

LLMs Do Not Think Step-by-step In Implicit Reasoning

Yijiong Yu^a

^aTsinghua University

Abstract

It has been well-known that Chain-of-Thought can remarkably enhance LLMs’ performance on complex tasks. However, because it also introduces slower inference speeds and higher computational costs, many researches have attempted to use implicit CoT, which does not need LLMs to explicitly generate the intermediate steps. But there is still gap between their efficacy and typical explicit CoT methods. This leaves us a doubt that, does implicit CoT really equal to explicit CoT? Therefore, in this study, we address this question through experiments. We probe the information of intermediate steps from the model’s hidden states when it is performing implicit CoT. The results surprisingly indicate that LLMs hardly think about intermediate steps, suggesting they may just rely on experience rather than strict step-by-step reasoning. Moreover, we find LLMs’ implicit reasoning capabilities are susceptible and unstable, reaffirming the necessity of explicit CoT to effectively support complex tasks.

1 Introduction

Advancements in Large Language Models (LLMs) have unveiled unprecedented capabilities in handling complex reasoning tasks. Chain-of-Thought (CoT) prompting (Wei et al., 2022; Yu et al., 2023), in particular, has demonstrated substantial improvements in the reasoning abilities of LLMs by explicitly mapping out intermediate reasoning steps. Moreover, recent works of CoT training, such as OpenAI o1 (Qin et al., 2024) further demonstrate the power of CoT.

However, the CoT approach, despite its efficacy, it notably incurs slower inference speeds and higher computational costs. These drawbacks have spurred some researches on alternative reasoning methodologies that bypass the explicit generation of intermediate tokens, leveraging the model’s inherent “vertical” reasoning capabilities through its internal processing layers. For example, (Deng

et al., 2024) remove the intermediate steps and fine-tune the model to let model learn implicit CoT, and (Deng et al., 2023) train a emulator which emulate the intermediate states in CoT reasoning and train a student model to generate answers from these implicit states. This form of reasoning does not need to output intermediate results as tokens, called implicit reasoning or vertical reasoning, which contrasts with the “horizontal” reasoning, i.e. typical CoT. Figure 1 shows the difference between explicit CoT and implicit CoT. Although the concept of “implicit CoT (reasoning)” is rarely directly mentioned, in many scenarios that require low latency, users usually ask LLMs to output the final answer directly, which actually has forced LLMs to adopt the implicit reasoning way.

Despite the theoretical appeal of implicit reasoning as a more efficient alternative to traditional CoT methods, empirical evidence suggests the performance of implicit CoT still lag behind explicit CoT. Moreover, though some previous researches have confirmed the concept of implicit reasoning and attempted to analyze its process and efficacy (Yang et al.; Wang et al.; Allen-Zhu and Li), they usually more focus on using knowledge-based problems to examine whether LLMs can recall their parametric knowledge during implicit reasoning, instead of investigating more basic and generic forms of multi-step problems such as arithmetic. So far, there is still no clear and widely accepted conclusion on the rationale of implicit reasoning.

This situation makes us wonder fundamental questions about the nature of the implicit reasoning, such as “Are LLMs doing the same thing in the processes of implicit and explicit CoT?” and “Can the hidden, internal and layer-by-layer processing truly serve as an equivalent to explicit CoT reasoning?” To answer these questions, our study designs an elaborate set of experiments aimed at uncovering the implicit reasoning processes within a large model, specifically targeting the process of

