RIT

LLM Fine-Tuning

Reza Fayyazi

Why Adaptation?

Base LLM

Predicts next word, based on text training data

Once upon a time, there was a unicorn
that lived in a magical forest with
all her unicorn friends

What is the capital of France?
What is France's largest city?
What is France's population?
What is the currency of France?

Instruction Tuned LLM

Tries to follow instructions

Fine-tune on instructions and
good attempts at following those instructions.

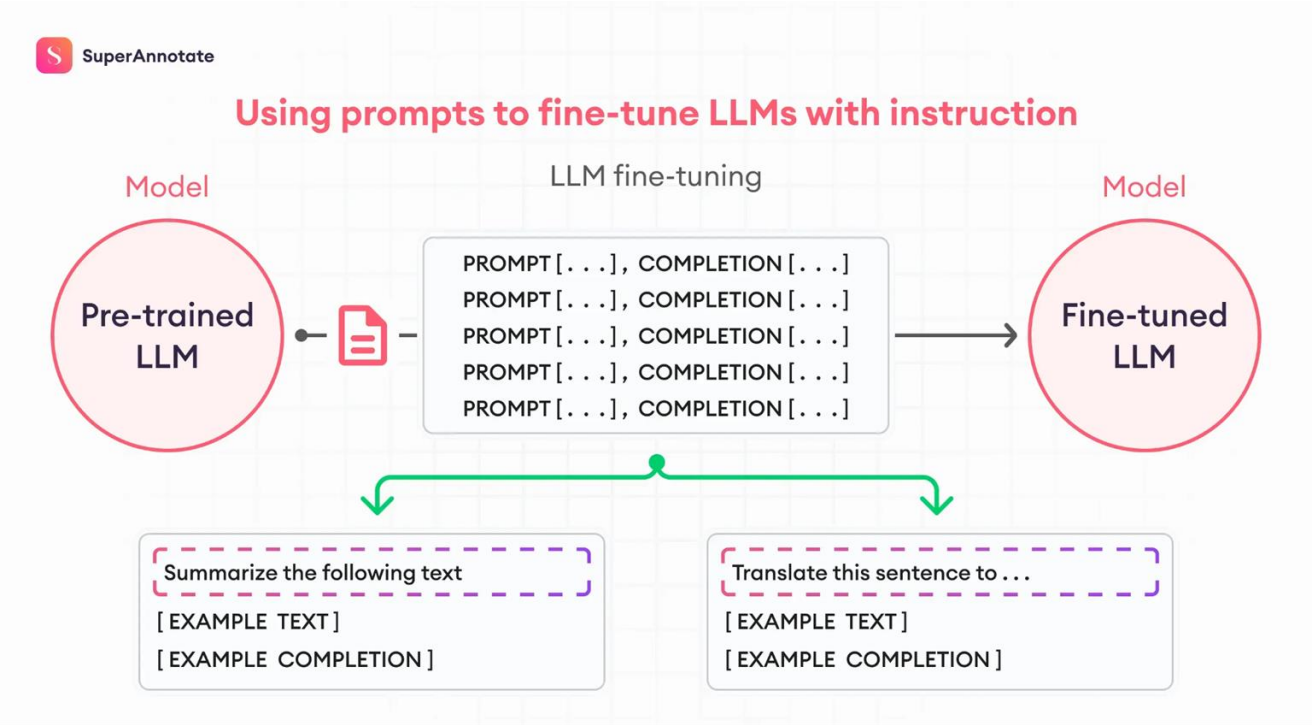
RLHF: Reinforcement Learning with
Human Feedback

Helpful, Honest, Harmless

What is the capital of France?
The capital of France is Paris.

