How do Language Models Solve Math Problems?

EECS487 FA23

UNIVERSITY OF MICHIGAN

Xiaoyan Bai smallyan@umich.edu

Yanran Lin yanran@umich.edu

Xinyu Bao xinyubao@umich.edu

Problem Statement and Overveiw

Problem Statement:

- Large Language Models are bad at correctly answering math problems.
- The most existing works only focus on verifying the right answers instead of analyzing the inner structure.
- Most evaluation did on attention are proved to be not as interpretative as we thought, since researchers found models rely less on the attention than we thought.

Overview:

- In this work, we analyzed what contributes to GPT2 model's decision making processing using LIME [1] and residual stream analysis.
- Residual connection in the Transformers can be thought as a information stream. LIME[1] is able to explain any black box classifier.
- With the improved transparency of the model, we can figure out suitable ways to improve performance.
- In the future, we will try generation tasks and other models.

Data

The dataset we chose is MathQA [2]. It is a mathematical reasoning dataset containing math word problems. The dataset includes annotated formula and rationale for each problem. We used LIME [1] and residual stream analysis to analyze 10 examples , each with 5 types of prompts, before finetune and after finetune.

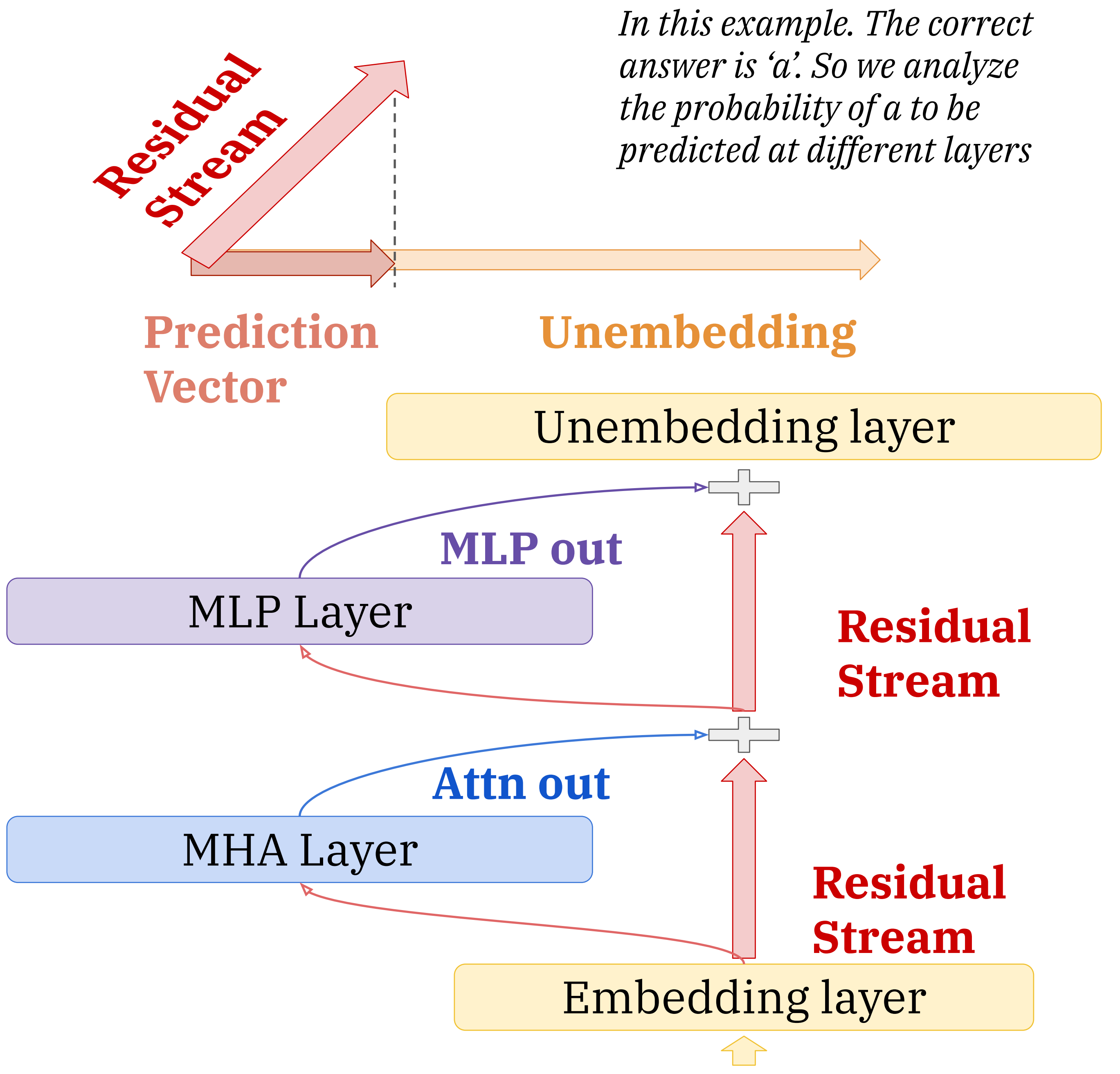
Problem + options	**Problem+rationale+options**
Problem+ rationale + answer + options	**Annotated_formula+options**
Problem + Annotated_formla + options	