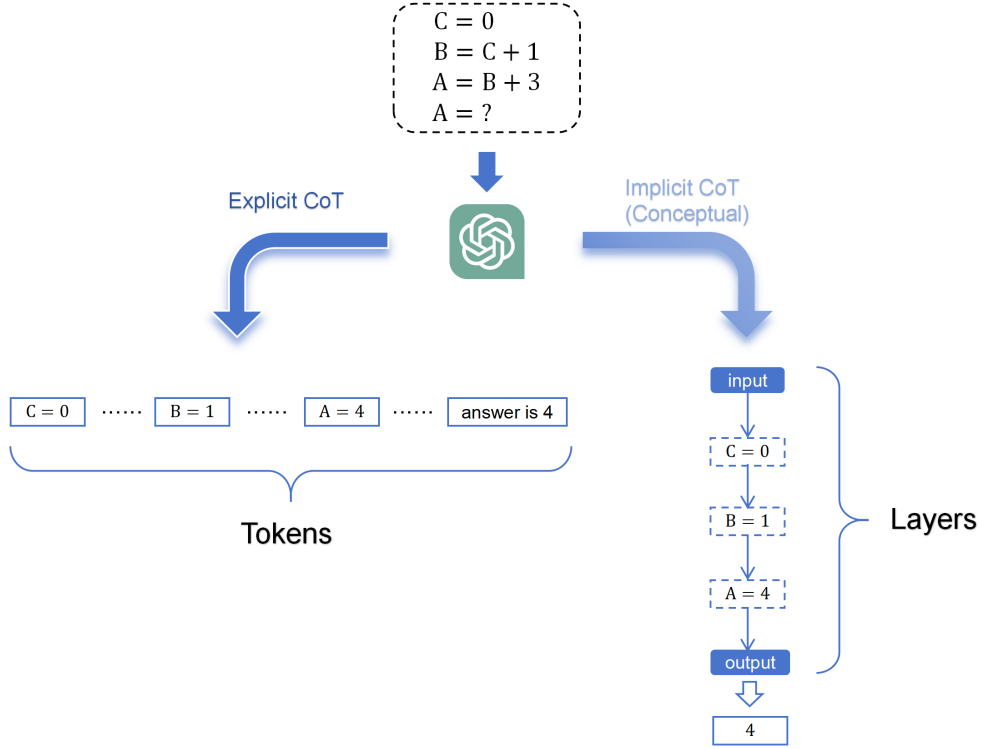


Figure 1: The examples of explicit CoT and implicit CoT. Explicit CoT is commonly used, which is completed by step-by-step output tokens. The process of implicit CoT is just a hypothetical or conceptual situation, which could be a layer-by-layer way.

handling multi-step arithmetic problems without resorting to outputting explicit intermediate steps.

In our experiment, we leverage a powerful open-source model, Qwen2.5-72B-Instruct (Team, 2024), with 80 layers, to tackle simple arithmetic problems that are easily solvable via typical CoT reasoning (Ye et al., 2024). However, we force the model to directly give the answer without outputting steps, so that we can examine whether these tasks can be addressed through implicit reasoning and how implicit reasoning happens. The arithmetic problems have a controllable number of reasoning steps, with each intermediate result being known. By investigating the hidden states associated with the final token of the given problem statement across layers and employing a simple linear classifier to probe those intermediate results, we aim to find out if the model really calculates the intermediate results in its implicit thinking process.

The experiment results are surprising and counter-intuitive: we find the model hardly calculates the intermediate results in implicit reasoning, despite it can often give the correct answer of the multi-step problem. Moreover, through slightly modifying the problem without even increasing its

difficulty, we find implicit reasoning is more unstable and susceptible. This finding suggests that in implicit reasoning, the model may not strictly follow a step-by-step reasoning process, but relies solely on an intuitive and direct way of thinking to complete the task, belonging to System 1 thinking (Kahneman, 2011), which is faster but less reliable.

In conclusion, we think, despite LLMs can often directly give the correct answer of a multi-step problem, especially those with very large sizes, they are not really doing step-by-step reasoning (at least in arithmetic problems), unless adopting explicit CoT. Implicit reasoning may just be an illusion created by LLMs’ powerful memory and rich experience, which is fundamentally different from conventional reasoning. Our study provides critical insights into the mechanics of implicit reasoning and emphasizes the ongoing necessity for explicit CoT methodologies in enhancing LLMs’ ability on complex tasks.

2 Approach

2.1 Experiment Design

To present the reasoning steps clearly, we adopt simple multi-step arithmetic problems with only

addition and subtraction. Usually, when given such problems, modern LLMs will automatically use a CoT manner to address them. To investigate the process of implicit reasoning, we use prompt to force the model to give the answer without using CoT. Therefore, an example of our prompt, which is a 5-step problem, is as follows:

$E = 8;$
 $D = E - 5;$
 $C = D + 2;$
 $B = C + 5;$
 $A = B - 1;$

Question: What is the value of A? You must answer directly with A=xxx.

Answer: A=

We randomly change the value in the problem to generate 2000 different samples, and each intermediate results are record. For example, the intermediate results of the above example should be $[8, 3, 5, 10, 9]$, i.e. the corresponding value of E, D, C, B and A. The result of the last step is the final answer.

The model will direct output the answer after our prompt, thus we take the last token of the prompt as our main research object and record its hidden states of each layers. Then, we adopt a typical linear probing method, which uses an 1-layer MLP, to predict each of the intermediate results from the hidden states. We control all of the intermediate values is within -10 to 10 so that the probe is a 21-class classifier (each value corresponds to one class).

