Let's Decrypt Dot by Dot: Decoding Hidden Computation in Transformer Language Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recent work has shown that transformer models can perform complex reasoning tasks using Chain-of-Thought (COT) prompting, even when the COT is replaced with hidden characters. This paper investigates methods to decode these hidden computations, focusing on the 3SUM task. We analyze a 34M parameter LLaMA model trained on hidden COT sequences and propose a novel decoding method that successfully recovers the original COT. Our findings provide insights into how transformers encode and process information in hidden COT sequences, offering new perspectives on model interpretability and the nature of computation in language models.

1 Introduction

Chain-of-Thought (COT) prompting has emerged as a powerful technique for improving the performance of large language models on complex reasoning tasks [1]. However, recent work by [2] demonstrates that these improvements can be achieved even when the COT is replaced with hidden characters (e.g., "..."), raising intriguing questions about the nature of computation being performed within these models.

This paper builds upon the findings of [2], focusing on the 3SUM task as a case study. We aim to decode the hidden computations embedded within the transformer architecture when trained on hidden COT sequences. Our work provides valuable insights into how these models encode and process information, potentially leading to improved model interpretability and more effective training strategies.

2 Background

2.1 The 3SUM Task

The 3SUM task involves finding three numbers in a given set that sum to zero. In this experiment, it serves as a proxy for more complex reasoning tasks to study the computational capabilities of transformer models [1]. For this experiment, we have used the same data representation method and generating process for 3SUM sequences as in [1].

2.2 Hidden Chain-of-Thought

In hidden Chain-of-Thought, intermediate reasoning steps are replaced with hidden characters (e.g., "..."). It has been observed in [2] that the models trained on hidden sequences still perform well, suggesting that meaningful computation occurs despite the lack of explicit reasoning steps.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

3 Methodology

We used a 34M-parameter LLaMA model with 4 layers, 384 hidden dimension, and 6 attention heads [3], the size of the training and test datasets and all other hyperparameters are kept the same as in the "Let's think dot by dot" paper [2]. Our analysis focused on three main areas: Layer-wise Representation Analysis, Token Ranking, and Modified Greedy Decoding Algorithm.

4 Results and Discussion

4.1 Layer-wise Analysis

Our analysis revealed a gradual evolution of representations across the model's layers. The initial layers primarily contained raw numerical sequences associated with the 3SUM problem's chain of thought. However, starting from the third layer, we noticed the emergence of hidden tokens. As we progressed through subsequent layers, we observed a steady transition from purely numerical sequences to an increasing prevalence of hidden characters.

This pattern suggests that the model develops the ability to utilize hidden tokens as proxies only in its deeper layers, which aligns with intuitive expectations of how neural networks process information. After conducting a comprehensive evaluation across numerous examples, we found that, on average, there is a marked increase in hidden token usage immediately following the second layer. Furthermore, the final layers of the model demonstrate extensive reliance on these hidden tokens.

For this analysis, we employed nostalgebraist's logit lens method [4].