Method1: Residual Analysis

We analyzed the accumulated **Residual Stream**. We checked how the **MLP out**, which is the output of MLP layer, and **Attn Out**, which is the output of self-attention layer, contribute to the **Residual Stream**.

We projected the **Residual Stream** to the **Unembedding** to get **Prediction Vector** . And we applied softmax to analyze the change of the probability at the last timestamp among different layers.

Probability = softmax(Prediction Vector [tokenizer(true predicted word)])

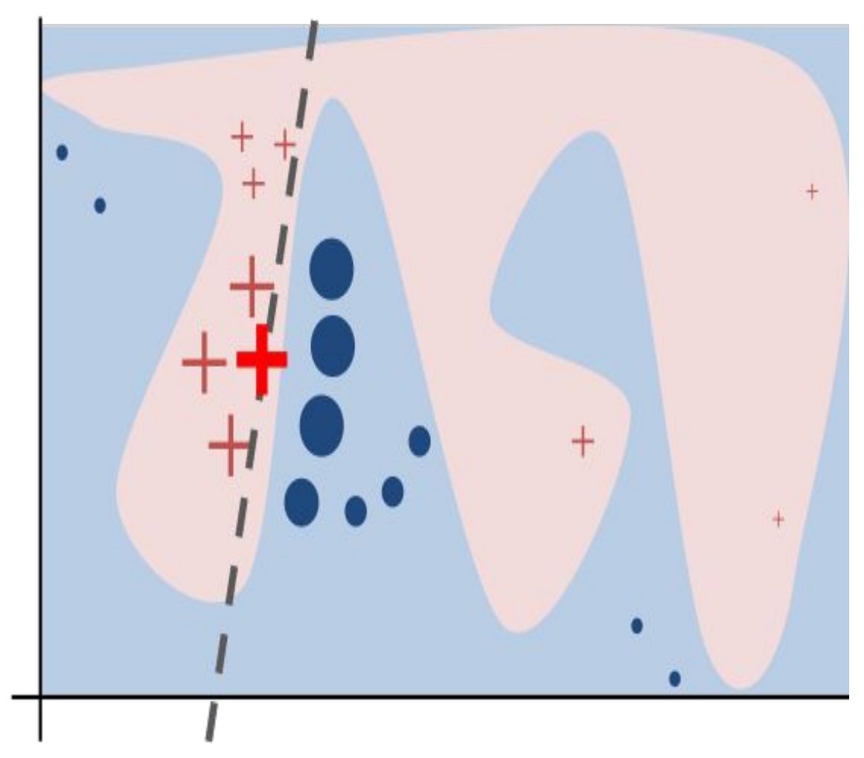


In this example. The correct answer is 'a'. So we analyze the probability of a to be predicted at different layers

*Prompt: "there are 10 girls and 20 boys in a classroom . what is the ratio of girls to boys ? Option: a) 1 / 2 , b) 1 / 3 , c) 1 / 5 , d) 10 / 30 , e) 2 / 5 Answer:"

Method2: LIME

We utilized **Local Interpretable Model-agnostic Explanations (LIME)** [1] to analyze what contribute to GPT2's final decision on math problems. We generated explanation for 10 examples in the MathQA[2] dataset using **LIME**. The visualizations are generated for prediction probabilities of each choice and the tokens contributing the most to the prediction.



