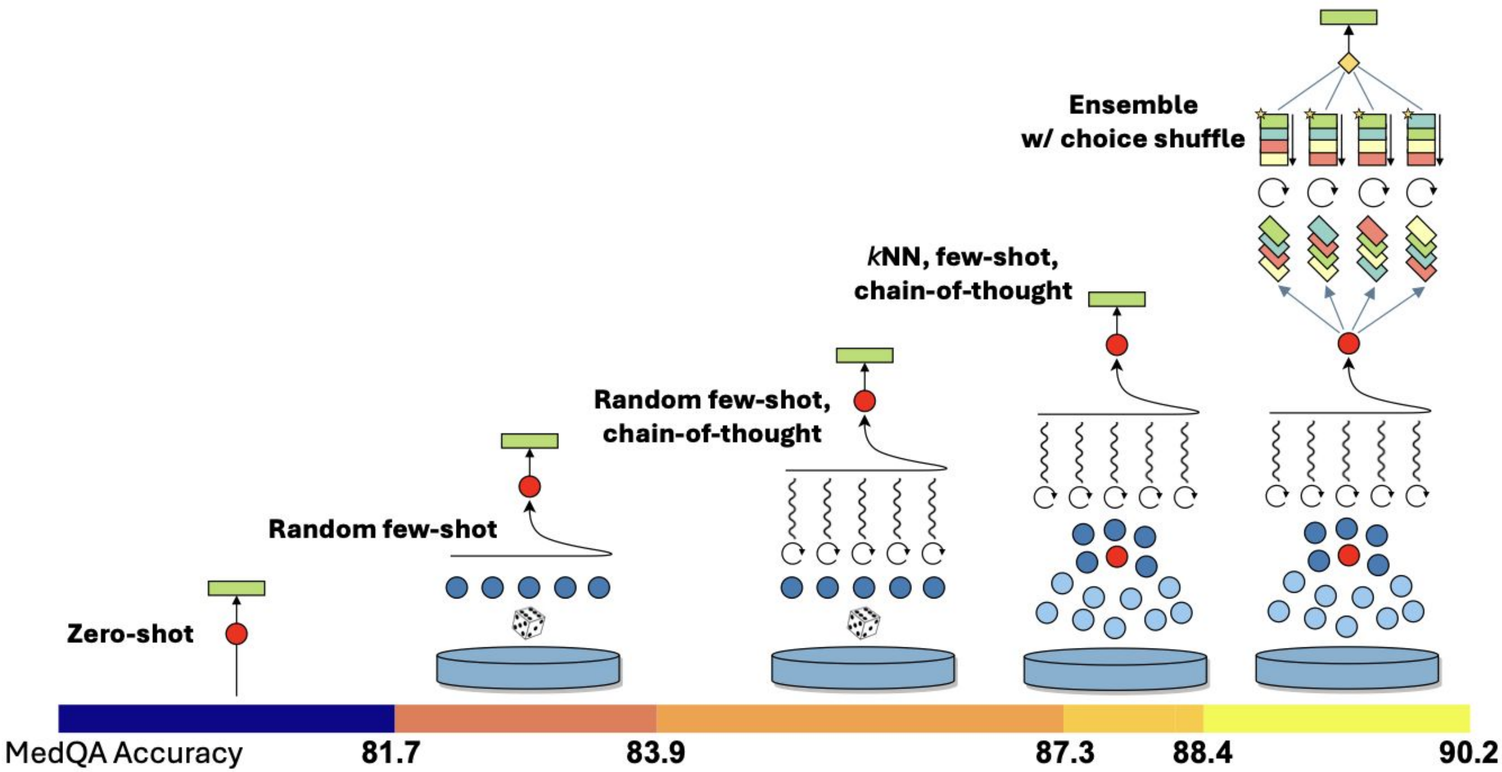
$$A^{\text{Final}} = \text{mode}(\{A_q^k\}_{k=1}^K),$$

where $\text{mode}(X)$ denotes the most common element in the set X.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*, 2023.

Case Study in Medicine Cont.