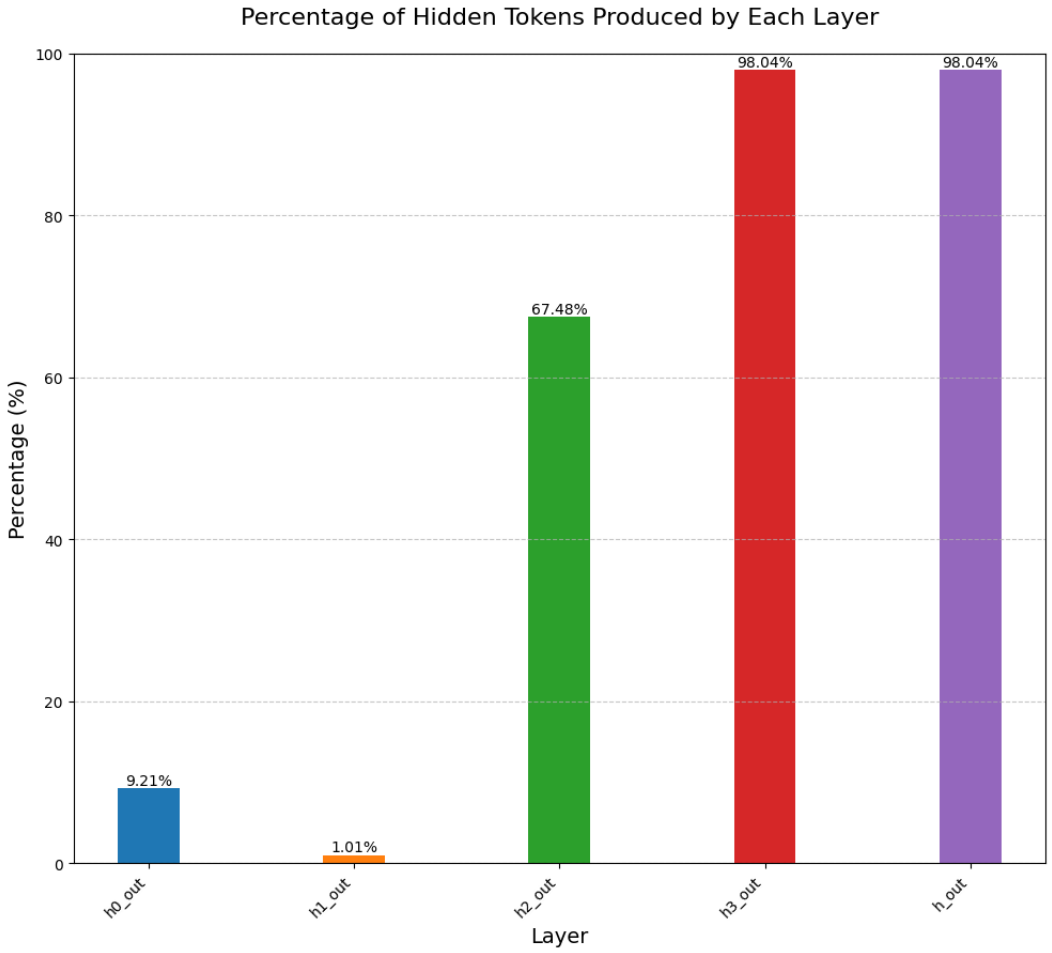

Figure 1: Hidden token occurrence percentages in generated sequences across layers

4.2 Token Rank Analysis

The top-ranked token was consistently the hidden character ("."), while lower-ranked tokens revealed the original, non-hidden COT sequences. This supports the hypothesis that the model replaces all computation with hidden tokens on top while keeping the original computation intact underneath. We have provided a sample snapshot of sequences decoded at each layer for both the top (rank-1) and rank-2 tokens in the appendix section.

4.3 Modified Greedy Decoding Algorithm

Based on the observations made in the token rank analysis, we implemented a modified greedy autoregressive decoding method. The steps include: performing standard greedy decoding, selecting the second-highest probability token when encountering a hidden token, and continuing this process for the entire sequence. This resulted in a 100% match in 3SUM task results with and without hidden tokens. To test the effectiveness of this method, we have also compared this with replacing the hidden token with randomly sampled token instead of the next highest (rank-2) token. The percentages of the original and these modified decoding methods are visualized as plots below.