Result1: Residual Analysis

- With original prompts (Fig.1), decisions are made in higher layers.
- With chain of thought prompting (Fig.2), differences appear in early layers.
- From decomposed residual stream (Fig 3), in later layers **Attn_Out** contributes more.

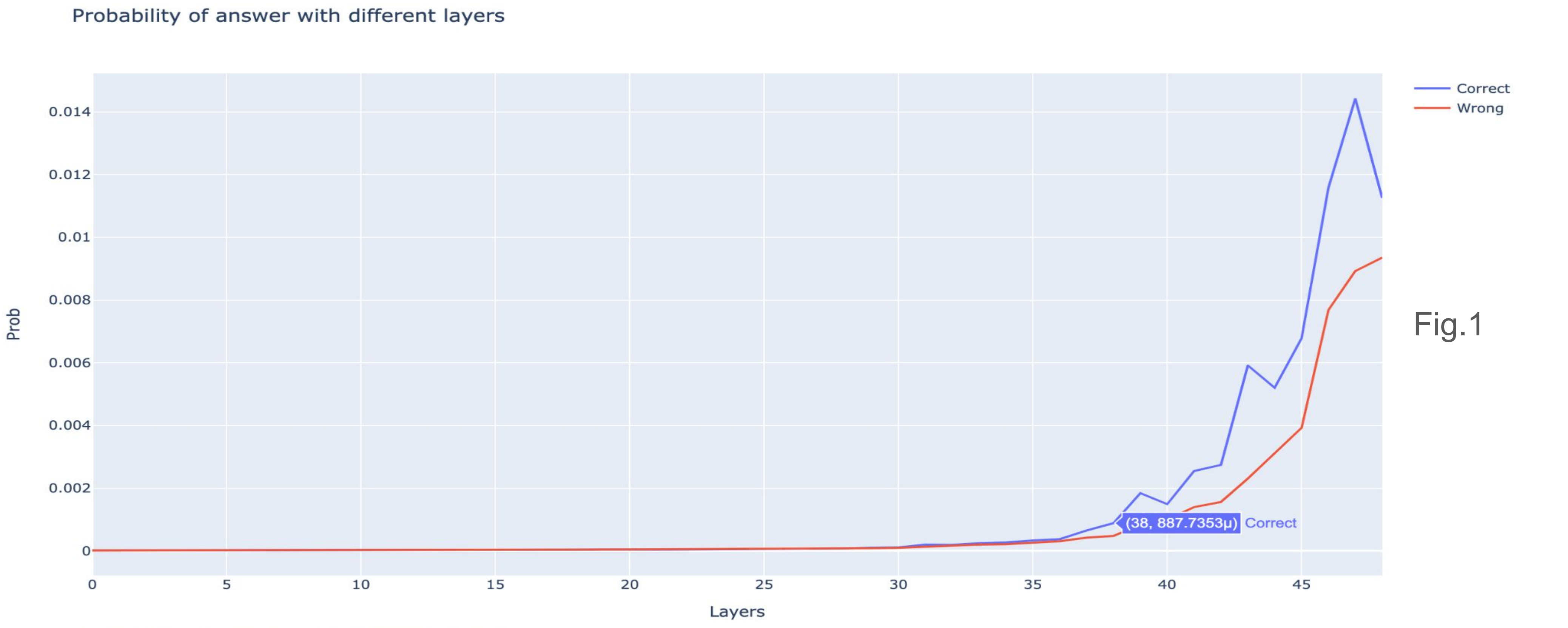


