Editing Large Language Models

Gokul Sunilkumar, Pooja Srinivasan, Sai Rithvik Kanakamedala, Saili Myana, Utsav Vachhani
Industry Mentors: Akshat Gupta, Simerjot Kaur
Faculty Mentor: Vivian Zhang

Data Science Capstone Project with JP Morgan Chase & Co.

Introduction

Large Language Models (LLMs) represent a substantial reservoir of information and generative capabilities. To enhance their factual accuracy, various editing methods have emerged, offering computationally efficient ways to update information. Our focus lies in analyzing the storage paradigm within autoregressive transformer models, utilizing causal tracing to quantify module contributions and understand their causal impact on predictions. Through a series of experiments, we delve into the localization of factual knowledge and the activation behaviors of these models.

Subject vs. Relation corruption

Causal tracing[1], our chosen algorithm for identifying knowledge storage layers, involves running the network multiple times with subject token corruptions. By restoring individual states, we pinpoint information crucial for restoring results. Extending this method, we introduce relation token corruptions, shedding light on additional MLP layers where specific facts may be stored.

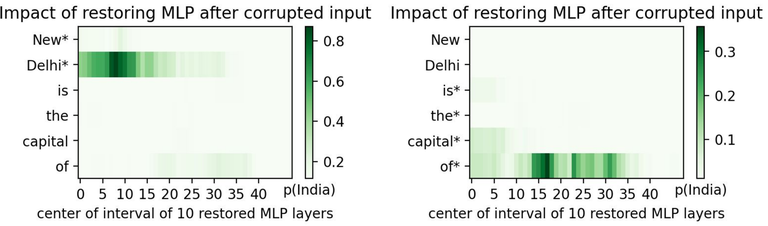

Figure 1. Causal tracing when corrupting subject and relation resp.

Knowledge Centers

In this experiment, we test the hypothesis that facts with different relations are stored by the model in different layers of the network. With a dataset featuring over 170 prompts for capital and sports categories, we meticulously analyze variations in activations, providing insights into the distribution of knowledge across layers.

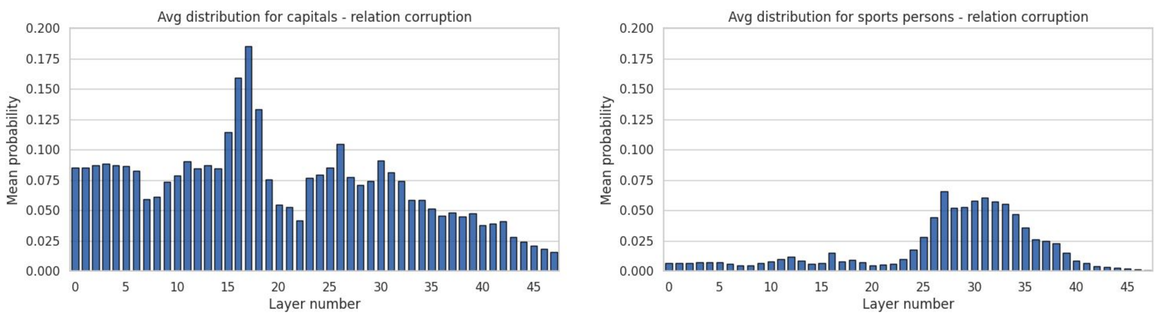

Figure 2. Probability distribution over layers for Capital prompts and sports prompts resp.

References

[1] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. "Locating and Editing Factual Associations in GPT." Advances in Neural Information Processing Systems 36 (2022).

Popular vs. Rare

In this experiment, we investigate variations in activations and layers, unveiling differences between predicting well-known facts and less common or rare information to enhance editing capabilities.

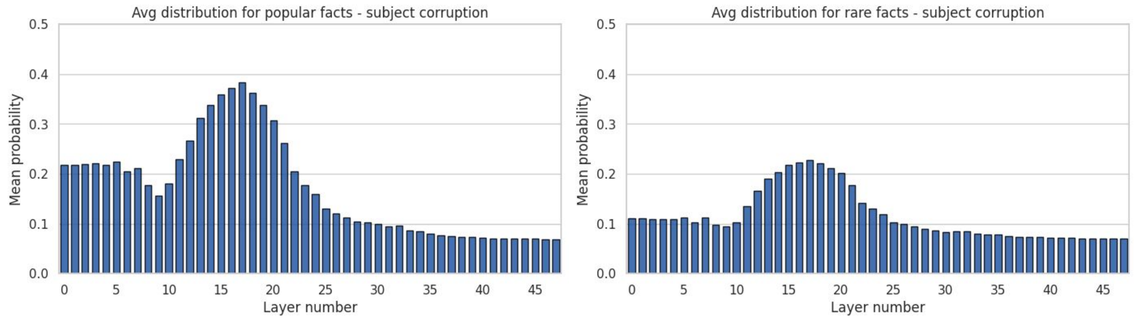

Figure 3. Probability distribution over layers for popular and rare prompts resp.

Correct vs Incorrect predictions

In this experiment, we discern the nuanced variations in activations during instances where the model accurately predicts the fact and where it makes incorrect predictions, shedding light on the model's performance dynamics.

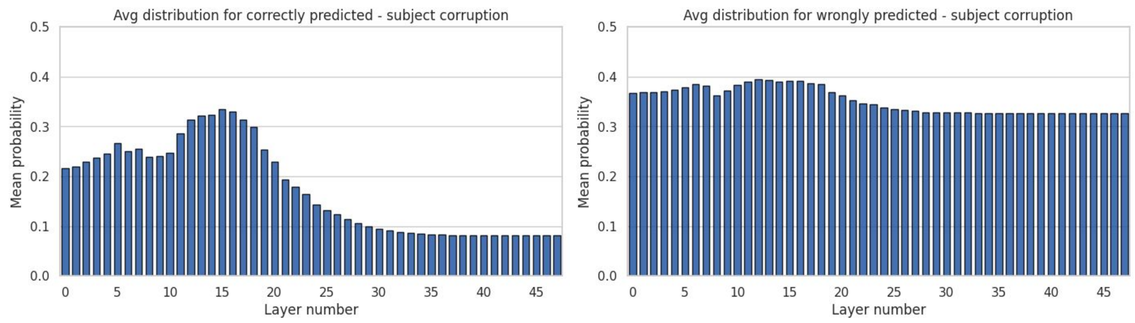

Figure 4. Probability distribution for correctly and incorrectly predicted prompts resp.

Conclusion

In conclusion, our work advances our understanding of knowledge storage within transformers as well as underscores the importance of considering the properties of a fact and its input tokens.

- We highlight the limitation of current editing approaches that focus on specific layers for all prompts, demonstrating that different types of prompts may be stored in distinct layers.
- Analyzing incorrectly predicted prompts also opens avenues for deeper investigations into the model's behavior and potential clues for improving accuracy based on activation patterns.

Acknowledgment: We express our sincere gratitude to Akshat Gupta and Simerjot Kaur from the JP Morgan AI Research Team for their invaluable guidance, which played a pivotal role in shaping the direction of our research endeavors.