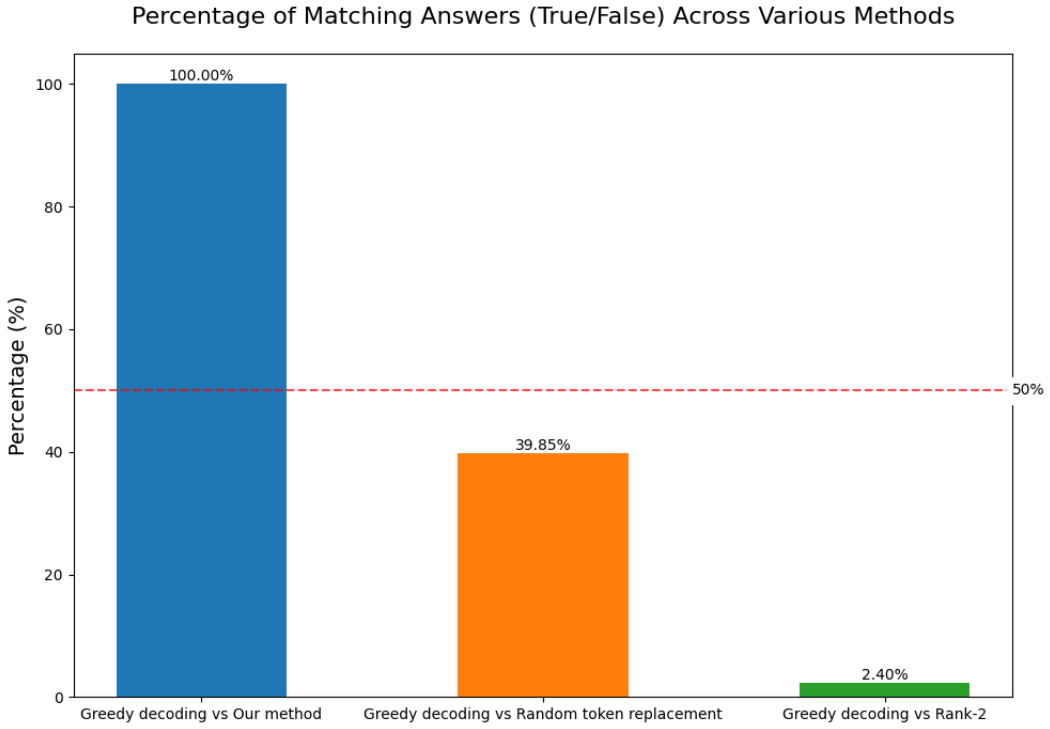

Figure 2: Comparison of decoding methods

5 Implications and Future Work

Our findings provide new tools for understanding internal reasoning processes and increase confidence in COT-based approaches for improving visibility. Future work should focus on developing better decoding methods or finding circuits that hide tokens, investigating generalizability to tasks beyond 3SUM (including natural language tasks), and improving token hiding methods (currently limited to one hidden token which is simple to decode).

6 Conclusion

We have presented a novel approach to understanding hidden computations in transformer models through the analysis of token rankings and layer-wise representations, and the development of a

modified decoding algorithm. Our insights into how models encode and process information in hidden COT sequences open new avenues for improving interpretability, efficiency, and safety in language models. This work strengthens the belief in COT visibility research for interpretability.

Supplemental Materials

The code used for the experiments and analysis in this paper is available on GitHub at `https://github.com/rokosbasilisk/filler_tokens/tree/v2`.

Appendix: Layerwise view of sequences generated via various decoding methods

```
h0_out: . [EOS] A 9 3 A [EOS] A A 4 A [EOS] 0 6 2 1 7 4 A A 1 1 3 3 3 A 4 6 6 6 9 3 4 4 4 4 [EOS] [EOS] [EOS] [EO
S] 3 . . 6 [EOS] [EOS] 3 9 9 [EOS] A A A 0-2 0 [EOS] [EOS] 4 4 [EOS] [EOS] [EOS] [EOS] 7 7 4 0 [EOS] [EOS] [EOS]
[EOS] 5 4 A 6 [EOS] [EOS] [EOS] [EOS] 0 1 [EOS] [EOS] [EOS] [EOS] A 4 [EOS] [EOS] [EOS] [EOS] [EOS] [EOS] [EOS] [E
OS] [EOS] [EOS] 0 0 [EOS] [EOS] [EOS]
h1_out: . 8 A 5 8 8 4 0 3 3 2 2 A A 3 3 2 2 A A 2 3 2 2 2 A A A A A 0 2 2 A 2 A 6 6 2 2 A 5 5 A A 3 A 0 A 9 A A A
A 7 3 8 2 5 A A A A 6 4 4 4 A A A A A A A A 4 4 A A 9 9 A 2 2 2 A A A 5 5 A A [EOS] [EOS] [EOS] [EOS] A A A A A [EO
S] False
h2_out: . . . . 0-7 True 4 . . . [EOS] . A . . . . . . . . . . . . A A 6 3 A . . . . . . . . . . A . . . . . A A T
rue . . . 0-9 0-9 . . . . . . . . A . [EOS] . . A A A . . A A [EOS] [EOS] . 6 . . . . . . A A . . . A [EOS] [EOS]
[EOS] [EOS] [EOS] A . . . [EOS] [EOS] False
h3_out: . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . A True
h_out: . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . A True
```

