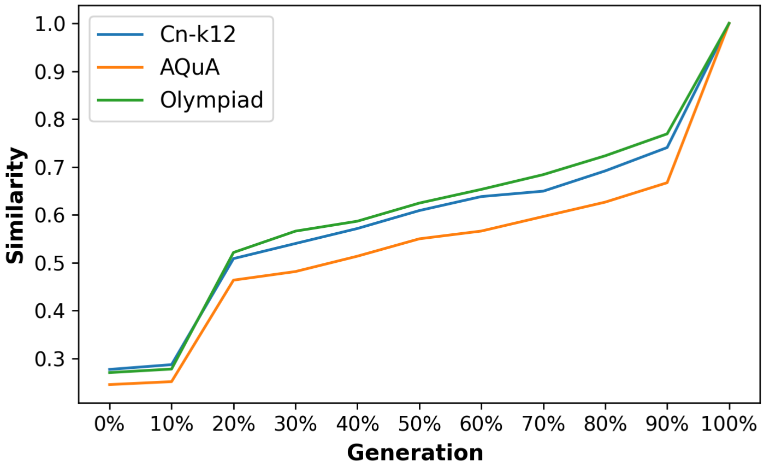

Figure 3: Similarity scores using SVCCA between hidden representations of each time step and the full generation, using Layer 14 for all three `Llama-3.1-8B` datasets.

during generation, as shown in Table 3. To do this, we capture the hidden states at various time intervals—after the LLM has generated 10%, 20%, and so on of the total content—and use these states to train the probe. The goal is to assess whether the LLM's understanding of success or failure becomes more apparent as more content is generated. This aligns with the idea that as more information is revealed, the model's understanding of success or failure becomes clearer, similar to how humans typically gain better insight into a task as they progress further. We selected the layer that performed best on the given dataset in prior evaluations for the prediction-over-time experiments. Interestingly, in the `Llama-3.1-8B`-generated dataset, BERT's performance tends to decline as more context is revealed, whereas for `Mistral-7B`, it remains relatively stable. While BERT can effectively process surface-level cues such as question structure, it appears less capable of tracking the evolving reasoning embedded in the chain-of-thought (CoT) generation. As the CoT unfolds and more tokens are added, the increasing complexity may exceed BERT's shallow interpretive capacity, limiting its ability to follow deeper logical developments. In contrast, using internal LLM representations as input significantly improves performance—e.g., from 60% to 69.4% on AQuA in the `Llama-3.1-8B`-generated dataset—and yields consistent gains across all `Mistral-7B` datasets. For Olympiad and cn-k12 with `Llama-3.1-8B`, results remain comparable. This demonstrates the representations' ability to capture more nuanced, cumulative reasoning and better reflect the underlying complexity of the CoT process, ultimately enabling more accurate success prediction. Surprisingly, `Mistral-7B`-generated AQuA dataset performs better as compared to the `Llama-3.1-8B`-generated one. Further inspection of the generations revealed that for the negative examples, `Mistral-7B` generations often assume that the provided question might be incorrect. These linguistic queues as shown in Table 4, might be used by BERT baseline as a feature to classify negative examples.

Complimentary Analysis for Probing Our probing results suggest that the initial CoT steps in the Olympiad and Cn-K12 `Llama-3.1-8B` generated datasets contain similar knowledge about CoT success as the later stages, which may indicate that they share an understanding of the calculations performed up to that point. However, prior research has highlighted several limitations of probing. For instance, when trained on a large number of examples, models using randomly initialized embeddings as input representations can outperform those using pretrained intermediate representations (Zhang and Bowman, 2018; Wieting and Kiela, 2019). To complement probing-based analyses, several alternative methods assess knowledge in model representations by measuring their correlations without relying on a classifier. These include SVCCA (Singular Vector Canonical Correlation Analysis) (Raghu et al., 2017), PWCCA (Projection Weighted Canonical Correlation Analysis) (Morcos et al., 2018), and CKA (Centered Kernel Alignment) (Kornblith et al., 2019), among others. Such methods are frequently used as alternatives for probing in NLP research (Saphra and

Example 1

1. First, we need to find the probability that Kyle and Catherine solve the problem but David does not.
2. The probability that Kyle solves the problem is 1/3, and the probability that Catherine solves the problem is 5/9.
3. The probability that David does not solve the problem is 1 - 2/7 = 5/7.
4. To find the probability that all three events happen, we multiply their individual probabilities: (1/3) * (5/9) * (5/7).
5. Simplify the expression: (5/27) * (5/7) = 25/189.
6. **The answer is E) 20 / 189. However, there seems to be a mistake in the options provided. The correct answer should be 25/189, not 20/189. So, the correct answer is E) 25 / 189.**

Example 2

1. The total investment is Rs. 70000 (A) + Rs. 120000 (B) = Rs. 190000.
2. The total profit is Rs. 26000.
3. To find the share of B, we need to divide the total profit by the total investment and then multiply it by B's investment.

Share of B = (Total profit / Total investment) * B's investment
= (26000 / 190000) * 120000
= (26/19) * 12000
= 24000 * (26/19)
= 24000 * (1.3158)
= 24009
So, the share of B is approximately Rs. 24009.
However, since the options provided are not exact, we can round the answer to the nearest option. The closest option is D) 24009. But since the question asks for one of the options, we cannot be certain that D) 24009 is the correct answer. It's a tricky question due to the rounding of the options.
Final answer: The share of B is approximately 24009, but the question asks for one of the options, so the exact answer cannot be determined from the given options.

