

Self-Training Elicits Concise Reasoning in Large Language Models

Tergel Munkhbat* Namgyu Ho* Seo Hyun Kim* Yongjin Yang
Yujin Kim Se-Young Yun†

KAIST AI

{tergelmunkhbat, itsnamgyu, shkimsally, yunseyoung}@kaist.ac.kr
<https://github.com/TergelMunkhbat/concise-reasoning>

Abstract

Chain-of-thought (CoT) reasoning has enabled large language models (LLMs) to utilize additional computation through intermediate tokens to solve complex tasks. However, we posit that typical reasoning traces contain many redundant tokens, incurring extraneous inference costs. Upon examination of the output distribution of current LLMs, we find evidence on their latent ability to reason more concisely, *relative* to their default behavior. To elicit this capability, we propose simple fine-tuning methods which leverage self-generated concise reasoning paths obtained by best-of-N sampling and few-shot conditioning, in task-specific settings. Our combined method achieves a 30% reduction in output tokens on average, across five model families on GSM8K and MATH, while maintaining average accuracy. By exploiting the fundamental stochasticity and in-context learning capabilities of LLMs, our self-training approach robustly elicits concise reasoning on a wide range of models, including those with extensive post-training.

1 Introduction

Chain-of-thought (CoT) reasoning has significantly improved the ability of large language models (LLMs) to perform complex tasks (Wei et al., 2022b). The effectiveness of CoT reasoning has been attributed to the additional computation allocated during inference, as each intermediate reasoning token enables the model to perform an additional forward pass through its parameters (Nye et al., 2021; Wei et al., 2022b). On the other hand, this inherently incurs additional inference cost and latency, roughly proportional to the number of output tokens (Agrawal et al., 2024; Ho et al., 2024).

We posit that current models often generate more tokens than necessary to accomplish the task, incurring extraneous inference costs. Typical reasoning chains sampled from the original model

*Equal contribution. †Corresponding author.



A candle melts by 2 centimeters every hour that it burns. How many centimeters shorter will a candle be after burning from 1:00 PM to 5:00 PM?



To find out how many centimeters shorter the candle will be, we need to determine how many hours it will burn from 1:00 PM to 5:00 PM. Step 1: Calculate the number of hours the candle will burn. From 1:00 PM to 5:00 PM, the candle will burn for 4 hours. Step 2: Determine how much the candle will melt in 4 hours. Since the candle melts by 2 centimeters every hour, we need to multiply the number of hours it will burn by the amount it melts per hour. Step 3: Multiply the number of hours by the amount the candle melts per hour. $4 \text{ hours} * 2 \text{ centimeters/hour} = 8 \text{ centimeters}$. Therefore, the candle will be 8 centimeters shorter after burning from 1:00 PM to 5:00 PM. The answer is 8



The candle burns for 4 hours. It melts 2 centimeters every hour. So, it will melt $2 * 4 = 8$ centimeters. The answer is 8 centimeters.

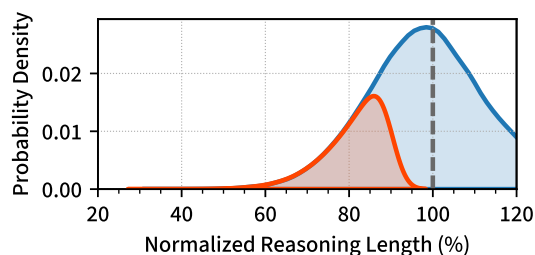


Figure 1: **Potential of concise reasoning in current LLMs.** Examples of reasoning samples with different lengths and the normalized length distribution of correct reasoning paths from Llama-3.1-8B on GSM8K. Typical samples from the *original model distribution* marked in blue, as shown in the blue box, contain redundant context as marked in gray. We fine-tune the model with relatively concise reasoning paths within the *subregion of the model output distribution* marked in orange. This elicits the model to reason more efficiently, as shown in the orange box, retaining only the essential parts while still leading to the correct answer.

distribution, exemplified in the blue box of Figure 1, includes many tokens that do not contribute to the solution, specifically verbose explanations and repetitive phrasing. Similar observations have also been made in prior work (Renze and Guven, 2024; Zhang et al., 2024; Chiang and Lee, 2024).

We argue that the redundancy in reasoning chains of current models is not surprising. LLMs have not been explicitly trained to utilize intermediate tokens for reasoning in the most efficient man-

ner. Rather, CoT reasoning is an emergent ability of LLMs (Wei et al., 2022a) first *uncovered* through novel few-shot prompting (Wei et al., 2022b). Post-hoc analysis suggests that reasoning ability is derived from procedural knowledge in pretraining data (Ruis et al., 2024), which is not optimized for conciseness. Therefore, it is natural that the default behavior of LLMs does not optimize for efficient use of tokens for reasoning. However, this does not preclude the possibility that LLMs *possess* the ability to reason more efficiently.

Indeed, the distribution of reasoning path lengths shown in Figure 1 reveals a compelling insight: the presence of shorter, more efficient reasoning paths within the model’s output distribution, marked in orange, suggests a latent capacity for concise reasoning. This observation motivates our central hypothesis: that by selectively leveraging these existing concise examples, we can fine-tune the model to elicit this latent ability and shift its default output distribution towards more efficient reasoning.

In this paper, we propose a simple yet effective fine-tuning method to elicit efficient, concise reasoning in models for a given target task. We note that existing zero-shot prompting methods fail to reliably elicit concise reasoning, exhibiting near negligible impact on task-specialized models (Section 2.2). Instead, we can leverage best-of-N (BoN) sampling and few-shot conditioning (FS) to reliably generate concise training data (Section 3). We can then apply standard fine-tuning to distill the length-reducing benefits of BoN sampling and few-shot prompting back into the model itself while avoiding inference-time overheads.

Across GSM8K and MATH datasets and a wide variety of model families, our unified few-shot conditioned best-of-N sampling (FS-BoN) method achieves a substantial reduction in output length of 30% on average, a 2.4x improvement over previous fine-tuning baselines (De Sabbata et al., 2024), while preserving overall accuracy. Our analysis shows that trained models adaptively adjust output length based on question complexity, preserving detail for difficult problems while simplifying responses to easier ones. We confirm that these results remain consistent across various model scales and diverse reasoning domains such as business, chemistry, and physics. This shows that fine-tuning with carefully curated, self-generated data can effectively unlock latent concise reasoning abilities within LLMs, leading to significantly more cost-effective inference for complex tasks.