Figure 3: Greedy Decoding

```
h0_out: 0-0 2 [EOS] 0-2 A 5 A 2 A 0-8 A 3 [EOS] 9 0-2 4 4 5 0-6 5 0-2 5 3 2 0-8 2 A 9 1 9 0-5 5 0-8 6 0-7 3 0-9 0-
8 2 0-6 0-6 9 [EOS] 9 0-2 0 0-5 0 [EOS] [EOS] [EOS] 2 0-6 2 0-7 5 0-6 0 A 5 [EOS] 1 0-6 2 0-4 5 0-7 . [EOS] 5 A 6
A 7 0-6 6 0 6 0-6 5 0-6 2 0-6 5 0-6 5 0-7 1 [EOS] 4 0-6 0-2 0-6
h1_out: 0-0 9 0-2 9 0-2 1 0-2 1 0-2 6 0-2 1 0-2 5 0-2 8 0-2 7 0-2 4 0-2 6 0-2 9 0-2 4 0-6 8 0-2 2 0-9 6 0-9 0 0-7
4 0-7 6 0-7 2 0-6 6 0-2 1 0-6 7 0-2 8 0-7 8 0-7 4 0-6 3 0-9 3 0-7 9 0-7 0 0-7 8 0-6 0 0-9 3 0-7 8 0-9 5 0-7 0 0-9
2 0-6 4 0-7 6 0-7 3 0-6 1 0-7 9 0-7 9 0-6 8 0-9 7 0-7 7 0-7
h2_out: 0-1 9 0-2 9 0-3 1 0-4 2 0-5 6 0-6 0 0-7 9 0-8 8 0-0 7 0-2 7 0-3 6 0-4 2 0-5 4 0-1 8 0-7 2 0-8 5 0-9 1 0-3
6 0-2 6 0-5 1 [EOS] 5 0-7 1 0-2 7 0-2 8 A 8 0-5 2 0-3 7 0-7 3 0-3 9 0-3 4 0-4 0 0-4 2 0-7 3 [EOS] 6 0-4 0 0-5 2 0-
7 2 0-8 5 0-5 5 0-7 3 0-8 1 0-6 0 [EOS] 7 0-7 5 0-8 7 0-3 4 0-4
h3_out: 0-0 9 0-0 9 0-0 1 0-4 2 0-0 1 0-0 1 0-0 5 0-8 8 0-0 7 0-1 0 0-1 6 0-1 0-5 0-5 4 0-1 2 0-1 2 0-1 3 0-9 1 0-
2 6 0-4 6 0-2 2 0-2 1 0-2 1 0-2 7 0-2 8 0-3 8 0-5 2 0-6 8 0-7 3 0-3 9 0-9 0 0-4 8 0-6 0 0-7 8 0-8 7 0-9 5 0-5 0 0-
7 2 0-8 4 0-9 0 0-6 3 0-8 6 0-9 1 0-8 9 0-9 4 0-9 7 0-3 7 [EOS]
h_out: 0-0 9 0-0 9 0-0 1 0-4 2 0-0 1 0-0 1 0-0 5 0-8 8 0-0 7 0-1 0 0-1 6 0-1 0-5 0-5 4 0-1 2 0-1 2 0-1 3 0-9 1 0-2
6 0-4 6 0-2 2 0-2 1 0-2 1 0-2 7 0-2 8 0-3 8 0-5 2 0-6 8 0-7 3 0-3 9 0-9 0 0-4 8 0-6 0 0-7 8 0-8 7 0-9 5 0-5 0 0-7
2 0-8 4 0-9 0 0-6 3 0-8 6 0-9 1 0-8 9 0-9 4 0-9 7 0-3 7 [EOS]
```