Table 4: Linguistic Patterns in the negative examples of `Mistral-7B`-generated AQuA dataset, which might be used by BERT classifier as a feature in our Prediction over Time experiments.

Lopez, 2018; Zhao et al., 2023, 2024). In our study, we employ SVCCA following Zhao et al. (2022), which investigated how domain-specific information is encoded in LLMs—a research focus related to ours.

We compute SVCCA scores between representations at each reasoning step[5] (10%, 20%, ..., 90%) with 100% (CoT completion), where higher scores indicate greater similarity in encoded information (Figure 3). Since the AQuA dataset consists of shorter questions and CoT sequences (see Table 2), one might expect its intermediate representations to be more similar across steps, as the reasoning process is more compact. However, we observe that AQuA exhibits lower SVCCA scores between earlier and later steps compared to Olympiad and Cn-K12. This strengthens our probing findings: when earlier steps contain information predictive of CoT success, representations remain more stable throughout the reasoning process. This suggests that intermediate representations may not only encode predictive information about the final answer correctness but also the final answer itself. If the model is implicitly performing parts of the final computation at earlier steps, we may be able to leverage this by directly prompting it to provide an answer before completing the full chain of thought.

Early Stopping in CoT Reasoning To examine whether the model's intermediate representations encode sufficient information for correct answers, we prompt it to generate an answer at various reasoning steps and evaluate its accuracy. As in our previous experiments, this intervention is performed in a zero-shot manner without explicitly training the model to follow such instructions. Specifically, we halt the `Llama-3.1-8B` reasoning process by providing the generated sequence up to a certain point, followed by the instruction: *"Stop all computation and give me the correct answer in 2–3 words, if you already know it"*. This allows us to assess whether the model can extract a final answer without completing the full chain of thought (See Appendix B for an illustration). We conduct human annotation on 100 samples per dataset at three timesteps[6], summarizing the results in Table 5. Our analysis examines how often the halted response remains consistent with the final, uninterrupted answer, how often it changes (inconsistent), and in what fraction of those cases the change leads to a corrected answer—where stopping CoT reasoning actually improves performance. Finally, we compare overall correctness rates to both the full CoT process and a setting without CoT. The relatively low consistency rate, even at 99% comple-

[5] See Appendix B for scores between all possible combinations of representations reasoning step

[6] We select two intervals where the classification model achieves its highest accuracy and one at the final step.

Dataset	Gen %	Consistent	Inconsistent	Corrected	Correct	W_CoT_Correct
AQuA	50	40	45	15	37	38
	70	48	41	11	37	
	99	81	14	5	59	
Olympiad	30	21	76	3	18	19
	50	28	70	2	22	
	99	57	41	2	33	
Cn-k12	30	37	55	7	32	24
	50	42	49	7	33	
	99	88	9	3	47	

Table 5: LLM's generation was artificially interrupted to halt computation and were asked to just provide their best guess on 100 samples at 3 different time steps defined Gen. Annotators marked an answer Consistent if it is same the answer it provides when it is allowed to continue generation and Inconsistent if the provided answer differ. There were cases where full generation led to incorrect solutions by LLM and these interruptions allowed LLM to generate the correct answer, which was given the label Correct by annotators.

tion—particularly in Olympiad (57%)—suggests that zero-shot early stopping is a suboptimal intervention. The model does not always converge to a stable answer, even when nearly the entire reasoning sequence is generated, highlighting the limitations of simply prompting it to stop early. However, despite this brittleness, halting CoT midway in AQuA and Cn-K12—the two datasets where later reasoning steps did not enhance CoT success predictability—still slightly outperforms the setting without CoT. This indicates that even incomplete CoT sequences can carry enough information to improve accuracy, revealing untapped potential in intermediate reasoning states. These findings suggest that while zero-shot interventions have limitations, more sophisticated approaches—such as training the model to generate concise reasoning chains through supervised learning or reinforcing brevity via RL-based rewards—could more effectively unlock this potential. By optimizing the model's ability to extract key reasoning steps without unnecessary verbosity, future methods could further bridge the gap between full CoT and early stopping while maintaining or even improving accuracy.[7]

6 Conclusion

We demonstrate that the success of the CoT reasoning process can be predicted from the internal representations of the LLM even before the generation of a single token. However, we also observe that, in some cases, the accuracy of this prediction does not improve when the classifier is exposed to intermediate reasoning steps. Using SVCCA, we show that early steps encode information that is more similar to the final steps in these cases. This raises the question of whether these early representations also contain valuable information about the final answer itself. Our initial experiments suggest that while this potential exists, zero-shot prompting may not fully unlock it. We hope our findings will inform future research aimed at making CoT more efficient without sacrificing accuracy, as the computational cost of CoT is significant.

Limitation

Manually annotating test examples for each dataset limits the generalizability of our study, as we use human evaluation for only a single LLM and focus on three math datasets in a zero-shot setting. Additionally, we use a temperature of zero to minimize stochastic noise in the generation process, which could accumulate over multiple reasoning steps. However, this setting may not fully capture the variability present in real-world, stochastic LLM usage. Furthermore, our method assumes white-box access to the model, which is not typically available for proprietary models.

Ethics Statement

Our research focused on evaluating the internal representation of a Large Language Model to better the notion of prediction for CoT Reasoning. During out research, we didn't perform any fine-tuning

[7] In some cases, halting CoT before completion improved answer correctness, as reflected in the *corrected* column in the table.