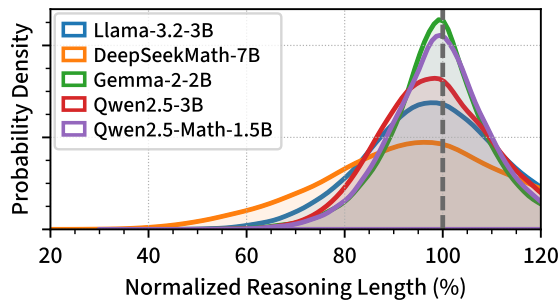


Figure 2: **Models already exhibit relatively concise reasoning in their output distributions to differing degrees.** Distribution of output token lengths in reasoning paths of GSM8K, normalized by the average length of correct reasoning paths for each question, across five main models. Reasoning paths to the left of the gray line are relatively shorter than the model’s *default output* (Section 2)

2 Preliminary Investigation

In this section, we probe the potential of current models to perform *concise reasoning* (Section 2.1) and investigate whether zero-shot prompting can reliably elicit such behavior (Section 2.2). We start with a definition of concise reasoning and key experimental setup.

Concise reasoning We define concise reasoning in relative terms: for a given model, concise reasoning is to correctly reason through a given problem using less *output* tokens, compared to its default output. *What is the default output?* For our preliminary investigation, we measure the default output length of a model on a given question by taking the average length of correct paths over multiple stochastic generations. We use this measure instead of greedy decoding length, as the greedy path may be incorrect. We also consider the greedy output length as a noisy estimate, as it is highly sensitive to superficial input perturbations and numerical errors (Holtzman et al., 2019).

Key experimental setup We analyze a wide range of moderately sized models, including both instruction-tuned language models and models specifically fine-tuned for mathematical reasoning, to assess robustness. We consider two reasoning tasks of moderate difficulty to assess preservation of model performance: GSM8K and MATH. We explain our full experimental setup in Section 4.

Method	Avg. Relative		Ll-3B		Ge-2B		Qw-3B		QM-1.5B		DM-7B	
	Acc	Len	Acc	Len	Acc	Len	Acc	Len	Acc	Len	Acc	Len
Default (Pang et al., 2024)	100.0	100.0	77.3	220.0	65.4	200.1	84.4	301.1	82.8	305.9	80.4	182.3
Be Concise (Renze and Guven, 2024)	99.9	88.5	78.2	168.4	64.2	163.1	83.8	258.2	82.9	304.1	80.7	180.6
Est. Budget (Han et al., 2024)	79.8	121.4	73.2	243.6	54.5	124.7	22.2	29.6	80.2	609.9	78.5	409.3
Fixed Budget (Nayab et al., 2024)	89.9	67.8	71.7	129.7	56.5	129.9	63.8	68.8	81.7	298.1	77.4	173.1
Hand Crafted 1	97.4	81.1	77.9	179.7	60.4	123.6	80.1	186.5	81.7	306.1	80.8	182.3
Hand Crafted 2	98.3	77.1	79.2	143.6	62.6	144.9	78.8	146.4	82.5	303.5	80.6	182.2
Hand Crafted 3	91.9	68.4	73.6	101.7	50.7	127.9	74.0	88.6	81.5	307.4	80.7	185.5
Hand Crafted 4	61.3	48.8	35.0	49.5	29.8	49.5	14.5	7.7	81.2	297.0	80.7	177.1

Table 1: **Inconsistency of zero-shot prompting methods.** Zero-shot prompting methods for concise reasoning on GSM8K. Relative accuracy (%) and length (%) compared to default prompting are shown, averaged over five main models: Llama-3.2-3B, Gemma-2-2B, Qwen2.5-3B, Qwen2.5-Math-1.5B, and DeepSeekMath-7B. Absolute accuracy (%) and length (tokens) are reported for each individual model. Values in red indicate **significant relative accuracy drop** greater than 5%. Values in gray indicate **negligible relative length reduction** less than 5%.

2.1 Reasoning Length Distribution

To assess the potential of current models to elicit concise reasoning, we analyze the extent of concise reasoning present within their output paths. Following our definition of concise reasoning in this context, we analyze the *normalized* length of *correct* reasoning paths, which is obtained by dividing the length (number of output tokens) of each path by the average length of correct paths for that question. We stochastically sample 16 paths per question on the GSM8K training dataset and visualize the distribution of normalized lengths using kernel density estimation (Scott, 2015). Alternative analysis on absolute token count and raw character count, as well as results on the MATH dataset, are present in Appendix Section D.1.

Models can reason with fewer tokens than their default to varying degrees. In Figure 2, the distribution of normalized lengths show substantial mass below 100%, indicating frequent occurrences of solutions shorter than the default length. This suggests that models have the capability to produce more concise solutions than their average output would suggest. Notably, DeepSeekMath-7B, which already uses the least tokens in absolute terms (see Figure 8), produces correct solutions using *less than half* of its average token count in 8.37% of cases, demonstrating substantial potential for more concise reasoning. We find similar patterns on the MATH dataset, detailed in Section D.1.

2.2 Efficacy of Zero-Shot Prompting

While several zero-shot prompting methods have been proposed to elicit concise reasoning, evaluation has been limited to few models. We reassess

their impact on output length and reasoning performance across diverse model families, evaluating accuracy and output length, relative to that of a strong default prompt (Pang et al., 2024). *Be Concise* (Renze and Guven, 2024) appends the phrase ‘be concise’ to the baseline prompt. *Estimated Budget* (Nayab et al., 2024) prompts the model to ‘use less than N tokens,’ where N is determined by the model itself. *Fixed Budget* (Han et al., 2024) prompts the model to ‘limit the answer length to 100 words.’ In addition, we introduce four novel prompts optimized on Llama-3.2-1B, as our best-effort attempt at prompt optimization. Full prompts and prompting details are outlined in Appendix F.

