selected based on empirical performance for each dataset. We perform a hyperparameter tuning specific to the task and detail the selected configurations in Appendix A. The classifier is trained for five epochs across all datasets and LLMs.

Success Prediction over Time While our primary methodology focuses on predicting CoT performance before LLM generation begins, we also explore how this prediction may change during the generation process. Identifying atomic steps in the generation can be challenging due to their varying lengths and styles, so for simplicity, we instead use percentages of the total number of tokens generated (10%, 20%, up to 90%) and concatenate them with the initial prompt (see Appendix C for illustration). This approach allows for a consistent evaluation across different generations. As in the previous approach, we extract hidden layers for partial generations at 10% intervals, enabling us to observe the evolution of internal representations. We apply a similar strategy to expand our test sets.

CoT	wo CoT	success rate		
		Cn-k12	**AQuA**	**Olympiad**
`Llama-3.1-8B`				
0	0	41.83%	47.46%	44.76%
0	1	8.11%	2.53%	4.77%
1	0	32.52%	43.32%	36.97%
1	1	17.44%	6.67%	12.92%
`Mistral-7B`				
0	0	44.07%	39.43%	41.43%
0	1	5.89%	10.54%	8.46%
1	0	32.33%	33.83%	28.08%
1	1	17.69%	16.2%	22.03%

Table 1: We depict the percentage of Problems the LLM was able to solve with and without any Chain-of-Thought prompting on a balanced dataset where CoT prompting helped solve 50% of the problems. We show a confusion matrix such that 0 means LLM was not able to solve the problem, and 1 means that it was able to solve the task.

4 Experimental Setup

4.1 Baseline

The prediction of CoT success may depend more on linguistic cues in the text than on internal LLM representations encoding intermediate computations. To explore this, we build on (Azaria and Mitchell, 2023), which examined whether LLMs store information in their internal states, albeit for different purposes. Our goal is to demonstrate that such information is indeed retained within the LLM's representations. To test this, we use BERT as a baseline, as it relies solely on textual tokens without access to internal LLM states, effectively functioning as a black-box access approach. Given BERT's strength as a text classifier, its ability to predict CoT success would indicate that textual cues alone suffice. We use the `google-bert/bert-base-uncased` variant with default settings, maintaining a consistent neural network-based classification setup while varying the input features.[2]

4.2 Large Language Models

For the hidden representations of `Llama-3.1-8B` and `Mistral-7B`, we use the `meta-llama/Llama-3.1-8B-Instruct` and `mistralai/Mistral-7B-Instruct-v0.3` checkpoints on huggingface respectively.

4.3 Datasets

We used three math problem datasets of varying difficulty in our experiments: World Olympiads Data (Olympiad) (LI et al., 2024), Chinese K-12 Exam (cn-k12) (LI et al., 2024), and AQuA (aqua) (Ling et al., 2017). To assess model behavior across different reasoning patterns, we ran each dataset with two different LLMs—`Llama-3.1-8B` (Grattafiori et al., 2024) and `Mistral-7B`—resulting in six distinct dataset variants. Since we enforce class balance (i.e., an equal number of correct and incorrect generations), the two versions of the same original dataset may contain different sets of questions, depending on the LLM's outputs. Given the difficulty level of questions within each dataset, the success rate using `Llama-3.1-8B`, for example, varies considerably, reaching 22%, 28%, and 62.3% on Olympiad, cn-k12, and AQuA, respectively. For all datasets and both LLMs, we ran inference to obtain balanced train (10k) and test (1k) sets with an equal distribution of positive and negative examples. Consequently, our classification model is trained on a 10000×4096 feature space for each dataset and LLM layer combination. We reserve 10% of the training set for validation to tune hyperparameters. The distribution of question and generation lengths is summarized in Table 2. Lastly, it is also important to evaluate how well the same

[2]Training features vary when using BERT embeddings (dimension 768) or LLM layers (dimension 4096).