Figure 4: Greedy Decoding with Rank-2 Tokens

```
h0_out: 0-0 5 [EOS] [EOS] [EOS] [EOS] [EOS] 6 A [EOS] [EOS] 0 A [EOS] 0-9 5 [EOS] [EOS] A 3 0-7 3 0-8 2 [EOS] 0-8
A 7 True True 0-6 [EOS] 0-6 0-8 0-6 2 [EOS] [EOS] [EOS] 2 0-6 9 0-6 5 0-6 0-8 0-6 2 A [EOS] [EOS] 2 0-6 6 A 6 [EO
S] [EOS] [EOS] [EOS] [EOS] 5 [EOS] 0 [EOS] 2 [EOS] True [EOS] [EOS] A 0-8 A 5 [EOS] [EOS] [EOS] [EOS] A 5 A 1 A [E
OS] A 6 0-7 [EOS] [EOS] [EOS] [EOS] [EOS] [EOS] [EOS]
h1_out: 0-0 2 0-0 3 [EOS] 5 0-0 2 0-0 7 0-0 0 0-0 9 0-0 5 0-0 3 0-0 0 0-7 2 0-7 4 [EOS] 8 0-2 8 0-2 0 0-6 5 0-8 8
0-2 0 0-2 1 0-2 0 0-6 1 0-7 4 0-6 0 0-7 2 0-6 3 0-6 2 0-7 8 0-7 6 0-6 2 0-6 4 0-6 9 0-7 2 0-7 2 0-9 6 0-7 0 0-7 9
0-7 8 0-6 5 0-9 5 0-6 2 0-6 8 0-6 0 0-6 6 0-7 5 0-7 4 0-9 4 0-6 False
h2_out: 0-1 1 0-0 3 0-0 5 0-0 1 0-0 1 0-0 1 0-0 5 0-0 5 0-9 3 0-1 0 0-1 2 0-1 0-9 0-1 8 0-6 9 0-1 0 0-1 3 0-9 8 0-
2 0 0-4 1 0-2 1 0-6 1 0-2 4 0-8 0 0-9 2 0-4 3 0-3 3 0-6 8 0-3 6 0-8 2 0-9 0 A 8 0-6 0 0-4 2 0-4 8 0-9 1 0-6 0 0-5
8 0-5 4 0-9 0 0-6 2 0-6 8 0-9 1 A 6 0-9 8 0-9 4 0-9 4 A A
h3_out: 0-0 1 0-2 3 0-3 5 0-0 6 0-5 6 0-6 2 0-7 9 0-0 5 0-9 3 0-2 7 0-1 9 0-4 0-9 0-1 8 0-6 8 0-7 4 0-8 6 0-1 8 0-
3 0 0-2 1 0-5 1 0-6 5 0-7 4 0-8 0 0-9 2 0-4 3 0-3 3 0-3 7 0-3 6 0-8 2 0-9 4 0-5 0 0-6 2 0-4 2 0-4 7 0-4 0 0-5 9 0-
5 8 0-8 5 0-5 5 0-7 2 0-6 6 0-6 9 0-7 6 0-7 8 0-8 4 0-9 4 A True
h_out: 0-0 1 0-2 3 0-3 5 0-0 6 0-5 6 0-6 2 0-7 9 0-0 5 0-9 3 0-2 7 0-1 9 0-4 0-9 0-1 8 0-6 8 0-7 4 0-8 6 0-1 8 0-3
0 0-2 1 0-5 1 0-6 5 0-7 4 0-8 0 0-9 2 0-4 3 0-3 3 0-3 7 0-3 6 0-8 2 0-9 4 0-5 0 0-6 2 0-4 2 0-4 7 0-4 0 0-5 9 0-5
8 0-8 5 0-5 5 0-7 2 0-6 6 0-6 9 0-7 6 0-7 8 0-8 4 0-9 4 A True
```