Output length reduction typically comes at the cost of accuracy. As shown in Table 1, most zero-shot prompting methods reduce reasoning length but often lead to notable accuracy loss. For example, the *Fixed Budget* prompt shortens reasoning by 32.2% but lowers accuracy by 10.1% relative to the baseline, on average. Our hand-crafted prompts confirm that greater length reduction generally results in higher accuracy degradation.

Zero-shot prompting is less effective on task-specific models. Zero-shot prompting methods show inconsistent performance across model families. While some prompts are effective on general-purpose models like Llama-3.2, they fail to elicit concise reasoning on math-specialized models like Qwen2.5-Math (Table 1). This suggests the internal representations of task-specific models are less responsive to zero-shot prompting. Similar trends are observed on the MATH dataset in Table 6.

3 Method

To address the limitations of zero-shot prompting, we propose a simple yet robust fine-tuning (FT) approach to reliably elicit concise reasoning in LLMs while preserving accuracy. We focus on *self-training* approaches, to unlock and refine the *latent concise reasoning abilities* of current LLMs, observed in Section 2.1. Self-training not only removes external dependencies, but we posit that it helps preserve reasoning capability, as the training data originates from the model’s own distribution.

3.1 Naive Best-of-N Sampling (BoN)

We first consider naive BoN sampling to collect relatively concise reasoning samples from the original output distribution, corresponding to the probability mass toward the left side of the distribution in Figure 2, for fine-tuning. Specifically, we generate N reasoning paths for each question in the original training dataset and select the shortest *correct* reasoning path for each question¹. In contrast to selecting the shortest subset of correct reasoning paths from a combined pool of questions, our question-wise selection scheme ensures supervision across a wide range of difficulty levels, as difficult questions may require longer absolute reasoning lengths. Corresponding experimental results are in Section D.3.

3.2 Few-Shot Conditioning For Effective Reduction

While naive BoN sampling is a straightforward approach, its sample inefficiency (Xiang et al., 2025) makes it infeasible to achieve significant length reduction beyond a certain point. Figure 3 shows a log-linear relationship between N and output length reduction, indicating that length reduction through BoN incurs exponential generation costs.

Few-shot conditioned sampling (FS) To mitigate the sample inefficiency of naive BoN, we can leverage few-shot prompting to bootstrap the reduction in output length. We consider three sources for few-shot exemplars: human-annotation (*FS-Human*), proprietary frontier LLMs (*FS-GPT4o*), and self-generated samples (*FS-Self*). We use human annotated examples from Wei et al. (2022b), as they are readily available and demonstrate very concise reasoning. We provide additional details on few-shot example acquisition in Appendix B.

¹Note that questions without any correct reasoning paths are excluded from the fine-tuning dataset.

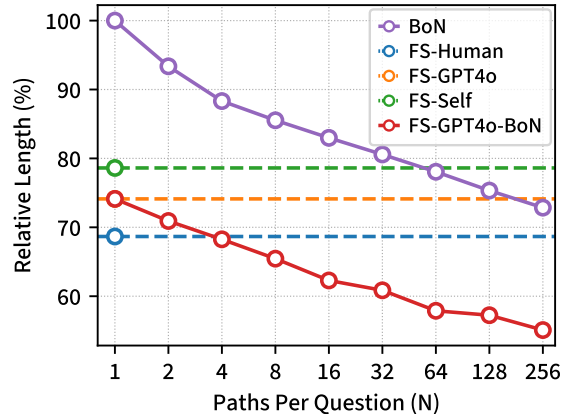


Figure 3: Sample efficiency of generation methods for self-training. Average relative length of reasoning paths collected by each method from 64 random questions in the training set. Relative length here is measured compared to the first path generated for naive BoN (only correct paths). Results are averaged over GSM8K and MATH across 4 models excluding DeepSeekMath-7B.

We find that few-shot prompting reliably elicits concise generations across all models considered in our study. As few-shot learning is a fundamental capability of LLMs (Brown et al., 2020), we expect this to further generalize to other models. Horizontal lines in Figure 3 shows that sampling with 8-shot conditioning reduces the length of reasoning paths significantly. Notably, FS-Human elicits more reduction than that achieved by BoN sampling with $N = 256$.

Few-shot conditioned BoN sampling (FS-BoN)

We can maximize length reduction by jointly applying BoN sampling on top of few-shot conditioning. We consider GPT-4o generated examples for this approach (*FS-GPT4o-BoN*), based on their strong accuracy preservation and length reduction in standalone few-shot conditioned self-training (see *FS-GPT4o* in Table 2). Figure 3 shows that the improvements from few-shot conditioning and BoN sampling are largely independent and additive, resulting in significant length reduction.

Why do we need (self-)training? Directly applying BoN and few-shot prompting at test time, without self-training, incurs prohibitive inference overhead due to repeated sampling and long prompts (Table 5). This defeats our ultimate goal of mitigating the inference costs of reasoning. On the other hand, self-training internalizes the length reduction benefits, allowing the model to reason concisely without additional inference-time overhead.

3.3 Sample Augmentation for Accuracy Boost

While few-shot prompting elicits concise reasoning, its adaptability is limited due to the small number of given examples. It may (1) prohibit generation of correct paths for very complex questions that require longer reasoning paths, while (2) eliciting extraneous steps that are not required for very easy questions. To address this, for each question, we augment the set of $\{1, N\}$ sample(s) generated for FS and FS-BoN, respectively, with N samples generated for naive BoN, and select the shortest correct path from the combined set. We find that this retains the length reduction from FS and FS-BoN while better preserving accuracy, likely due to better coverage of difficult questions. We apply this augmentation by default to few-shot conditioned methods, with its effectiveness ablated in Figure 4.

4 Experimental setup

Models To account for realistic task-specific deployment settings, we select recent moderately sized post-trained models. We also consider math-specialized models to evaluate on models that have been optimized for specific task domains. We consider five main models for our key experiments: Llama-3.2-3B (Dubey et al., 2024), Gemma-2-2B (Team et al., 2024), Qwen2.5-3B (Yang et al., 2024a), Qwen2.5-Math-1.5B (Yang et al., 2024b), and DeepSeekMath-7B (Shao et al., 2024). We investigate scaling on Llama-3.2- $\{1B, 3B\}$ and Llama-3.1-8B.