We use 1600 samples to train the classifier for 10 epochs and 400 samples for testing its accuracy. And respectively for each hidden state of the 20 groups, we use it as the input feature to train an individual classifier. In training and testing, we set the result of each step as the label respectively to also train an individual classifier. Therefore, we finally get $20 * num_steps$ classifiers. If the classifier of the k -th hidden state shows high accuracy in the n -th step, it represents the model has calculated the result of the n -th step in the k -th hidden state.

We choose a large model, Qwen2.5-72B-Instruct (Team, 2024), to perform implicit reasoning, because we find small 7B level models can hardly do a multi-step problem correctly without CoT, while a 70B level model can achieve an accuracy of over 50%. Because the 72B model has 80 layers, to reduce the computing cost, we average the hidden

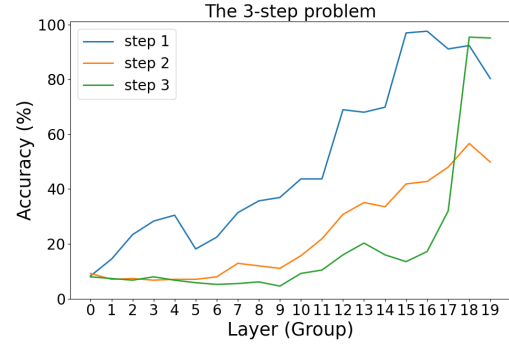


Figure 2: The accuracy of probing the result of each step in a 3-step problem.

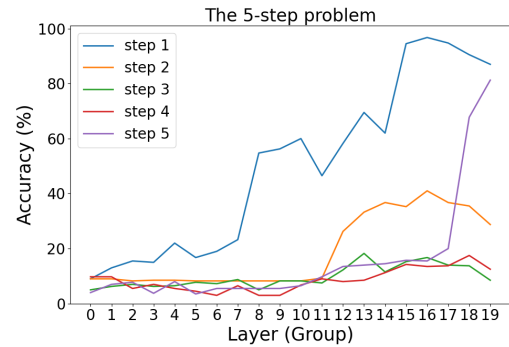


Figure 3: The accuracy of probing the result of each step in a 5-step problem.

states across every 4 consecutive layers. So, after the model processes our prompt, we get 20 groups of hidden state per sample, as well as its answer.

By default, in all generation experiments, the temperature is set to 0.

2.2 Results of Probing Intermediate Steps

The results in Figure 2 and Figure 3 shows the accuracy of probing the intermediate result of each step when the problem is 3-step or 5-step. It is clear that, in both situations, the results of the first step and the last step can always be probed successfully in the back layers, indicating the model does memorize the input value (i.e. the result of the first step) and does conceive the final answer (i.e. the result of the last step). And the result of the second step can also be detected to some extent, which suggests that LLMs may indeed have the ability to perform a 2-hop reasoning (the 3-step problem actually only needs 2 hops because the result of the first step is already given) in implicit reasoning. This phenomenon is consistent with the conclusion of previous research (Yang et al.).

However, by contrast, in the 5-step problem, the

results of other steps (3rd and 4th) can hardly be detected. It looks that the curve of the last step just surges in the last layers, even without waiting for the processing of the 3rd or 4th step. Therefore, the model may not calculate the results of 3rd and 4th steps at all.

This finding indicates that, in generic cases where the problem is more than 2-hop, there is actually not a specific state where the model calculates the results of the intermediate steps, even though it can give the correct answer of the multi-step problem. It actually skips the intermediate steps and come up with the final result directly. Therefore, we posit that perhaps due to a large model’s strong abstraction and memory abilities, it has learned a large number of answers to mathematical problems during the training stage. Therefore, it can almost directly map problems with multiple steps to their answers through intuition and experience, thus producing an “implicit reasoning” effect. But in fact, its mechanism is not equivalent to the explicit CoT process at all.