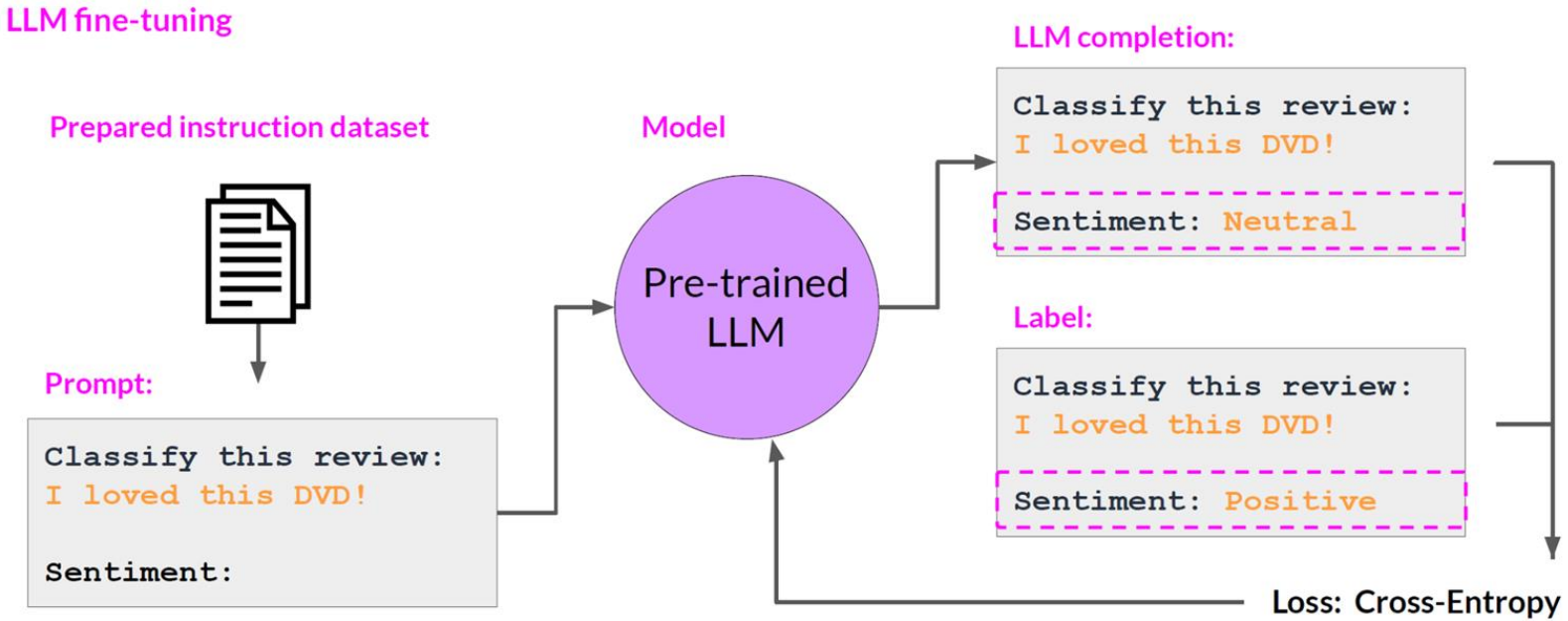
[1] https://www.deeplearning.ai/short-courses/chatgpt-prompt-engineering-for-developers/