Tasks and datasets We focus on challenging reasoning tasks where (1) CoT reasoning significantly improves model performance, (2) only the final answer is relevant, and (3) models achieve moderate performance. Reasoning length reduction is desirable under the first and second conditions as it can reduce inference latency without affecting utility. The third condition is necessary to assess accuracy preservation. We consider two mathematical reasoning datasets: GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021), where the models achieve accuracies of 40–90% and 20–70%, respectively. For evaluation, we utilize the test set of the GSM8K and MATH500 dataset (Lightman et al., 2024). We explain details in Section C.1.

Evaluation metrics We evaluate methods using two primary metrics: *accuracy* and *length*. Accuracy is evaluated using Python-based parsing code, described in Section C.2. Length is defined as the average number of output tokens in all reasoning

paths, including incorrect ones, as output tokens incur inference costs in deployment scenarios regardless of their correctness. We focus on number of *output* tokens for clear comparison, as number of input tokens are similar across methods, and output tokens affect wall-clock latency to a higher degree (Agrawal et al., 2024). We further justify this choice in Section C.3. We also employ *relative accuracy* and *relative length* metrics to better evaluate how well each method elicits concise reasoning while maintaining accuracy. Specifically, relative accuracy is the ratio of the given method’s accuracy to the baseline accuracy, and relative length is the ratio of the given method’s average length to the baseline average length, using a strong baseline zero-shot prompt (Pang et al., 2024). We use greedy decoding throughout evaluation to ensure reproducibility².

Baseline methods For baselines, we consider zero-shot prompting (Section 2.2) and fine-tuning directly on ground-truth answers, and with external supervision from human- and GPT-4o-generated concise reasoning paths. For existing fine-tuning methods on concise reasoning, we reproduce Rational Metareasoning (RM) (De Sabbata et al., 2024), which follows a similar approach to our naive BoN method but with two key differences: (1) a reward function that balances efficiency and utility (i.e., output length and accuracy), and (2) iterative training via expert iteration (Zelikman et al., 2022).

Self-training budget allocation The cost of self-training methods can be broken down into (1) generation and (2) fine-tuning. We aim to match the budget across methods in terms of generation (number of paths generated per question) since fine-tuning costs are relatively negligible (see Table 5). For naive BoN, we generate 16 paths per question. For RM (De Sabbata et al., 2024), we generate $N = 4$ paths per iteration across 4 iterations. We generate a single path for standalone few-shot conditioning (FS) and 16 paths for FS-BoN, both augmented with 16 additional paths from the default output distribution (used for naive BoN). Considering the large budget of FS-BoN with augmentation, we evaluate a *Budget-Matched* setting with 8 paths each from the few-shot conditioned distribution and default distribution, as shown in Table 2. Details are provided in Section C.4.

²Note that this contrasts with Section 2 where we took the average length of *correct* samples over *stochastic* generations per question, for precise question-level length normalization.

Dataset	GSM8K				MATH				
	Method	Acc	Len	Rel. Acc	Rel. Len	Acc	Len	Rel. Acc	Rel. Len
<i>Zero-Shot Prompting</i>									
Baseline	78.06	241.87	100.00 (0.00)	100.00 (0.00)	46.40	480.37	100.00 (0.00)	100.00 (0.00)	
Be Concise	77.98	214.87	99.85 (1.18)	88.46 (10.37)	47.76	446.09	102.71 (7.59)	92.66 (7.46)	
Hand Crafted 2 (ours)	76.72	184.13	98.27 (3.67)	77.10 (22.27)	46.84	404.85	101.62 (4.79)	85.26 (15.97)	
<i>FT - External Data</i>									
Direct Answer	19.70	3.17	24.88 (5.03)	1.36 (0.40)	15.08	6.98	35.16 (10.34)	1.44 (0.73)	
Human CoT	65.73	127.85	83.82 (7.28)	54.95 (13.17)	33.88	243.54	75.61 (13.56)	53.14 (13.87)	
GPT4o CoT	76.36	156.24	97.65 (3.63)	67.60 (16.70)	40.44	399.80	90.52 (15.07)	87.21 (22.22)	
<i>FT - Best-of-N Self-Generation</i>									
Naive BoN	77.12	214.22	98.79 (1.64)	87.17 (8.79)	47.64	433.26	101.74 (7.04)	89.89 (3.99)	
Rational Metareasoning	76.15	207.49	97.21 (5.74)	84.93 (5.09)	47.56	432.56	103.02 (6.56)	90.56 (5.25)	
<i>FT - Few-Shot Conditioned Self-Generation (ours)</i>									
FS-Human	76.66	161.72	98.06 (3.28)	67.96 (16.62)	46.44	421.54	99.69 (6.97)	87.78 (5.98)	
FS-GPT4o	78.07	175.54	99.94 (1.69)	73.15 (13.49)	47.36	421.21	101.87 (5.33)	87.58 (6.60)	
FS-Self	77.27	190.03	98.86 (2.51)	77.51 (9.18)	48.00	426.67	102.67 (5.24)	88.50 (4.49)	
<i>FT - Few-Shot Conditioned Best-of-N Self-Generation (ours)</i>									
FS-GPT4o-BoN	75.88	153.38	97.00 (4.11)	64.25 (16.66)	47.36	364.33	102.56 (6.24)	76.30 (10.56)	
└ Budget-Matched	76.24	160.59	97.44 (3.67)	67.15 (16.41)	47.52	384.43	101.58 (9.53)	80.43 (9.04)	

Table 2: Evaluation of zero-shot prompting, fine-tuning (FT), and our proposed methods on GSM8K and MATH datasets, averaged across five main models. We report absolute accuracy (%) and length (tokens), along with relative accuracy (%) and length (%) compared to baseline prompting. Values in parentheses show standard deviations.

5 Results

5.1 Main results

Our main results, presented in Table 2 and Figure 4, demonstrate the performance of our self-training methods against baseline approaches.

