Knowledge Emergence in Chain-of-Thought Reasoning

Surya Appana [1] **Abhay Paidipalli** [1] **Dhruv Pendharkar** [1]

[1]University of California, Berkeley

suryacalnet@berkeley.edu, abhayp@berkeley.edu, dhruvpendharkar@berkeley.edu

Introduction

Chain-of-thought (CoT) prompting is the predominant inference-time technique to enable reasoning capabilities in large language models (LLMs). Instead of being prompted to solve a complex problem at once, an LLM is asked to break down the problem into a series of ”reasoning” steps, potentially coupled with one or more examples to guide its reasoning. This technique has been demonstrated to significantly improve performance on complex tasks [4], especially on math and logic problems [3].

Understanding why and how CoT prompting affects reasoning capabilities is critical to the deployment of LLMs in practice, especially in light of AI safety concerns. However, a precise mechanistic understanding of CoT-based reasoning is limited, with the majority of interpretability work for CoT reasoning focused on observational or correlational investigations.

Main Contributions

We focus on the following primary research questions:

1. **Answer emergence from CoT Prompting:** How does CoT prompting affect the emergence of answer tokens in intermediate layers/computations between non-CoT, zero-shot CoT, and few-shot CoT (1-shot and 2-shot) prompting?
2. **Reasoning Trace Causal Analysis:** How much are reasoning traces from CoT causally load-bearing for response generation?

We first show that the pattern of answer token emergence is heterogenous across prompt regimes and depends on the dataset and model size. Furthermore, we show that CoT surface tokens are causally load-bearing for difficult problems, but less so for easy ones. Crucially, the model is not simply reading off its internal state — the model is also, to some degree, conditioned on the tokens it has already produced, even when those tokens are wrong.

Methods

- **Datasets:** GSM8K and PrOntoQA
- **Models:** `Meta-Llama-3-8B-Instruct/Qwen2.5-14B-Instruct` (for answer emergence) and `Qwen-2.5-3B-Instruct/Qwen-2.5-7B-Instruct` (for reasoning trace analysis).
- **Logit lens:** applied to intermediate layers to observe the trend in answer token probability/rank across prompt regimes.
- **Injection procedure:**
  1. Generate a full baseline CoT using the model under standard instruction-following prompting (system prompt instructs step-by-step reasoning ending with ”The answer is [X]”).
  2. Segment the CoT into discrete reasoning steps (newline-separated for GSM8K).
  3. Select an injection point: early (after step n/4), middle (after step n/2), or late (after step n−2).
  4. Generate a wrong intermediate answer at that step and replace the step text accordingly.
  5. Reconstruct the prompt as: original prompt + CoT prefix up to injection + injected wrong step, then continue generation.
  6. Extract the final answer from the continuation and classify the outcome.
- **Linear probing:** Extract hidden states at the first generated token after injection for linear probe analysis.

Context-Based Knowledge Emergence

We first examine the pattern of answer emergence from CoT prompting across `Llama-3-8B` and `Qwen-2.5-14B` models, noting the probability rank of the answer token in intermediate layers. Notably, we observe the following:

- The trend in the correct answer token rank is heterogenous across prompt regimes, model sizes, and datasets, diverging at later model layers.
- In most cases, correct token rank is lower under CoT prompting, indicating a delayed answer emergence until the end of the model.

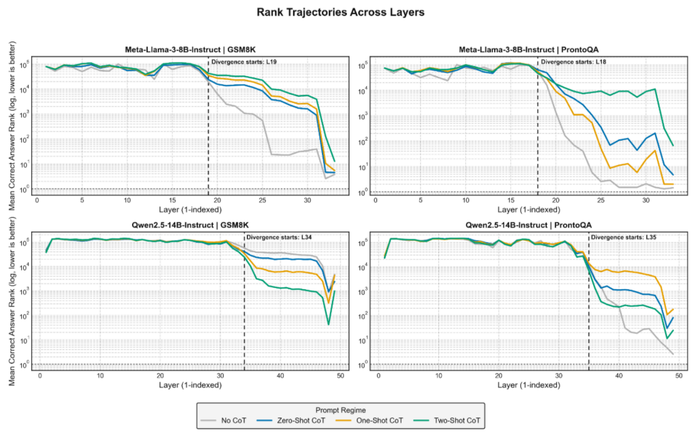

Figure 1:Correct answer token rank across model layers. We see a divergence in answer token emergence at later layers.