2.3 Result of Modifying the Problem Presentation

To further show the difference in mechanism between implicit reasoning and explicit reasoning, we slightly modify the problem by 2 methods: 1. reversing the order of the equations; 2. dividing all values by 10. Thus we obtain 3 types of problem presentations. The examples of the modified problem are as follows:

Reverse
$A = B - 1;$ $B = C + 5;$ $C = D + 2;$ $D = E - 5;$ $E = 8;$
Divide
$E = 0.8;$ $D = E - 0.5;$ $C = D + 0.2;$ $B = C + 0.5;$ $A = B - 0.1;$

For humans or LLMs performing CoT, such modifications hardly increase any difficulty of the problem, because the reasoning steps are not changed at all. However, for a model which only relies on intuition and experience, they will result in a significant difference between the form of the given

Prompt	Implicit		Explicit	
	3-step	5-step	3-step	5-step
original	85.01	53.95	100.00	100.00
reverse	70.62	13.71	100.00	100.00
divide	69.86	37.28	100.00	100.00

Table 1: The accuracy (%) of Qwen2.5-72b-instruct under different problem presentations using implicit or explicit reasoning on 3-step and 5-step problems.

problem and the form in experience. Through this comparison, we can be more convinced of the difference in mechanism between implicit and explicit reasoning.

We evaluate the model’s performance in under the 3 types of problem presentations while keeping the original values of the problems the same. We test with prompt styles of both the implicit reasoning way (as shown in section 2.1) and the explicit reasoning way (adding “let’s think step by step”).

From the results in Table 1, we can clearly see that, compare to the original problems, the modified problems significantly degrade the performance when using implicit reasoning. While the performance of explicit reasoning is always perfect. This contrast further demonstrate our inference that, in implicit reasoning, the model is actually answering directly by experience and intuition, but not by reasoning step-by-step. This cause the way of implicit reasoning less robust and less reliable.

3 Conclusion

In this study, we investigate the mechanism of LLMs doing implicit reasoning, and get a non-trivial finding that, unlike some previous studies which envisioned implicit reasoning as a substitute for explicit reasoning, implicit reasoning cannot be on par with explicit reasoning methods because it actually does not follow a step-by-step process but just intuitively thinks of the answer, which makes it less reliable. This finding remind us that there is no free lunch, that is, under current technological conditions, there may not be a perfect solution that can make LLMs output very few tokens while keeping the accuracy on solving complex problems. When you ask LLMs to give the answer directly, you should know that it has not actually undergone a real reasoning. Scaling the test-time by using explicit CoT may still be the most feasible method to further propel the capabilities of LLMs at present.

4 Limitations

We only test one type of reasoning problems. However, for more complex problems, there may be multiple situations of the intermediate steps and the final answer, making the experiment hard to design.

The fact that the intermediate results cannot be detected by linear probing cannot completely confirm they do not exist. Maybe they are encoded in another form, which is different from that of the results of the first and last step, and is more difficult to be detected. However, we have not yet found a better detection method.

References

- Zeyuan Allen-Zhu and Yuanzhi Li. [Physics of language models: Part 3.2, knowledge manipulation](#). *Preprint*, arxiv:2309.14402 [cs].
- Yuntian Deng, Yejin Choi, and Stuart Shieber. 2024. [From explicit CoT to implicit CoT: Learning to internalize CoT step by step](#). *Preprint*, arxiv:2405.14838 [cs].
- Yuntian Deng, Kiran Prasad, Roland Fernandez, Paul Smolensky, Vishrav Chaudhary, and Stuart Shieber. 2023. [Implicit chain of thought reasoning via knowledge distillation](#). *Preprint*, arxiv:2311.01460.
- Daniel Kahneman. 2011. Thinking, fast and slow. *Farrar, Straus and Giroux*.
- Yiwei Qin, Xuefeng Li, Haoyang Zou, Yixiu Liu, Shijie Xia, Zhen Huang, Yixin Ye, Weizhe Yuan, Hector Liu, Yuanzhi Li, and Pengfei Liu. 2024. [O1 Replication Journey: A Strategic Progress Report – Part 1](#). *arXiv preprint*. ArXiv:2410.18982.
- Qwen Team. 2024. [Qwen2.5: A party of foundation models!](#) Section: blog.
- Boshi Wang, Xiang Yue, Yu Su, and Huan Sun. [Grokking transformers are implicit reasoners: A mechanistic journey to the edge of generalization](#). *Preprint*, arxiv:2405.15071.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. [Do large language models latently perform multi-hop reasoning?](#) *Preprint*, arxiv:2402.16837.
- Tian Ye, Zicheng Xu, Yuanzhi Li, and Zeyuan Allen-Zhu. 2024. [Physics of Language Models: Part 2.1, Grade-School Math and the Hidden Reasoning Process](#).
- Zihan Yu, Liang He, Zhen Wu, Xinyu Dai, and Jiajun Chen. 2023. [Towards Better Chain-of-Thought Prompting Strategies: A Survey](#). *arXiv preprint*. ArXiv:2310.04959.