Figure 5: Our Method: Greedy Decoding with Hidden Tokens Replaced by Rank-2 Tokens

```
h0_out: 0-0 5 [EOS] [EOS] A 5 [EOS] 2 A [EOS] A 3 A 6 0-9 4 4 7 A 5 0-2 2 A 5 True 2 A 5 1 5 0-5 [EOS] 0-8 6 0-6 0
-5 0-9 0-8 [EOS] 2 0-6 5 0-6 2 0-6 0-8 0-2 [EOS] [EOS] A [EOS] 2 A 2 A 6 [EOS] 5 A [EOS] [EOS] 5 [EOS] 0 [EOS] 2
[EOS] 0-8 0-7 5 0-6 0-8 0-2 5 [EOS] 6 0 [EOS] A 5 A 5 A 5 A 5 0-7 1 [EOS] True [EOS] 4 0-6
h1_out: 0-0 2 0-2 9 0-0 1 0-0 2 0-2 7 0-2 0 0-0 5 0-0 5 0-0 3 0-2 0 0-7 6 0-2 9 0-2 4 0-2 9 0-2 0 0-2 6 0-9 0 0-2
0 0-2 1 0-2 2 0-6 1 0-2 1 0-6 0 0-6 2 0-6 3 0-6 4 0-6 3 0-7 3 0-6 2 0-6 4 0-7 9 0-7 0 0-9 3 0-7 6 0-9 5 0-7 9 0-9
2 0-6 5 0-9 6 0-6 2 0-6 1 0-7 0 0-7 6 0-9 8 0-6 7 0-7 7 0-6
h2_out: 0-1 1 0-2 9 0-3 1 0-0 2 0-0 1 0-0 1 0-0 5 0-8 8 0-0 3 0-2 0 0-3 2 0-1 0-9 0-5 8 0-6 9 0-1 0 0-1 5 0-1 8 0-
3 6 0-4 6 0-2 1 [EOS] 5 0-7 1 0-8 7 0-9 8 A 3 0-3 2 0-6 8 0-3 3 0-8 2 0-3 0 0-4 8 0-6 0 0-4 3 A 7 0-9 0 0-5 2 0-7
2 0-5 4 0-5 6 0-6 3 0-6 1 0-6 1 A 7 0-7 5 0-8 4 0-9 4 0-4
h3_out: 0-0 9 0-0 9 0-0 5 0-4 2 0-0 7 0-6 2 0-7 5 0-8 8 0-9 7 0-2 7 0-3 9 0-1 0-9 0-5 4 0-1 8 0-1 4 0-1 6 0-1 1 0-
2 0 0-2 1 0-5 0 0-6 5 0-2 4 0-8 0 0-9 8 0-3 8 0-3 2 0-3 8 0-7 3 0-8 9 0-3 4 0-5 8 0-4 0 0-4 8 0-8 7 0-4 0 0-6 9 0-
7 8 0-8 4 0-5 0 0-7 3 0-6 6 0-9 0 0-8 6 0-7 8 0-8 7 0-0 7 A
h_out: 0-0 1 0-2 9 0-0 1 0-4 2 0-5 7 0-0 1 0-7 5 0-8 5 0-9 7 0-2 0 0-3 9 0-1 0-9 0-1 4 0-1 8 0-1 4 0-8 3 0-9 8 0-3
6 0-2 6 0-5 1 0-2 1 0-7 1 0-2 7 0-9 8 0-3 8 0-5 2 0-6 8 0-7 6 0-8 2 0-9 4 0-5 0 0-6 2 0-7 2 0-4 7 0-9 0 0-5 2 0-5
8 0-8 4 0-5 0 0-7 3 0-8 1 0-6 0 0-8 9 0-9 8 0-9 7 0-9 7 [EOS]
```

Figure 6: Greedy Decoding with Hidden Tokens Replaced by Randomly Selected Tokens

References

[1] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Zhao, Sifei Zhu, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022. URL https://arxiv.org/abs/2201.11903.

[2] Jacob Pfau, William Merrill, and Samuel R Bowman. Let's think dot by dot: Hidden computation in transformer language models. *arXiv preprint arXiv:2404.15758*, 2023. URL https://arxiv.org/abs/2404.15758.

[3] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. URL https://arxiv.org/abs/2302.13971.

[4] nostalgebraist. Interpreting GPT: the logit lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/, 2020. Accessed: 2024-04-25.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the main contributions of the paper, including the analysis of a 34M parameter LLaMA model, the proposal of a novel decoding method, and insights into how transformers process hidden COT sequences.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations in the "Implications and Future Work" section, mentioning the current limitation to one hidden token and the need for further investigation into generalizability beyond the 3SUM task.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This paper focuses on empirical analysis and does not present formal theoretical results or proofs.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides details on the model architecture, dataset, and methodology used.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

Justification: While the paper describes the methodology in detail, it does not provide direct access to the code or data used in the experiments.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper provides details on the model architecture, including the number of layers, hidden dimension, and attention heads. It also mentions that other hyperparameters are kept the same as in the referenced "Let's think dot by dot" paper.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: While the paper presents quantitative results, it does not explicitly report error bars or statistical significance tests for the experiments.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The paper does not provide detailed information about the specific compute resources used for the experiments.

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research presented in this paper does not appear to violate any ethical guidelines. It focuses on model interpretability and does not involve sensitive data or potentially harmful applications.

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: While the paper discusses implications and future work, it does not explicitly address potential positive and negative societal impacts of the research.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper does not involve the release of models or datasets that pose a high risk for misuse.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

Justification: The paper cites the original sources for the LLaMA model and the 3SUM task methodology. However, specific license information for these assets is not mentioned in the paper.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduces a novel decoding method, which is described in detail in the methodology section.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This research does not involve crowdsourcing or human subjects.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This research does not involve human subjects, so IRB approval was not required.