LLM Fine-Tuning Process



[2] https://www.superannotate.com/blog/llm-fine-tuning

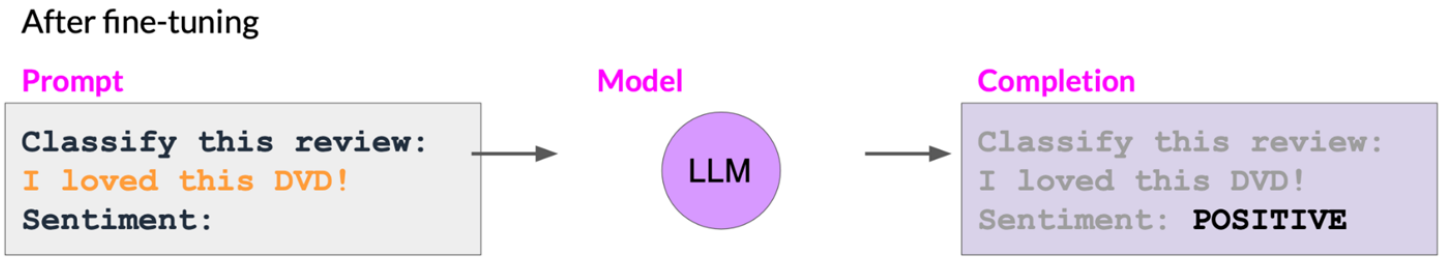
LLM Fine-Tuning Process



[3] https://www.coursera.org/learn/generative-ai-with-llms/

Catastrophic Forgetting

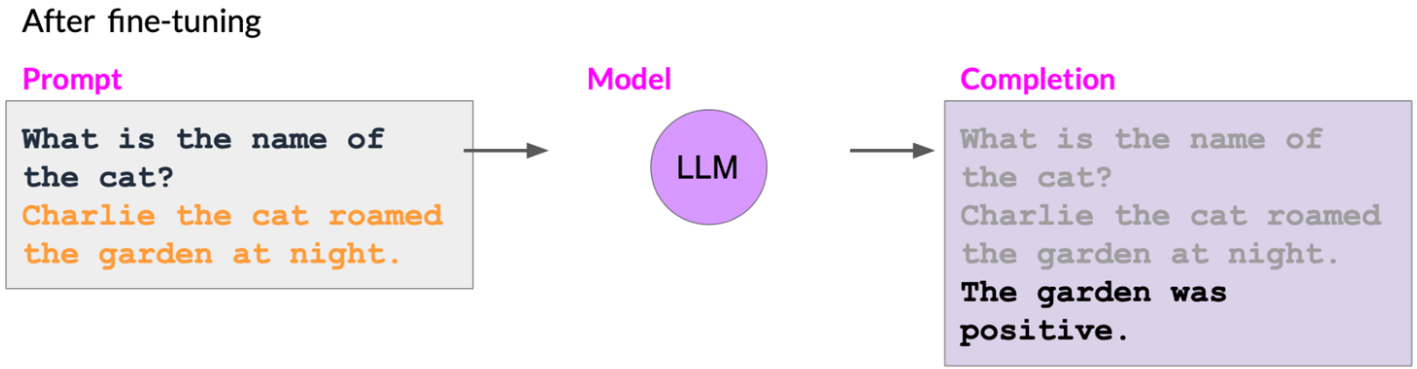
- Fine-tuning can significantly increase the performance of a model on a specific task…



