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

Aryasomayajula Ram Bharadwaj
Independent Researcher
ram.bharadwaj.arya@gmail.com

August 23, 2024

Abstract

Transformer models can perform complex reasoning with Chain-of-Thought (COT) prompting. COT can be replaced with hidden characters while maintaining performance. This paper investigates methods to decode these hidden computations, focusing on the 3SUM task using a 34M parameter LLaMA model. We propose a novel decoding method to recover the original COT, providing insights into how transformers encode and process hidden COT information. Our work offers new perspectives on model interpretability and computation in language models.

1 Introduction

COT prompting improves the performance of large language models[?]. Recent work shows that these improvements persist when COT is replaced with hidden characters, raising questions about the nature of computation in these models[?]. This paper builds on findings by Pfau et al., focusing on the 3SUM task as a case study. We aim to decode hidden computations in the transformer architecture, with potential for improved model interpretability and training strategies.

2 Background

2.1 The 3SUM Task

The 3SUM task involves finding three numbers in a set that sum to zero. It serves as a proxy for more complex reasoning tasks and is used to study the computational capabilities of transformer models[?]. An example of a 3SUM sequence is provided in the methodology section.

2.2 Hidden Chain-of-Thought

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

3 Methodology

We used a 34M-parameter LLaMA model with 4 layers, 384 hidden dimension, and 6 attention heads[?]. 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

We observed a progressive transformation of representations across layers. Initial layers contained pure number sequences related to 3SUM's COT, with hidden tokens appearing from the third layer onward. We observed a gradual replacement of number sequences with hidden characters.

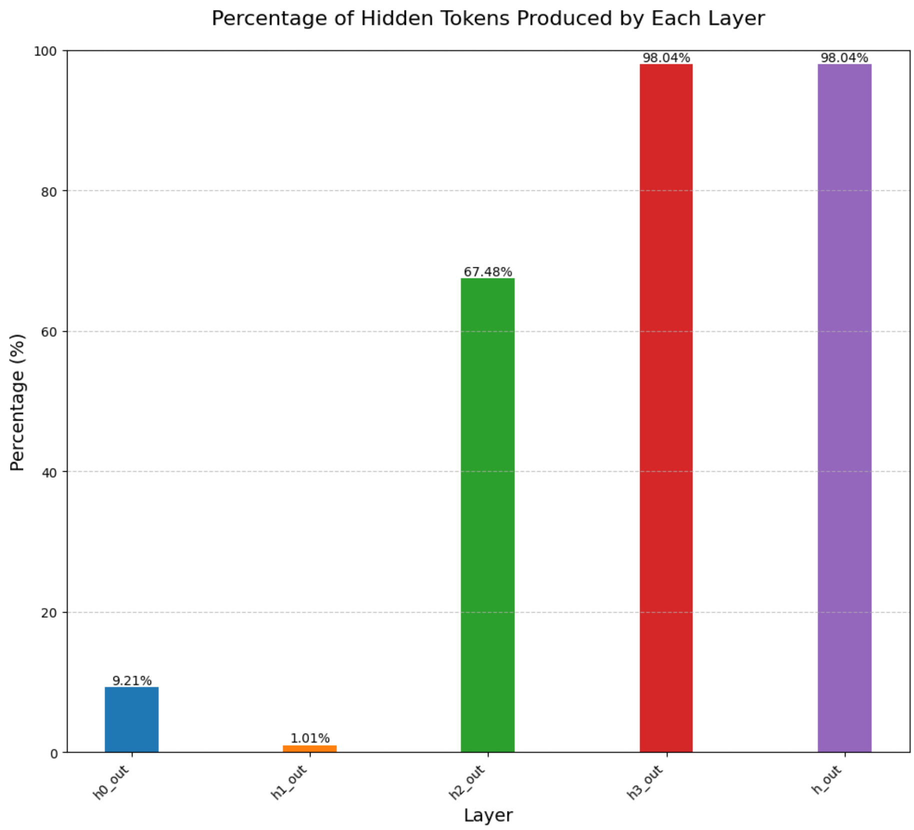

Figure 1: Hidden token percentages across layers

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.

4.3 Modified Greedy Decoding Algorithm

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. Random token replacement was less effective than choosing the next highest ranked token.

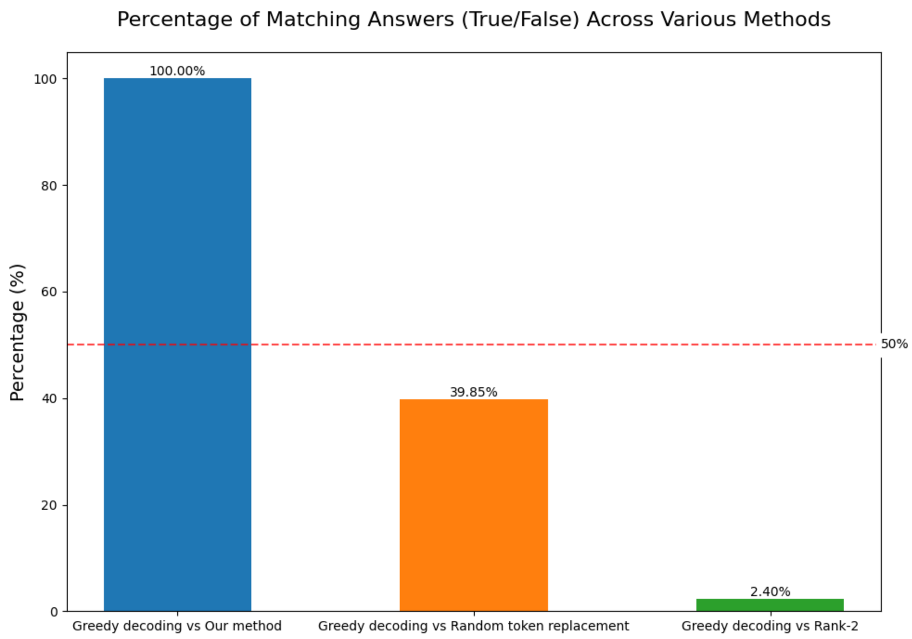

Figure 2: Comparison of decoding methods

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[**?**].

A Layerwise comparison of 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 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

B References

1. Pfau, J., Merrill, W., & Bowman, S. R. (2023). Let's Think Dot by Dot: Hidden Computation in Transformer Language Models. arXiv:2404.15758.

2. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. arXiv:2201.11903.

3. Sanford, C., Hsu, D. J., & Telgarsky, M. (2023). Representational strengths and limitations of transformers. arXiv:2306.02896.

4. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.

5. Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., ... & Hubinger, E. (2023). Measuring faithfulness in chain-of-thought reasoning. arXiv:2307.13702.