Visualizing Token Importance in Large Language Models

Guided Research Project Description

Ruben Kaiser

1 Introduction and Background

Large Language Models (LLMs) have revolutionized natural language processing, demonstrating remarkable capabilities in various tasks. However, their inner workings often remain opaque, presenting challenges for understanding, interpretation, and inhibiting trust. This guided research aims to develop and apply visualization techniques to elucidate token-level attention within LLMs, offering valuable insights into their decision-making processes when generating outputs.

Recent advancements (e.g.[1, 2, 3]) in model interpretability have shown promise in unraveling the complexities of LLMs. Despite these advancements, there is still a need for more accessible and visually intuitive methods to illustrate token attention in contemporary language models.

To bridge this gap, this research proposes a comprehensive visualization framework for token importance, which then can be leveraged to assess existing prompting strategies and illuminate the inner workings of language models.

2 Research Objectives

The primary objectives of this research are:

1. To develop a comprehensive visualization tool that accurately and intuitively represents token importance in LLMs across various tasks and contexts.
2. To analyze and compare token importance patterns across different types of language tasks, models, and input structures.
3. To investigate the relationship between token importance and output quality, relevance, and adherence to task instructions.

3 Methodology

3.1 Tool Development

An initial proof-of-concept tool has been developed in JupyterLab that generates heatmap visualizations of token importance in LLMs. A significant focus of the implementation part of this research will be on enhancing the robustness and scalability of this tool to accommodate a wider range of models and tasks as well as more intuitive visualisations, especially for larger outputs.

3.2 Visualization and Analysis

Above mentioned tool generates token importance heatmaps for various prompts, allowing for in-depth analysis of the resulting patterns. Example visualizations are provided in the Appendix (Figures 1 and 2), demonstrating the tool's capabilities for different types of prompts.

Key insights from these visualizations include:

- The model's adherence to specific instructions (e.g., "one word" or "six-word story").
- Varying importance of input tokens throughout the generation process.
- The model's focus on thematic elements (e.g., "sky" for color, "love" for the story).
- Shifts in token importance as the generation progresses.

3.3 Comparative Studies

The research will conduct comparative analyses to:

- Examine token importance patterns across different task types.
- Investigate the impact of prompt structure on token importance and model output.
- Compare token importance patterns across different model architectures and sizes.
- explore the relationship of output relevance for different prompting structures

4 Expected Outcomes and Impact

This research is expected to yield several outcomes:

1. A robust, open-source tool for visualizing token importance in LLMs.
2. New insights into the decision-making processes of LLMs, potentially revealing biases, limitations, and areas for improvement.
3. A deeper understanding of how different model architectures and sizes process and prioritize information.

The impact of this research extends beyond academic circles. By providing a window into the inner workings of LLMs, this work has the potential to foster greater trust and understanding of these powerful tools. It might influence the development of more interpretable and reliable AI systems, contributing to the broader goal of responsible AI development.

5 Conclusion

This guided research on visualizing token importance in LLMs represents a significant step towards demystifying the inner workings of these complex systems. By leveraging innovative visualization techniques and rigorous analysis, it aims to provide valuable insights that can drive advancements in model interpretability and AI research. The potential applications of this work span across academia, industry, and public understanding of AI, making it a timely and impactful contribution to the field of natural language processing and artificial intelligence.

References

[1] Chandan Singh et al. "Rethinking Interpretability in the Era of Large Language Models". In: *ArXiv* abs/2402.01761 (2024). URL: https://api.semanticscholar.org/CorpusID:267412530.

[2] Haoyan Luo and Lucia Specia. "From Understanding to Utilization: A Survey on Explainability for Large Language Models". In: *ArXiv* abs/2401.12874 (2024). URL: https://api.semanticscholar.org/CorpusID:267095032.

[3] Haiyan Zhao et al. "Explainability for Large Language Models: A Survey". In: *ACM Transactions on Intelligent Systems and Technology* 15 (2023), pp. 1 –38. URL: https://api.semanticscholar.org/CorpusID:261530292.

Appendix: Visualization Examples

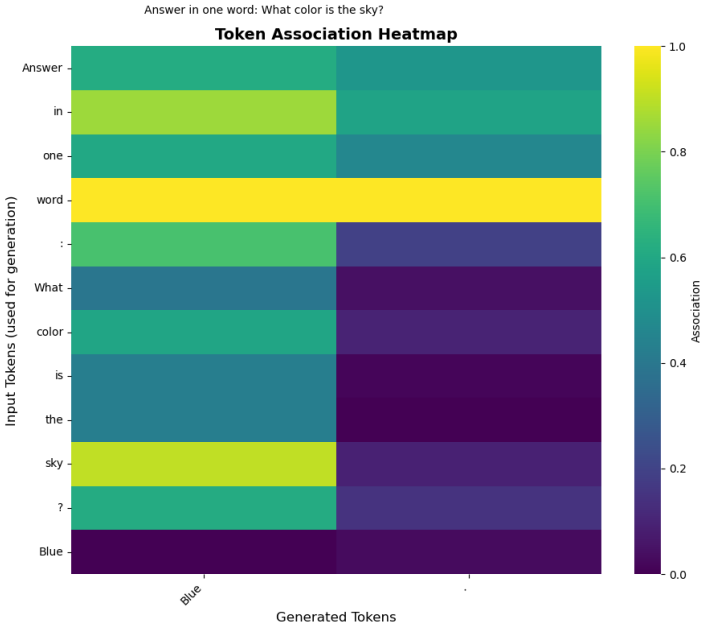

Figure 1: Token Association Heatmap for "Answer in one word: What color is the sky?"

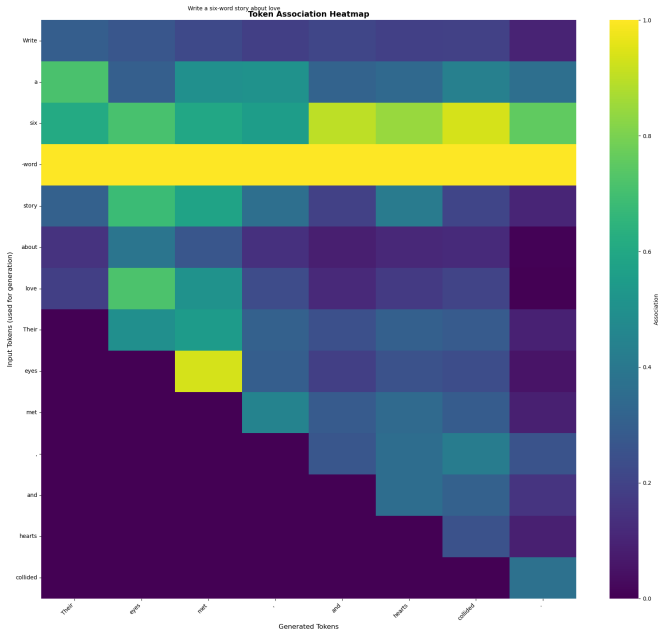

Figure 2: Token Association Heatmap for "Write a six-word story about love"