[3] https://www.coursera.org/learn/generative-ai-with-llms/

Catastrophic Forgetting

- ...but can lead to reduction in ability on other tasks



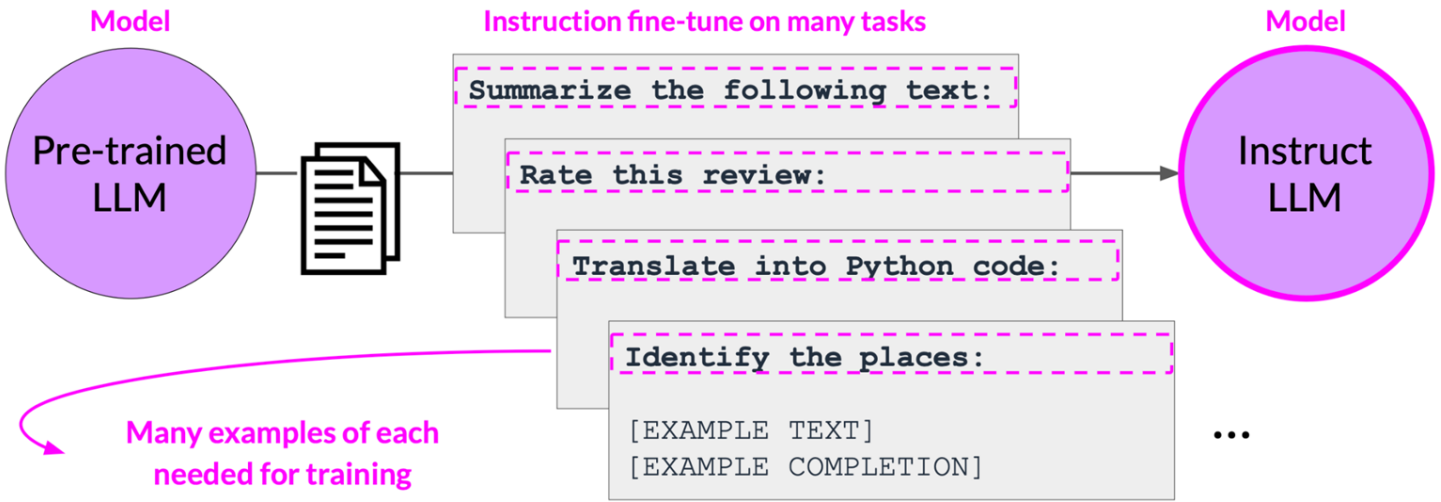
[3] https://www.coursera.org/learn/generative-ai-with-llms/

How to Avoid Catastrophic Forgetting

- Fine-tune on multiple tasks at the same time
- Consider Parameter Efficient Fine-tuning (PEFT)

[3] https://www.coursera.org/learn/generative-ai-with-llms/

Multi-task Fine-Tuning



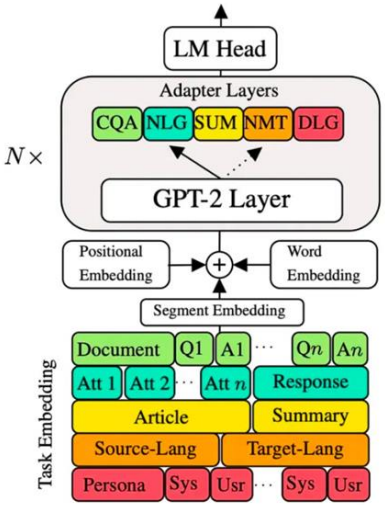
[3] https://www.coursera.org/learn/generative-ai-with-llms/

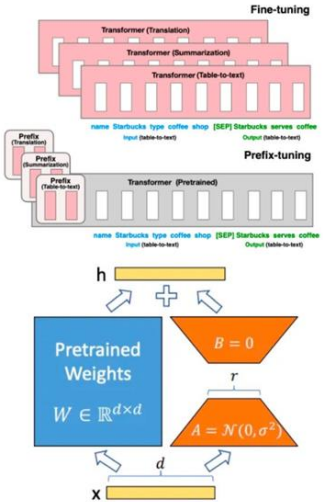
Computational challenges still remain!

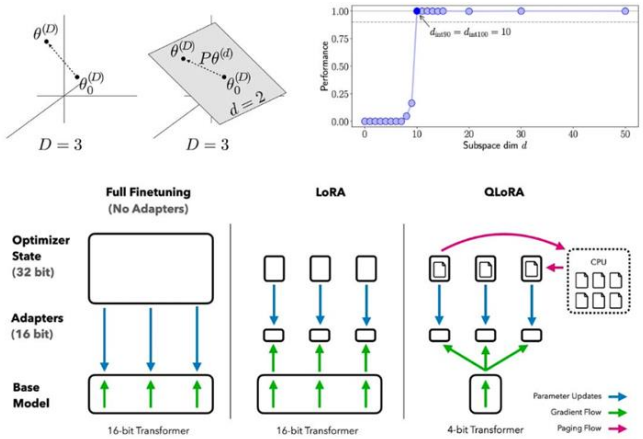
Parameter-Efficient Fine-Tuning (PEFT) Techniques