Naive BoN fine-tuning is effective but sample inefficient. Naive BoN fine-tuning effectively reduces output length without significantly degrading model performance. This also holds true for Qwen2.5-Math-1.5B and DeepSeekMath-7B (Table 13 and Table 14), which failed to achieve length reduction through zero-shot prompting. However, the length reduction from naive BoN with $N = 16$ is limited to 12% on average. Furthermore, as illustrated in Figure 3, achieving more compression with BoN becomes progressively less efficient.

Iterative baseline yields similar results as naive BoN fine-tuning. Rational Metareasoning, an iterative baseline, yields similar relative length reduction and relative accuracy to BoN fine-tuning. This suggests that the utility reward proposed by De Sabata et al. (2024) may not effectively achieve both accuracy gains and token length reduction.

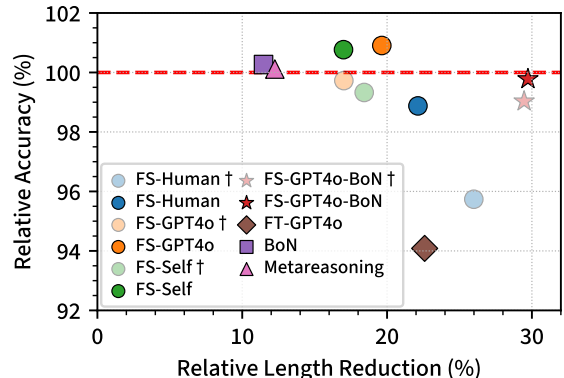


Figure 4: Tradeoff between relative accuracy and length reduction for main methods. Results are averaged over GSM8K and MATH across five main models. Matching colors and shapes indicate the same FS prompt. FS conditioning without augmentation (\dagger) are marked with lighter colors. Relative length reduction refers to $100 - \text{relative length} (\%)$.

Few-shot conditioning outperforms BoN in length reduction. The results demonstrate that few-shot conditioning achieves a greater relative length reduction compared to naive BoN, including math-specialized models (Table 13 and Table 14). This is in line with the superior length reduction of

Question

A robe takes 2 bolts of blue fiber and half that much white fiber. How many bolts in total does it take?

Reasoning Paths

[Zero-Shot] To find the total number of bolts needed, we need to calculate the amount of white fiber first, since it's half the amount of blue fiber. Step 1: Determine the amount of blue fiber needed. The robe takes 2 bolts of blue fiber. Step 2: Calculate the amount of white fiber needed. The robe takes half the amount of white fiber as blue fiber. Half of 2 bolts is $2/2 = 1$ bolt. Step 3: Add the amount of blue fiber and white fiber to find the total. Total bolts = blue fiber + white fiber. Total bolts = 2 bolts + 1 bolt. Total bolts = 3 bolts. The answer is 3.

[FS-GPT4o-BoN] The robe takes 2 bolts of blue fiber. It takes half that much white fiber, which is $2/2 = 1$ bolt. Add the blue and white fiber together: $2 + 1 = 3$ bolts. The answer is 3.

Table 3: Example rationales generated from the original Llama-3.1-8B model (**Zero-Shot**) and the model fine-tuned using our method (**FS-GPT4o-BoN**). The question is randomly sampled from GSM8K. For the same question, one can observe that our rationale is much more concise, removing redundant parts that explain the question.

few-shot conditioning, compared to naive BoN as shown in Figure 3. Notably, self-training on generations conditioned on human-annotated examples (FS-Human) achieves an average relative length of 67.96% on GSM8K, compared to 87.17% with naive BoN.

Self-training better preserves accuracy than training with external data. Table 2 shows fine-tuning with external data (*FT-External Data*) leads to a significant reduction in relative length but causes a severe drop in relative accuracy. Figure 4 further highlights the accuracy preservation of self-training: fine-tuning with external concise reasoning supervision from GPT-4o (FT-GPT4o) lies below the Pareto-curve of relative accuracy and relative length reduction, established by our self-training methods.

Few-shot conditioned BoN achieves best length reduction while maintaining accuracy. FS-BoN elicits the largest length reduction among our self-training methods, while maintaining relative accuracy, on average. Notably, for math-specialized models, FS-GPT4o-BoN achieves the greatest reduction among all methods, except those fine-tuned on external data which greatly sacrifice the accuracy (Table 13 and Table 14). Moreover, robustness testing with different few-shot exemplar sets confirms consistent performance, maintaining accuracy while effectively reducing reasoning length even with varied exemplars (Section D.9).

Augmentation boosts accuracy for few-shot conditioning. Figure 4 compares few-shot conditioning, i.e., FS and FS-BoN, with and without augmentation (\dagger). Augmentation improves accuracy by providing solutions for previously unsolvable hard

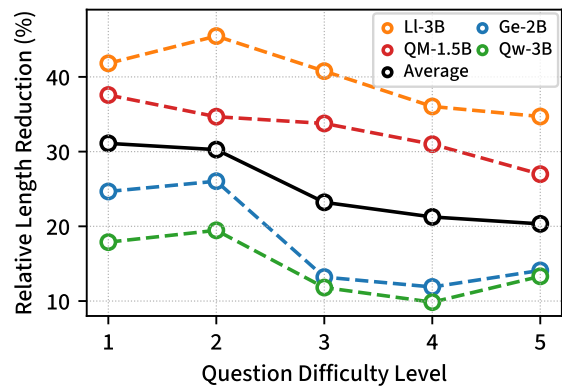


Figure 5: **Tokens are reduced adaptively according to question difficulty.** Token reduction rate for each difficulty level on MATH, for 4 models individually and averaged. Relative length reduction refers to 100 - relative length (%).

questions as discussed in Section 3.3. While augmentation may slightly affect reduction rates, they remain superior to naive BoN and RM. Even when matching the budget (*Budget-Matched*) with other self-training methods in Table 2, it achieves the greatest length reduction among them with minimal accuracy degradation. The effect of augmentation on training data length is analyzed in Section D.5.

Cross-task evaluation demonstrates broader applicability. We conducted cross-task evaluation by testing models fine-tuned on one dataset against another. Our method shows strong generalization, achieving 10-12% length reductions with minimal accuracy drops when applied to out-of-domain mathematical problems (detailed analysis in Section D.7).

