A video chain-of-thought dataset with active annotation tool. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 92–101, Bangkok, Thailand. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

John Wieting and Douwe Kiela. 2019. No training required: Exploring random encoders for sentence classification. *arXiv preprint arXiv:1901.10444*.

Ancheng Xu, Minghuan Tan, Lei Wang, Min Yang, and Ruifeng Xu. 2024a. NUMCoT: Numerals and units of measurement in chain-of-thought reasoning using large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14268–14290, Bangkok, Thailand. Association for Computational Linguistics.

Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. 2024b. Faithful logical reasoning via symbolic chain-of-thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13326–13365, Bangkok, Thailand. Association for Computational Linguistics.

Kelly W Zhang and Samuel R Bowman. 2018. Language modeling teaches you more syntax than translation does: Lessons learned through auxiliary task analysis. *arXiv preprint arXiv:1809.10040*.

Zheng Zhao, Yftah Ziser, and Shay Cohen. 2022. Understanding domain learning in language models through subpopulation analysis. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 192–209, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Zheng Zhao, Yftah Ziser, and Shay B Cohen. 2024. Layer by layer: Uncovering where multi-task learning happens in instruction-tuned large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15195–15214, Miami, Florida, USA. Association for Computational Linguistics.

Zheng Zhao, Yftah Ziser, Bonnie Webber, and Shay Cohen. 2023. A joint matrix factorization analysis of multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12764–12783, Singapore. Association for Computational Linguistics.

Mingyu Zheng, Hao Yang, Wenbin Jiang, Zheng Lin, Yajuan Lyu, Qiaoqiao She, and Weiping Wang. 2023. Chain-of-thought reasoning in tabular language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11006–11019, Singapore. Association for Computational Linguistics.

A Technical Details

Classification Models We performed hyperparameter-tuning for each dataset (and each layer) when training the classification models. We illustrate the best hyperparameters in Table 7 and used them in our Prediction over time experiments.

B Additional Results

Complimentary Analysis for Probing We compute SVCCA scores between all possible combinations of representations reasoning step (10%, 20%, ..., 90%) and with 100% (CoT completion) and show them in Figure 4.

Early Stopping in CoT Reasoning We show an illustration of Early Stopping in CoT Reasoning and the LLM response in Table 8. It can be seen that despite being prompted to stop calculations to 30% of it's original generatation, it still generates the correct answer.

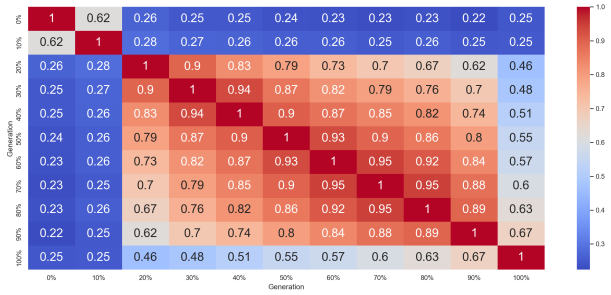
Classification Performance Breakdown Following our earlier evaluations, we perform an in-depth analysis of the classification model's performance on a test set of 1,000 examples, summarized in Table 9. Unlike the BERT baseline, which performs well on true positives (TP) only, our model excels at identifying both true negatives (TN) and true positives (TP). For example, on the Llama-3.1-8b generated Cn-k12 dataset, BERT identified 250 true negatives, while our model identified 317. Similarly, on AQuA, BERT's true negatives were 131, while our model achieved 177, and on Olympiad, BERT had 315 true negatives, compared to 388 for our method. These results show that our model not only performs well on true positives but also significantly outperforms BERT in detecting true negatives. This suggests that our method captures

Cn-k12	AQuA	Olympiad
layer 31	layer 32	layer 13
layer 10	layer 10	layer 28
layer 11	layer 13	layer 32
layer 13	layer 12	layer 10
layer 17	layer 5	layer 3

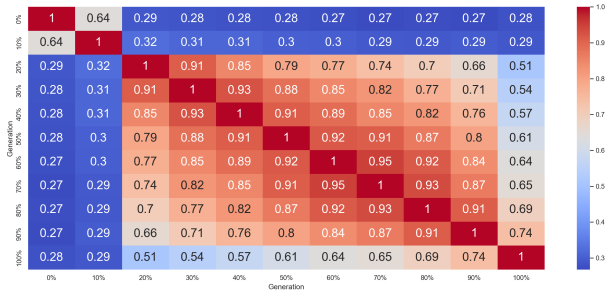
Table 6: Five most dissimilar layers using SVCCA (Singular Vector Canonical Correlation Analysis) between hidden representation with and with CoT prompt per layer, averaged over 1000 samples.

Dataset	batch size	weight init	learning rate	optimizer	threshold
		BERT (baseline)			
AQuA	128	HE_uniform	0.001	**sgd**	0.6
Olympiad	128	HE_uniform	0.001	adam	0.6
Cn-k12	128	-	0.001	adam	0.6
		Our Model			
AQuA	32	HE_uniform	0.001	**adam**	0.5
Olympiad	128	HE_normal	**0.001**	adam	0.5
Cn-k12	16	HE_uniform	0.001	sgd	0.5

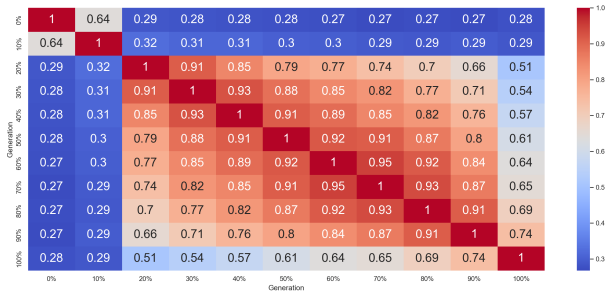
Table 7: Accuracy of the classification model before it starts generating and over time when model has generated x% of the answer.



(a) AQuA



(b) Cn-k12



(c) Olympiad

Figure 4: Similarity Scores using Singular Vector Canonical Correlation Analysis (SVCCA) of the internal representation of LLM through Prediction over Time using Layer 14.

A high school has a total of 900 students, among which there are 300 freshmen, 200 sophomores, and 400 juniors. Now, a stratified sampling method is used to select a sample of 45 students. How many students should be selected from each grade, respectively? A: 15, 5, 25 B: 15, 15, 15 C: 10, 5, 30 D: 15, 10, 20 Let's think step by step: To determine the number of students to be selected from each grade, we need to calculate the proportion of students in each grade and then apply this proportion to the total sample size of 45 students. Step 1: Calculate the proportion of students in each grad *Stop all computation and give me the correct answer in 2- 3 words, if you know it already.* Answer **15, 10, 20**
Reference Answer The correct answer is D.

Table 8: An Illustration of a sample showing how we artificially halt LLM generation when it has generated 30% of the answer.

more relevant predictive signals, enabling a more nuanced understanding of the factors that drive Chain-of-Thought success.

LLM Layer Analysis We also include the representation of LLM without any CoT reasoning in our experiments. In Table 6, we illustrate the 5 most dissimilar layers. Since they were both encoded using the same problems, the only different is the inclusion of CoT Reasoning, the disimilarity among the layers may just a more notable role in LLM's internal notion of CoT Reasoning.

C Dataset Examples

Success Prediction over Time An illustration of samples from the Success Prediction over Time method is shown in Table 10 for 2 different generation steps.