DEEP (LEARNING) FOCUS

Easily Train a Specialized LLM: PEFT, LoRA, QLoRA, LLaMA-Adapter, and More

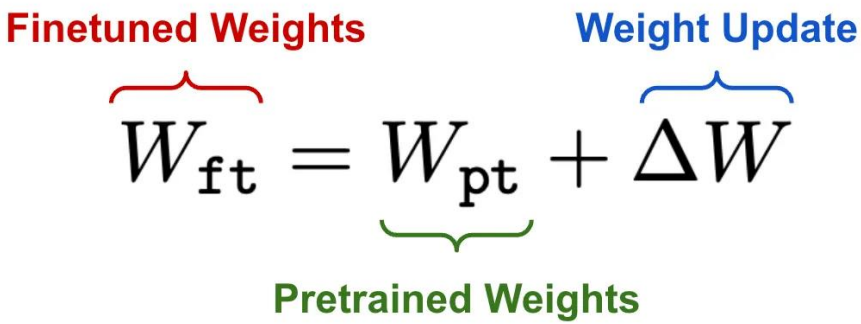






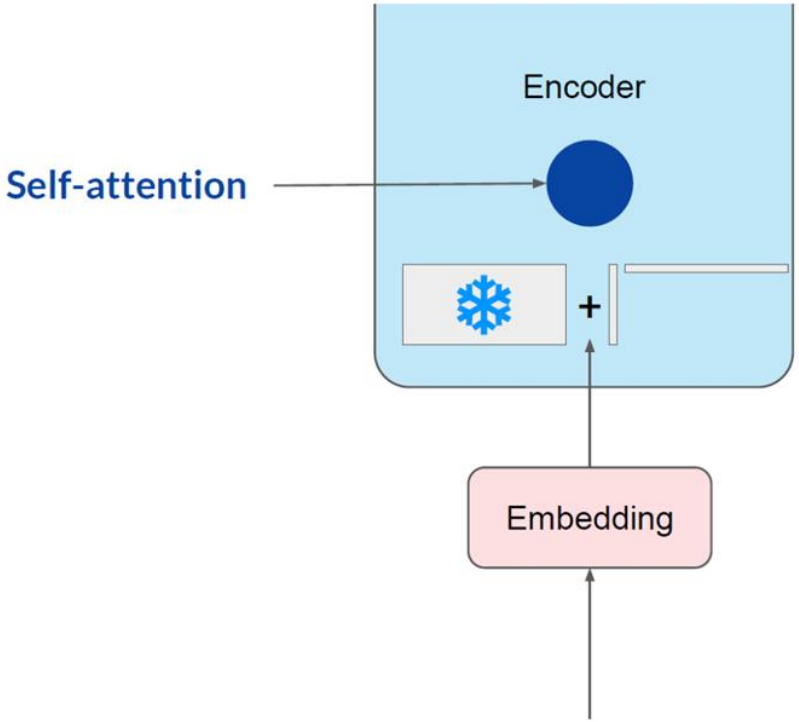
[4] https://cameronrwolfe.substack.com/p/easily-train-a-specialized-llm-peft

PEFT



[3] https://www.coursera.org/learn/generative-ai-with-llms/

LoRA [9]



1. Freeze most of the original LLM weights.
2. Inject 2 **rank decomposition matrices**
3. Train the weights of the smaller matrices