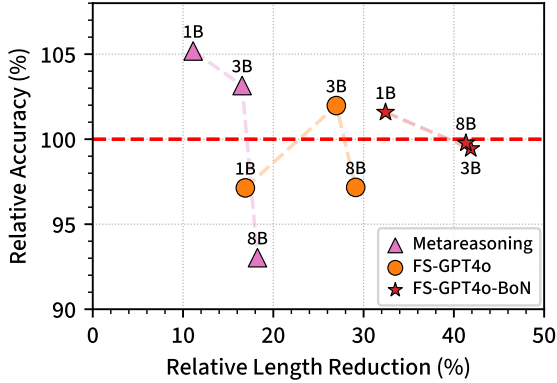


Figure 6: Scaling study on baseline and few-shot conditioned self-training methods. Results are averaged over GSM8K and MATH for Llama 1B, 3B, and 8B. Relative length reduction refers to $100 - \text{relative length}$ (%).

Broader domain evaluation demonstrates consistent effectiveness. We evaluate our method on diverse reasoning domains (business, chemistry, physics) from MMLU-Pro (Wang et al., 2024a) that benefit from chain-of-thought reasoning (Sprague et al., 2024) in Appendix D.8. Our method demonstrates consistent effectiveness across all three domains, achieving both accuracy improvements (16.51% average) and substantial length reduction (26.82% average). Notably, our method provides over 3x better length reduction efficiency compared to naive BoN while maintaining competitive accuracy gains.

5.2 Analysis

In this section, we analyze the length reduction effects in depth. We exclude DeepSeekMath-7B from quantitative analysis due to cost.

Tokens are reduced adaptively according to question complexity. The MATH dataset’s difficulty levels range from 1 (basic algebra) to 5 (advanced calculus and complex mathematical reasoning). As shown in Figure 5, our method adaptively reduces tokens based on question difficulty, with higher difficulty leading to less reduction. The higher reduction (20%–40%) at easier difficulty levels (1–2) suggests that the original model outputs for these easier questions contained unnecessary tokens. This reveals a gap in current models’ ability to tailor their inference budget to problem complexity. Our method effectively closes this gap by reducing redundancy, allowing for more precise token allocation based on question difficulty.

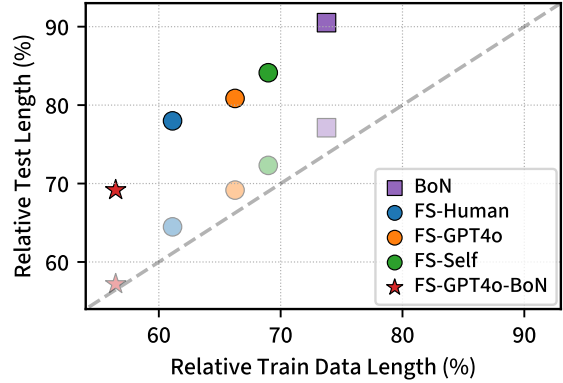


Figure 7: **Fine-tuning effectively transfers the length reduction to the model.** Correlation between the relative length of train data and test output averaged over GSM8K and MATH across 4 models. Training length includes only correct solutions. Solid points represent test lengths including all generated outputs, while lighter points indicate test lengths of correct solutions only.

Self-training maintains consistency across model scales. We conduct a scaling study on Llama-3.2-1B, 3B, and Llama-3.1-8B to examine consistency across different model sizes (Figure 6). Overall, token reduction increases as the model size increases, while the maintenance of accuracy does not show a strong correlation with model size. RM exhibits lower reduction rates compared to our few-shot conditioned self-training methods across all models and shows a decrease in accuracy with increasing model size. Our standalone few-shot conditioning method (FS-GPT4o) also shows a similar trend in length reduction, but better preserves accuracy. Our joint FS-GPT4o-BoN method achieves the greatest reduction across all models, maintaining relative accuracy across different model sizes, especially in the largest 8B model.

Sample study Table 3 presents qualitative examples of reasoning paths generated by the model before and after fine-tuning with our method. The original reasoning exhibits verbosity, containing redundant processes such as question confirmation and repeated instructions. In contrast, the reasoning generated by our method includes only the necessary steps, significantly reducing the number of tokens while still arriving at the correct answer. More examples are provided in the Appendix E.

Length reduction is transferred through fine-tuning. As shown in Figure 7, fine-tuning with shorter rationales results in shorter model outputs, showing a strong correlation between test and train-

ing lengths. We note that the length of test outputs (incorrect and correct) are longer than the length of training samples (only correct) on average. This is mainly because incorrect paths are generally longer than correct ones. We find a closer correspondence between train length and test length of correct samples only, indicated by the lighter datapoints. This discrepancy suggests the need to terminate incorrect paths early to minimize redundant inference overhead. We consider this for future work.

Practical efficiency gains We also conducted wall-clock time and memory usage measurements after fine-tuning models with our method. The results show that latency is reduced by 15.38% to 52.94%, while memory usage decreases by 2.50% to 6.26% across two different models (detailed analysis provided in [Section D.6](#)). These improvements directly translate to lower computational costs and faster inference in real-world deployments.

6 Discussion

Default reasoning behavior of LLMs Previous research has shown that the CoT reasoning ability of LLMs originates from procedural knowledge in pretraining data ([Ruis et al., 2024](#)). Modern LLM training pipelines utilize high-quality math, code, and synthetic reasoning data to enhance reasoning, but these do not promote conciseness ([Dubey et al., 2024](#); [Yang et al., 2024a](#)). Furthermore, recent ‘thinking’ models are reinforced to use additional tokens to improve reasoning performance, rather than save on token budget ([OpenAI, 2024](#)). Thus, current LLMs naturally exhibit redundant reasoning. We believe that incorporating concise reasoning supervision or rewards in training pipelines can be beneficial for model efficiency, especially for ‘thinking’ models with lengthy internal reasoning.