Fig.1

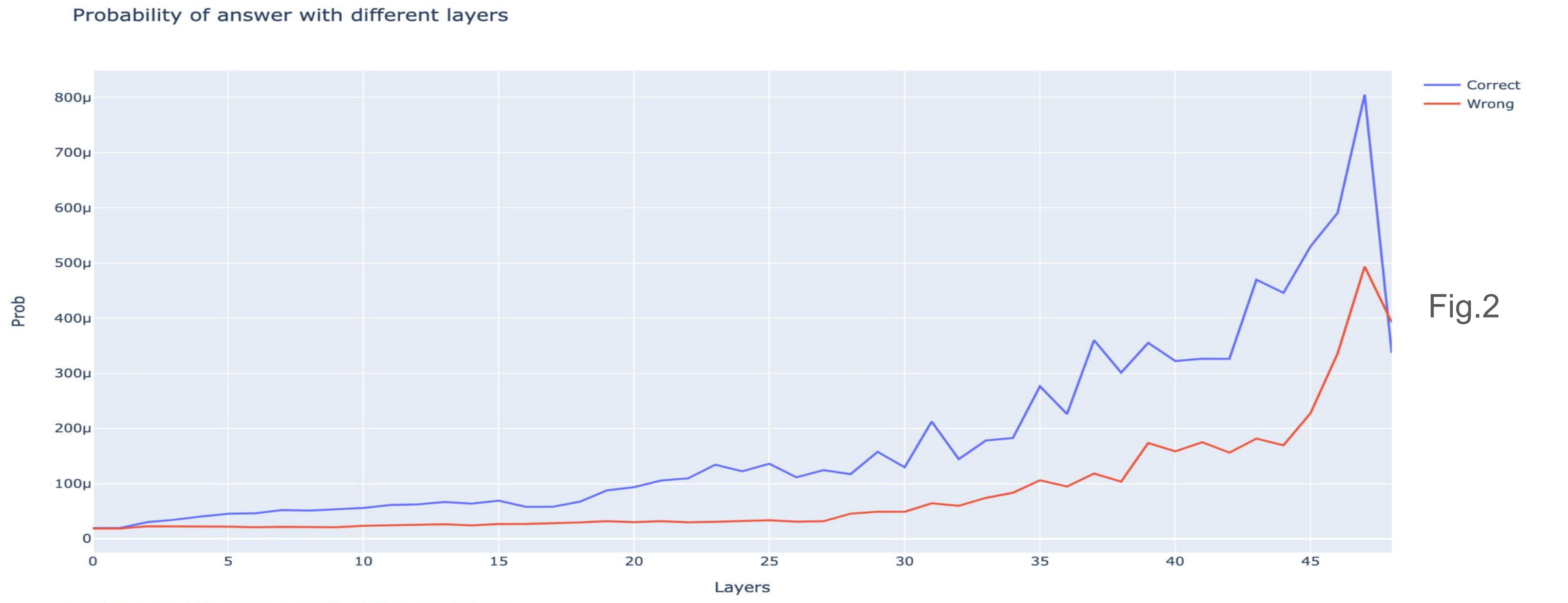


Fig.2

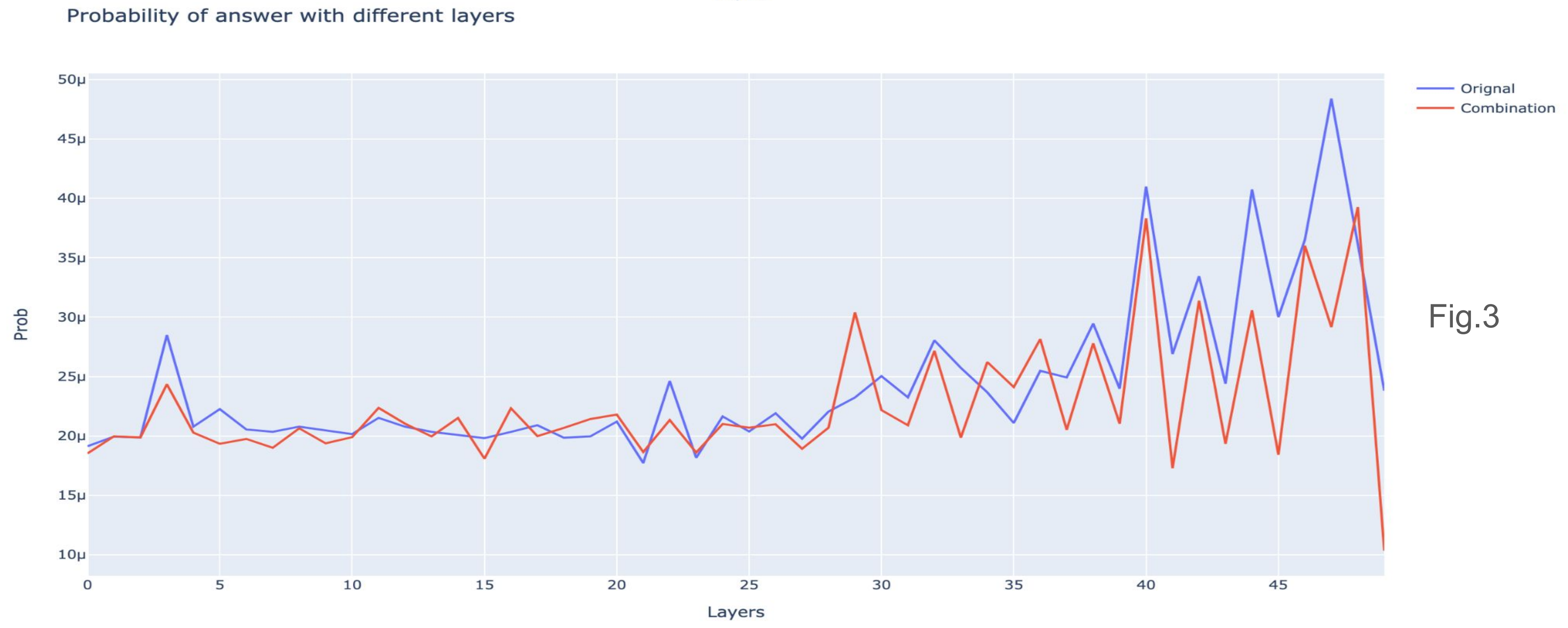


Fig.3

Result2: LIME

- LIME shows that the pre-trained model doesn't understand the classification task (Fig.1).
- LIME shows that the fine-tuned model understands the classification task better (Fig.2).