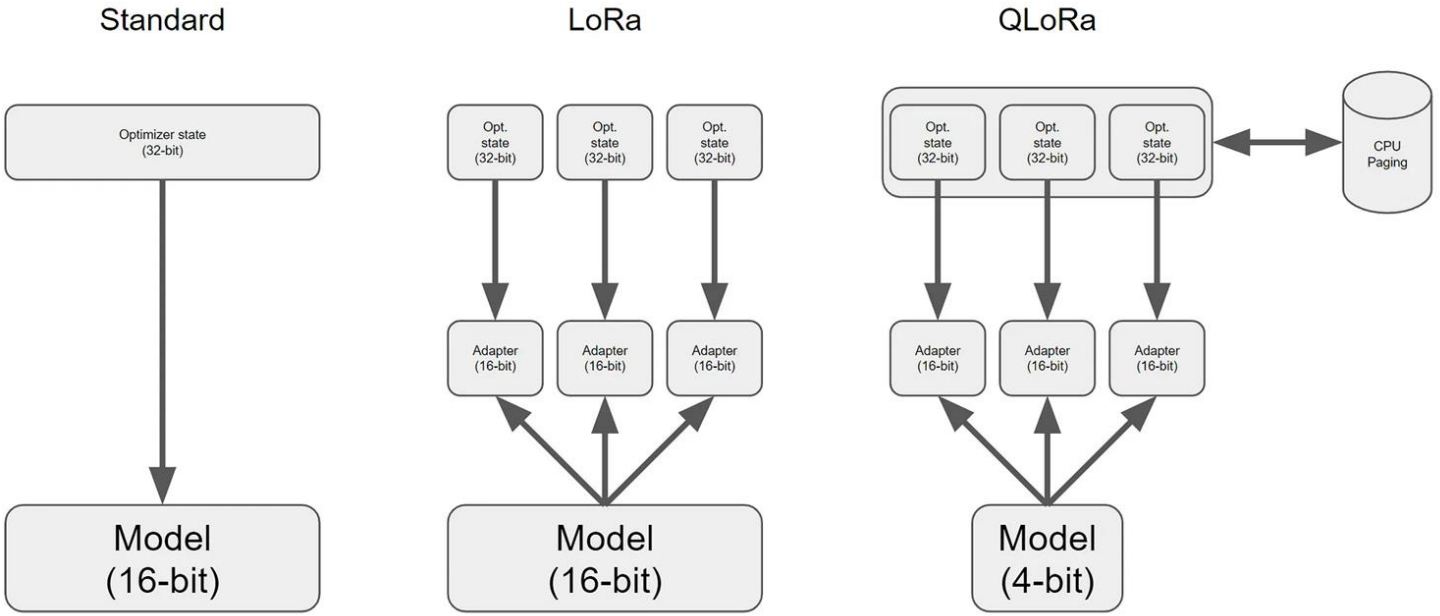
Steps to update model for inference

1. Matrix multiply the low rank matrices

2. Add to original weights

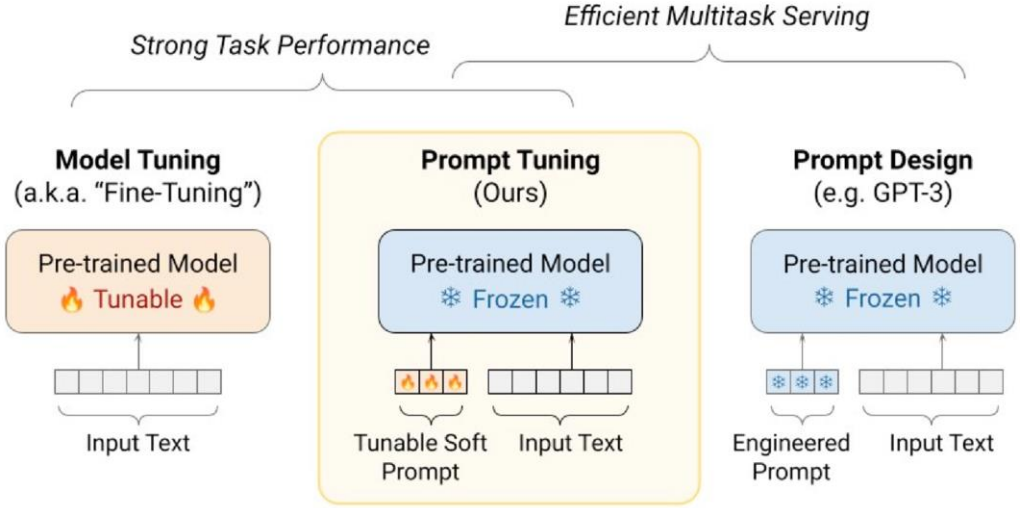
[3] https://www.coursera.org/learn/generative-ai-with-llms/
[5] Hu, E. J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. In International Conference on Learning Representations.
[6] https://www.youtube.com/watch?v=dA-NhCtrrVE
[7] https://www.youtube.com/watch?v=t509sv5MT0w