Lightweight fine-tuning for concise reasoning elicitation In this paper, we focused on standard fine-tuning based on self-generated samples. Our analysis shows that LLMs *already* possess the ability to correctly reason in a relatively concise manner. Therefore, we posit that lightweight fine-tuning is sufficient to achieve significant reasoning length reduction, echoing the Superficial Alignment Hypothesis in the field of LLM alignment ([Zhou et al., 2023](#)). Indeed, our post-hoc analysis in [Figure 7](#) shows that standard fine-tuning can reduce the output length of models in proportion to that of training samples.

7 Conclusion

We tackle redundancy in CoT reasoning by hypothesizing that LLMs possess a latent capacity for concise reasoning, as evidenced by shorter correct reasoning paths within their output distributions. We introduce fine-tuning methods that leverage self-generated data from best-of-N sampling and few-shot conditioning to elicit this capacity. Our FS-BoN method significantly reduces reasoning length by 30% while maintaining accuracy. This implies fine-tuning with curated self-generated data can reliably unlock latent concise reasoning, enabling more efficient inference.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-00075, Artificial Intelligence Graduate School Program (KAIST) and No. RS-2024-00457882, AI Research Hub Project) and Artificial intelligence industrial convergence cluster development project funded by the Ministry of Science and ICT (MSIT, Korea) & Gwangju Metropolitan City.

We would like to express our gratitude to Taehyeon Kim, Murad Aghazada, Sunwoo Hong, and Minu Kim for their valuable feedback and assistance throughout this research project. Their insightful comments and suggestions significantly contributed to the quality of this work.

Limitations and Future Work

Advanced Training Schemes

While our focus was on standard fine-tuning with self-generated samples, exploring advanced reinforcement learning (RL) based training schemes could potentially maximize efficiency further. Our preliminary experiments with expert iteration and BoN sampling (without a utility reward) showed promising results, boosting both accuracy and length reduction compared to non-iterative BoN. Further investigation of iterative or more advanced RL methods is warranted.

Few-Shot Prompting Exploration

Although we leveraged few-shot prompting to bootstrap length reduction during data generation, we did not explore advanced few-shot prompting methods as a middle ground between zero-shot and fine-tuning. While direct few-shot prompting with our prompts achieved comparable length reduction to fine-tuning on single-path few-shot generations, it incurred a slight accuracy loss. Critically, fine-tuning enabled us to incorporate BoN sampling and sample augmentation, leading to significantly greater length reduction without sacrificing accuracy. Moreover, our robustness analysis across different exemplar sets confirms consistent performance. However, future work on integrating advanced exemplar selection (Fu et al., 2022) and many-shot prompting (Agarwal et al., 2024) could further enhance our fine-tuning approach.

Extended Scaling Studies

Our scaling study was limited to Llama 3.x models at 1B, 3B, and 8B parameters. While these model sizes are relevant for our task-specific setting, further empirical studies are needed to evaluate the effectiveness of our method on models exceeding 8B parameters. Investigating scaling trends across a wider range of model sizes is crucial for understanding the full potential of our approach.

Concise Reasoning in General LLMs

This study focused on task-specific fine-tuning. While existing zero-shot and fine-tuning methods often struggle with reliability and efficacy even in such settings, and our method proved effective, generalizing our approach to a broader range of tasks is an important direction. Exploring techniques like multi-task training could enable efficient reasoning

without task-specific tuning. This could be particularly beneficial for ‘thinking’ models (Guo et al., 2025) where reasoning enhances the final response, similar to our math reasoning tasks but in a more general context. Furthermore, while prior work has explored reducing the number of reasoning stages (Chen et al., 2024), our method focuses on reducing verbal redundancy within sentences, offering a complementary approach.

References

- Rishabh Agarwal, Avi Singh, Lei Zhang, Bernd Bohnet, Luis Rosias, Stephanie C. Y. Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, John D. Co-Reyes, Eric Chu, Feryal M. P. Behbahani, Aleksandra Faust, and Hugo Larochelle. 2024. [Many-shot in-context learning](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Amey Agrawal, Nitin Kedia, Ashish Panwar, Jayashree Mohan, Nipun Kwatra, Bhargav S. Gulavani, Alexey Tumanov, and Ramachandran Ramjee. 2024. [Taming throughput-latency tradeoff in LLM inference with sarathi-serve](#). In *18th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2024, Santa Clara, CA, USA, July 10-12, 2024*, pages 117–134. USENIX Association.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. 2024. [Do NOT think that much for 2+3=? on the overthinking of o1-like llms](#). *CoRR*, abs/2412.21187.
- Cheng-Han Chiang and Hung-yi Lee. 2024. [Over-reasoning and redundant calculation of large language models](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 2: Short Papers, St. Julian's, Malta, March 17-22, 2024*, pages 161–169. Association for Computational Linguistics.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- C Nicolò De Sabbata, Theodore R Sumers, and Thomas L Griffiths. 2024. Rational metareasoning for large language models. *arXiv preprint arXiv:2410.05563*.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2022. Complexity-based prompting for multi-step reasoning. In *The Eleventh International Conference on Learning Representations*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Tingxu Han, Chunrong Fang, Shiyu Zhao, Shiqing Ma, Zhenyu Chen, and Zhenting Wang. 2024. Token-budget-aware llm reasoning. *arXiv preprint arXiv:2412.18547*.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. [Measuring mathematical problem solving with the MATH dataset](#). In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.
- Namgyu Ho, Sangmin Bae, Taehyeon Kim, Hyunjik Jo, Yireun Kim, Tal Schuster, Adam Fisch, James Thorne, and Se-Young Yun. 2024. [Block transformer: Global-to-local language modeling for fast inference](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.
- Mingyu Jin, Qinkai Yu, Dong Shu, Haiyan Zhao, Wenyue Hua, Yanda Meng, Yongfeng Zhang, and Mengnan Du. 2024. [The impact of reasoning step length on large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1830–1842, Bangkok, Thailand. Association for Computational Linguistics.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. [Large language models are zero-shot reasoners](#). 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*.
- Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, et al. 2024. Training language models to self-correct via reinforcement learning. *arXiv preprint arXiv:2409.12917*.
- Xin Lai, Zhuotao Tian, Yukang Chen, Senqiao Yang, Xiangu Peng, and Jiaya Jia. 2024. Step-dpo: Step-wise