Dataset	# Question Tokens			# Generation Tokens		
	avg	min	max	avg	min	max
`Llama-3.1-8B`						
AQuA	42.96	5	308	287.24	17	512
Olympiad	67.76	3	1109	401.56	2	512
Cn-k12	76.38	7	520	327.46	4	512
`Mistral-7B`						
AQuA	84.23	38	344	255.07	35	509
Olympiad	86.10	9	1108	368.47	40	511
Cn-k12	88.94	5	679	304.96	29	511

Table 2: Average, Minimum and Maximum token count of Questions and LLM generation using `Llama-3.1-8B` and `Mistral-7B` tokenizers on all three datasets. Generation Tokens are capped at 512 using max_new_tokens = 512.

LLM performs when solving the same problems without chain-of-thought (CoT) prompting. We present this comparison in a confusion matrix in Table 1, which shows that, for instance, without CoT, `Llama-3.1-8B` achieved only 9.2%, 17.69%, and 25.55% success rates on AQuA, Olympiad, and cn-k12, respectively, substantially lower than the 50% CoT accuracy we enforce when collecting a balanced dataset.

Olympiad: The Math Olympiad is a competitive examination designed to evaluate students' mathematical skills and competencies. We use the olympiad dataset from the NuminaMath-CoT (LI et al., 2024) collection of the dataset. This dataset is a collection of problems and respective answers following a CoT format collected from international/national contests as well as forums, books, and summer school materials.

Cn-k12: This large-scale Chinese K-12 education math exercise dataset was translated and re-aligned to English using GPT-4. This dataset is also part of the NuminaMath-CoT (LI et al., 2024) collection of datasets, and reference answers follow a CoT format.

AQuA: This dataset (Ling et al., 2017) consists of algebraic math problems, each presented with a step-by-step reference solution. Unlike the other two datasets, it includes multiple-choice options. We use the original dataset released by the authors.

4.4 Manual Annotation

Given the large dataset size, we used `GPT-4o mini`[3] to label the training and validation sets for the first language model (`Llama-3.1-8B`). To ensure the reliability of our evaluation, we manually annotated the test set using annotators selected from a local university's STEM Master's program, compensated at 16 euros per hour. Annotators evaluated the correctness of each generation based on the question and a reference answer. On a test set of 1,000 samples, human annotations agreed with LLM-generated ones on 90.9%, 94.8%, and 93.4% for AQuA, Cn-K12, and Olympiad, respectively. These numbers indicate that although LLM-generated labels are generally reliable, the disagreement with human judgment suggests caution when using them for evaluation. For the second language model (`Mistral-7B`), all datasets (training, validation, and test) were annotated using `GPT-4o mini`. Given the imperfect agreement observed between human and `GPT-4o mini` annotations for the first model, these results should be interpreted with appropriate caution.

[3] https://platform.openai.com/docs/models#gpt-4o-mini

5 Results and Discussion

Since the test set for `Llama-3.1-8B` was manually annotated by human experts, it is considered more reliable than the `GPT-4o mini`-annotated test set used for `Mistral-7B`. As a result, in certain analyses where only one LLM is presented, we focus solely on `Llama-3.1-8B` to ensure more reliable evaluation.

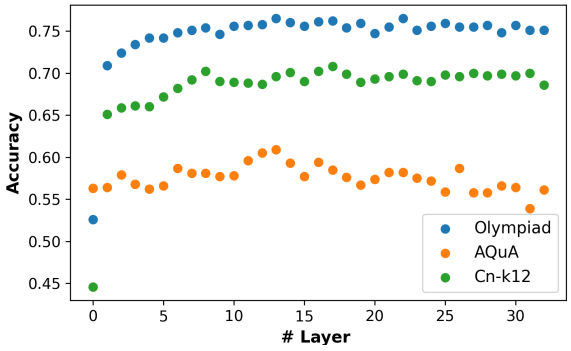

Figure 2: Accuracy on the test set when the probe is trained and tested on hidden representations from each layer. Results are shown for all 33 layers of Llama 3.1 8B Instruct across all three datasets.