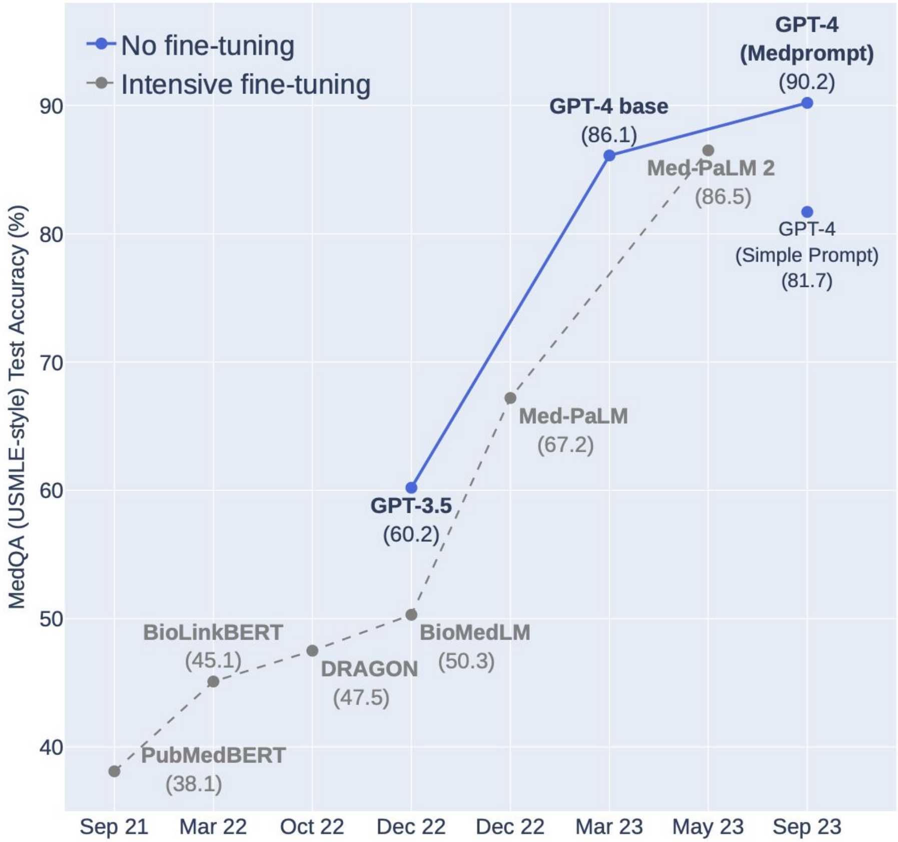
- MedPrompt



Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*, 2023.

Case Study in Medicine Cont.

- Results on MedQA



Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*, 2023.

Other Methods / Applications

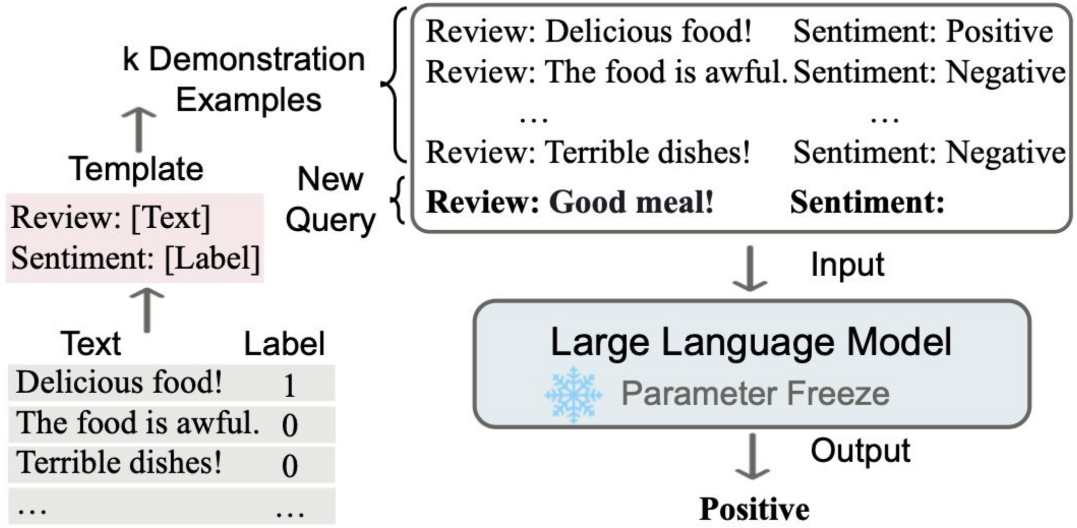
- Zero Shot Chain of Thought
  - https://arxiv.org/pdf/2205.11916.pdf
- Automatic Chain of Thought
  - https://arxiv.org/abs/2210.03493
- Self Consistency with COT
  - https://arxiv.org/abs/2203.11171
- Visual In-Context Learning
  - https://arxiv.org/abs/2301.13670, https://arxiv.org/abs/2304.04748
- Multi-Modal In-Context Learning
  - https://arxiv.org/abs/2204.14198, https://arxiv.org/abs/2206.06336
- Speech In-Context Learning
  - https://arxiv.org/abs/2303.03926
- Graph Classification prompt tuning
  - https://arxiv.org/abs/2310.17394,https://www.sciencedirect.com/science/article/abs/pii/S030645732300376X

References

- https://medium.com/@shahshreyansh20/prompt-tuning-a-powerful-technique-for-adapting-llms-to-new-tasks-6d6fd9b83557#:~:text=Prompt%20tuning%20is%20a%20technique,towards%20generating%20the%20desired%20output.
- https://research.ibm.com/blog/what-is-ai-prompt-tuning
- https://cobusgreyling.medium.com/prompt-engineering-text-generation-large-language-models-3d90c527c6d5
- https://huggingface.co/docs/peft/conceptual_guides/prompting
- https://arxiv.org/pdf/2101.00190.pdf
- https://arxiv.org/pdf/2311.16452.pdf
- https://docs.cohere.com/docs/prompt-engineering
- https://kili-technology.com/large-language-models-llms/the-ultimate-guide-to-fine-tuning-llms-2023
- https://docs.cohere.com/docs/prompt-engineering
- https://blog.research.google/2022/02/guiding-frozen-language-models-with.html
- https://arxiv.org/abs/2104.08691
- https://arxiv.org/abs/2304.05970
- https://arxiv.org/abs/2201.11903
- https://arxiv.org/abs/2311.16452

In Context Learning

- Learn from analogy
- Allows language models to learn tasks given only a few examples in the form of demonstration.



Q. Dong, L. Li, D. Dai, C. Zheng, Z. Wu, B. Chang, X. Sun, J. Xu, and Z. Sui, "A survey for in-context learning," *arXiv preprint arXiv:2301.00234*, 2022.