Reasoning Trace Causal Analysis

We further examine whether reasoning traces from CoT are causally load-bearing for response generation. We interrupt the model’s CoT at a controlled point, inject a wrong intermediate answer, and measure whether the model recovers the correct final answer, propagates the error, or collapses entirely.

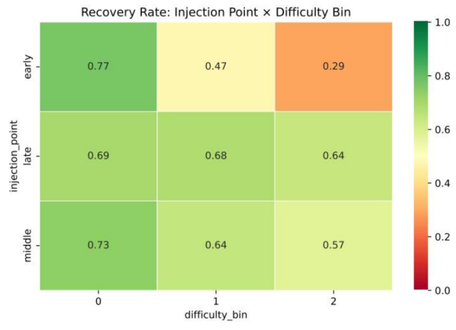

Figure 2:Recovery Rate: Injection Point × Difficulty Bin (Qwen2.5-7B, GSM8K). Each cell shows the proportion of trials with full recovery. Colour scale: red = 0.0, green = 1.0.

Recovery rate by injection point and difficulty

Recovery rates are substantially higher for easy problems (bin 0) under early and middle injection — 0.77 and 0.73 respectively. For hard problems (bin 2), early injection drops recovery to 0.29, indicating that the model is more reliant on its own step-by-step trace for complex reasoning and is less able to internally correct a wrong early step.

Under late injection, recovery rates are substantially more stable across difficulty bins (0.69, 0.68, 0.64) than for early or middle injection, suggesting that late injection is less disruptive because the model can more reliably override a single wrong terminal step based on prior context.

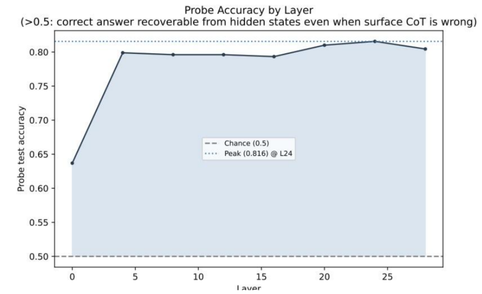

Figure 3:Linear Probe Accuracy by Layer (Qwen2.5-7B, GSM8K). The probe predicts whether the model will recover the correct answer from hidden states extracted at the injection point.

Probing Analysis: Hidden State Dissociation

Probe accuracy rises sharply from 0.64 at layer 0 to approximately 0.80 by layer 4, and remains broadly stable above 0.79 throughout the middle and later layers, peaking at 0.816 at layer 24 (of 28 total transformer layers). This means that from very early in the forward pass at the injection point, the model’s hidden representations carry substantial information about whether it will ultimately recover the correct answer — information that is largely independent of what the surface CoT token says at that position.

Combining probe predictions with behavioural outcomes further reveals a meaningful proportion of trials where the probe at peak layer predicts recovery (probe confidence > 0.6) but the model’s output instead propagates the injected wrong answer.

References

[1] Xi Chen, Aske Plaat, and Niki van Stein.
How does chain of thought think? mechanistic interpretability of chain-of-thought reasoning with sparse autoencoding, 2025.

[2] Subhabrata Dutta, Joykirat Singh, Soumen Chakrabarti, and Tanmoy Chakraborty.
How to think step-by-step: A mechanistic understanding of chain-of-thought reasoning, 2024.

[3] Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett.
To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning, 2025.

[4] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou.
Chain-of-thought prompting elicits reasoning in large language models, 2023.