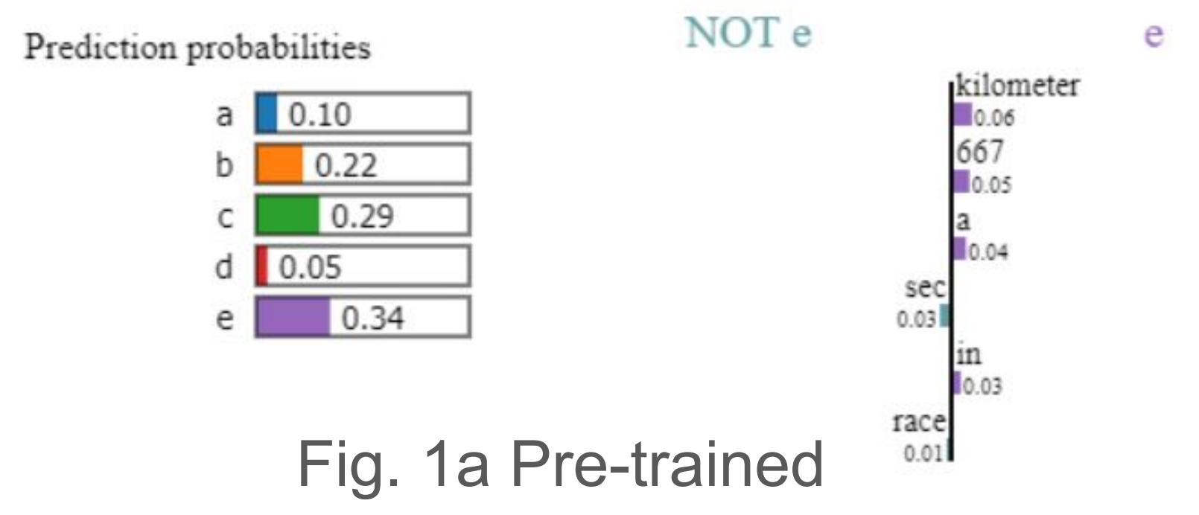


Fig. 1a Pre-trained

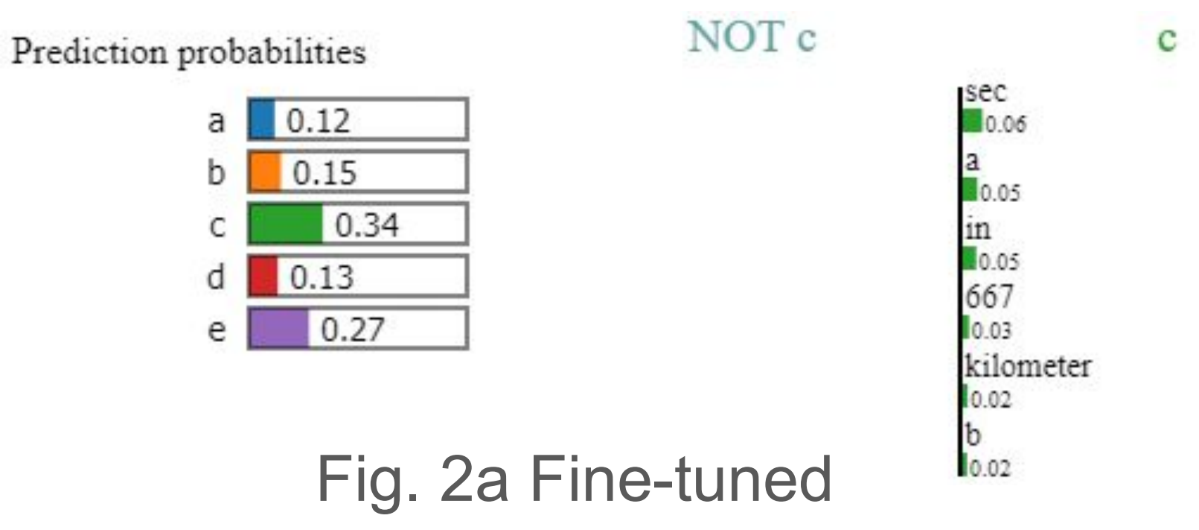
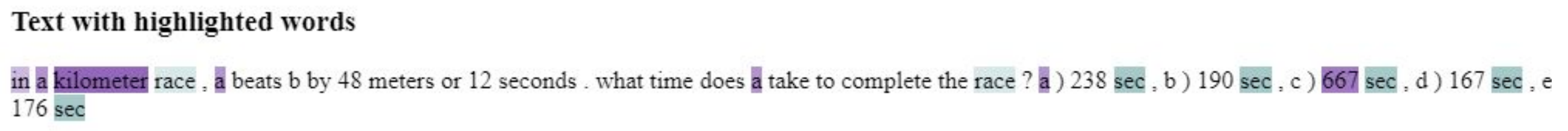


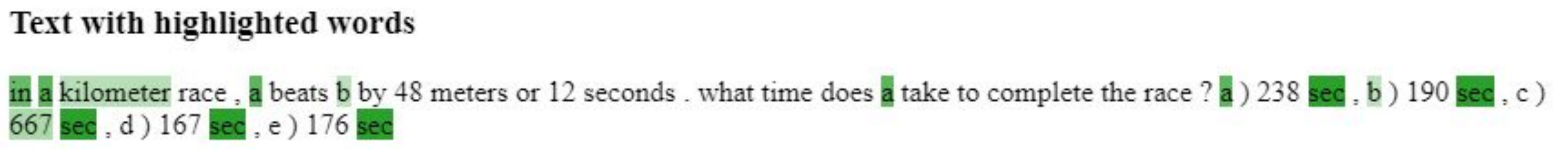
Fig. 2a Fine-tuned



Text with highlighted words

in a kilometer race , a beats b by 48 meters or 12 seconds . what time does a take to complete the race ? a) 238 sec , b) 190 sec , c) 667 sec , d) 167 sec , e 176 sec

Fig. 1b Pre-trained



Text with highlighted words

in a kilometer race , a beats b by 48 meters or 12 seconds . what time does a take to complete the race ? a) 238 sec , b) 190 sec , c) 667 sec , d) 167 sec , e) 176 sec

Fig 2b. Fine-tuned

Conclusion and Future Work

- In conclusion, the pre-trained GPT2 model doesn't understand the classification task and the fine-tuned one works slightly better. The accuracy of the fine-tuned one is still not significantly better than a random guess. From the residual stream experiment, the decision is made in later layers and from attention layers.
- In the future, we plan to switch from the classification task to a generation task for the MathQA dataset to analyze how the probability of the next tokens change. We may test the model with simpler mathematical problems to see if the model understand the task. We can further analyze which head contributes more to the decision.

[1] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016.

[2] Amini, Aida, et al. "Mathqa: Towards interpretable math word problem solving with operation-based formalisms." arXiv preprint arXiv:1905.13319 (2019).

* The prompt is from MathQA dataset