QLoRA [9]



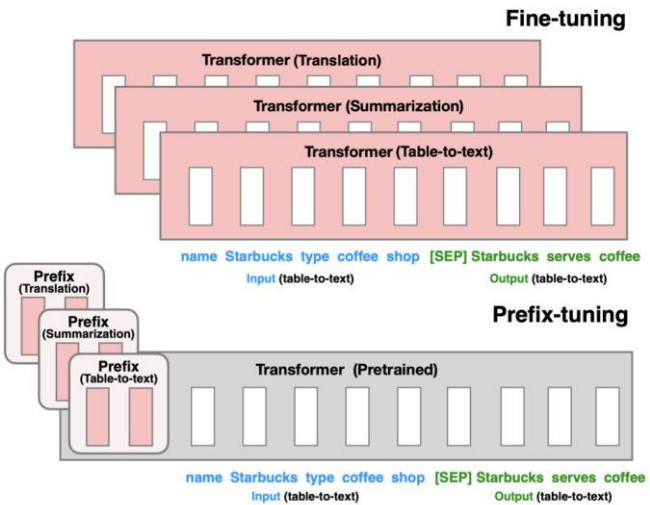
[8] Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2024). Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, *36*.
[9] https://towardsdatascience.com/qlora-fine-tune-a-large-language-model-on-your-gpu-27bed5a03e2b

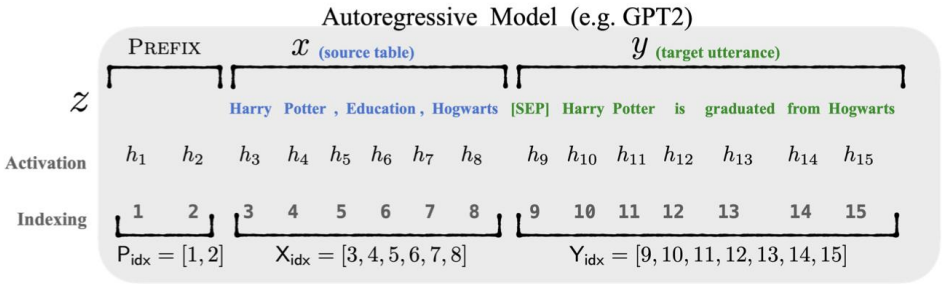
Prompt-Tuning [9]



[10] Lester, B., Al-Rfou, R., & Constant, N. (2021, November). The Power of Scale for Parameter-Efficient Prompt Tuning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 3045-3059).
[11] https://cobusgreyling.medium.com/prompt-tuning-hard-prompts-soft-prompts-49740de6c64c
[12] Wang, Y., Chauhan, J., Wang, W., & Hsieh, C. J. (2024). Universality and limitations of prompt tuning. Advances in Neural Information Processing Systems, 36.'
[13] https://www.youtube.com/watch?v=HkZOGGvZzg4
[14] https://fnl.es/Science/Papers/Prompt+Engineering/Prompt+Tuning

Prefix-Tuning [14]





$$h_i = \begin{cases} P_\theta[i,:], & \text{if } i \in \mathsf{P}_{\mathsf{idx}}, \\ \mathrm{LM}_\phi(z_i, h_{<i}), & \text{otherwise.} \end{cases}$$

[14] Li, X. L., & Liang, P. (2021, August). Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 4582-4597).

Thank you!

Reza Fayyazi
rf1679@rit.edu