5.1 Prediction before Generation

Main Results We evaluate the information contained in the internal representation of the LLM before it begins generating and present the results in Table 3, specifically under the %0 column. Given that 50% of answers in the dataset are correct (bal-

	Before Gen		Over Time									
Dataset	top-5 layers	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
`Llama-3.1-8B` *(human-annotated)*												
BERT (baseline)												
AQuA	-	53.50	54.2	55.8	54.1	54.1	53.5	51.7	51.6	53.7	51.7	50.6
Olympiad	-	69.10	67.9	68.7	66.2	66.4	67.4	65.1	67.9	66.8	67.1	66.9
Cn-k12	-	66.20	57.8	64.4	64.1	63.8	64.3	63.8	62.7	61.7	61.1	62.4
Our Model												
AQuA	11, 12, 13, **14**, **16**	60	51.5	60	60.5	63.2	60.11	60.8	63.7	62.9	65.7	**69.4**
Olympiad	8, **14**, **16**, 17, 31	76.4	75.1	75.8	73.8	74.6	**76.5**	73.8	75.3	75.3	73.8	75.9
Cn-k12	13, **14**, **16**, 17 22	69.10	67.7	69.9	69.2	67.7	67.2	70	**71.6**	70.7	68.9	70.9
`Mistral-7B` (`GPT-4o mini` *annotated)*												
BERT (baseline)												
AQuA	-	60.1	57.8	59.5	63.4	65.2	63.7	67.8	66.7	66.8	68.1	71.8
Olympiad	-	68.8	68.1	69.6	68.8	68.3	68.6	66.4	67.1	68.3	68.6	69.6
Cn-k12	-	65.5	63.5	65.8	65.1	64.4	64.8	66.5	66.7	67.3	67.8	68.5
Our Model												
AQuA	15, 16, **18**, 23, **28**	64.7	54.4	66.1	66.9	66.4	65.1	65.8	64.1	66.6	67.4	**80.2**
Olympiad	7, 9, **18**, 26, **28**	71.8	71.7	72.0	72.3	74.1	74.2	75.6	74.5	75.3	75.5	**75.9**
Cn-k12	12, 14, **18**, 21, 24	67.1	68.0	68.4	66.7	67.7	67.8	67.1	68.1	68.6	67.6	**71.4**

Table 3: Classification model accuracy before generation begins and as it progresses, measured at different completion percentages. The best-performing layers and time steps are highlighted in bold.

anced data), a random classifier would achieve an accuracy of 50%, which serves as the baseline for interpreting the results. The BERT baseline, which relies exclusively on token representations on CoT prompt with question, consistently outperformed random chance across all six datasets. On generations from `Llama-3.1-8B`, BERT achieved accuracy scores ranging from 53.5% to 69.1%. For `Mistral-7B`, BERT's accuracy ranged from 60.1% to 68.8%. However, the predictive power appears relatively weaker for the AQuA dataset[4]. Our suggested method, which uses the LLM's internal representations, outperformed the BERT baseline across all datasets, demonstrating their importance for predicting CoT success before generation begins. For instance, on `Llama-3.1-8B` datasets, our model achieved 60.0% accuracy on AQuA, significantly higher than BERT, and 76.4% on Olympiad, compared to BERT's 69.1%. We observe similar patterns for the `Mistral-7B` datasets. The large variance in results (ranging from 60.0% to 76.4%) suggests that the effectiveness of the internal representations may be influenced by factors such as task complexity or dataset characteristics, which warrant further investigation. See Appendix B for details on Classification Performance Breakdown.

LLM Layers Analysis In our experiments, we evaluated the accuracy of the classification model when trained on the hidden representations of each layer. We show the accuracy per layer of `Llama-3.1-8B` in Figure 2. It can be seen that the middle such as layers 11 - 14 and layers 16 - 17, and in some cases the last layers of the LLM seem to play a role in it's internal notion of Success or Failure. As the in shown Table 3, we find layer 14 and layer 16 to be consistent among all three datasets suggesting these layers to be more involved in the notion of CoT prediction. Our findings are inline with those of Azaria and Mitchell 2023, who show similar results regarding LLM's notion of truthfulness and show layer 16 and in some cases the last layers of LLM to be most knowledgable for their task.

5.2 Prediction over Time

Main Results We evaluate the information contained in the internal representation of the LLM

[4]We hypothesize that accuracy on `Llama-3.1-8B`-generated AQuA dataset might be lower as the sampled questions might contain less linguistic patterns than the other datasets.