- preference optimization for long-chain reasoning of llms. *arXiv preprint arXiv:2406.18629*.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. [Let's verify step by step](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Haotian Luo, Li Shen, Haiying He, Yibo Wang, Shiwei Liu, Wei Li, Naiqiang Tan, Xiaochun Cao, and Dacheng Tao. 2025. O1-pruner: Length-harmonizing fine-tuning for o1-like reasoning pruning. *arXiv preprint arXiv:2501.12570*.
- William Merrill and Ashish Sabharwal. 2023. The expressive power of transformers with chain of thought. *arXiv preprint arXiv:2310.07923*.
- Sania Nayab, Giulio Rossolini, Marco Simoni, Andrea Saracino, Giorgio Buttazzo, Nicolamaria Manes, and Fabrizio Giacometti. 2024. Concise thoughts: Impact of output length on llm reasoning and cost. *arXiv preprint arXiv:2407.19825*.
- Franz Nowak, Anej Svete, Alexandra Butoi, and Ryan Cotterell. 2024. [On the representational capacity of neural language models with chain-of-thought reasoning](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 12510–12548. Association for Computational Linguistics.
- Maxwell I. Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. 2021. [Show your work: Scratchpads for intermediate computation with language models](#). *CoRR*, abs/2112.00114.
- OpenAI. 2024. [Openai o1 system card](#). This report outlines the safety work carried out prior to releasing OpenAI o1-preview and o1-mini, including external red teaming and frontier risk evaluations according to our Preparedness Framework.
- Richard Yuanzhe Pang, Weizhe Yuan, He He, Kyunghyun Cho, Sainbayar Sukhbaatar, and Jason Weston. 2024. [Iterative reasoning preference optimization](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Xuan Ren, Biao Wu, and Lingqiao Liu. 2024. [I learn better if you speak my language: Understanding the superior performance of fine-tuning large language models with llm-generated responses](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 10225–10245. Association for Computational Linguistics.
- Matthew Renze and Erhan Guven. 2024. [The benefits of a concise chain of thought on problem-solving in large language models](#). *CoRR*, abs/2401.05618.
- Laura Ruis, Maximilian Mozes, Juhan Bae, Siddhartha Rao Kamalakara, Dwarak Talupuru, Acyr Locatelli, Robert Kirk, Tim Rocktäschel, Edward Grefenstette, and Max Bartolo. 2024. Procedural knowledge in pretraining drives reasoning in large language models. *arXiv preprint arXiv:2411.12580*.
- David W Scott. 2015. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2024. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. *arXiv preprint arXiv:2409.12183*.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.
- Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. 2025. Kimi k1.5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. [Self-consistency improves chain of thought reasoning in language models](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, et al. 2024a. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Ze Zhong Wang, Xingshan Zeng, Weiwen Liu, Yufei Wang, Liangyou Li, Yasheng Wang, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2024b. Chain-of-probe: Examining the necessity and accuracy of cot step-by-step. *arXiv preprint arXiv:2406.16144*.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama,

- Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. [Emergent abilities of large language models](#). *Trans. Mach. Learn. Res.*, 2022.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022b. [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*.
- Heming Xia, Yongqi Li, Chak Tou Leong, Wenjie Wang, and Wenjie Li. 2025. Tokenskip: Controllable chain-of-thought compression in llms. *arXiv preprint arXiv:2502.12067*.
- Violet Xiang, Charlie Snell, Kanishk Gandhi, Alon Albalak, Anikait Singh, Chase Blagden, Duy Phung, Rafael Rafailov, Nathan Lile, Dakota Mahan, et al. 2025. Towards system 2 reasoning in llms: Learning how to think with meta chain-of-thought. *arXiv preprint arXiv:2501.04682*.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024a. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.
- An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. 2024b. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.
- Tian Ye, Zicheng Xu, Yuanzhi Li, and Zeyuan Allen-Zhu. 2024. Physics of language models: Part 2.2, how to learn from mistakes on grade-school math problems. *arXiv preprint arXiv:2408.16293*.
- Edward Yeo, Yuxuan Tong, Morry Niu, Graham Neubig, and Xiang Yue. 2025. Demystifying long chain-of-thought reasoning in llms. *arXiv preprint arXiv:2502.03373*.
- Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. [Star: Bootstrapping reasoning with reasoning](#). 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*.
- Yusen Zhang, Sarkar Snigdha Sarathi Das, and Rui Zhang. 2024. Demystify verbosity compensation behavior of large language models. *arXiv preprint arXiv:2411.07858*.
- Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. [LIMA: less is more for alignment](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

A Related work

A.1 Chain-of-thought Reasoning

CoT reasoning enables LLMs to solve diverse tasks (Wei et al., 2022b; Kojima et al., 2022) by leveraging an arbitrarily large amount of computation for next-word prediction (Merrill and Sabharwal, 2023; Nowak et al., 2024; Xiang et al., 2025). Recently, additional lines of research have expanded (OpenAI, 2024; Guo et al., 2025) basic CoT by adding in-context search (Ye et al., 2024; Kumar et al., 2024), often termed as meta-CoT (Xiang et al., 2025) or long-CoT (Yeo et al., 2025). This paper focuses on original CoT framework to investigate the simplest ways to reduce the number of tokens, which could potentially be applied to long-CoT as well.

A.2 Concise Reasoning

Many studies have shown that LLMs tend to overthink during CoT, often including unnecessary words in their reasoning (Zhang et al., 2024; Chiang and Lee, 2024; Wang et al., 2024b), which can negatively impact efficiency. However, simply reducing the number of tokens often leads to significantly lower performance (Merrill and Sabharwal, 2023; Jin et al., 2024). To address this, some works have explored prompting techniques to elicit concise reasoning (Renze and Guven, 2024; Nayab et al., 2024; Han et al., 2024), but these methods are highly model-dependent. Another line of work has focused on fine-tuning methods to balance efficiency and reasoning quality (De Sabbata et al., 2024), but these approaches are insufficient in reducing token usage. In this work, we investigate how to efficiently generate concise reasoning samples and use them for fine-tuning, improving model efficiency without sacrificing accuracy.

Several concurrent works also address redundancies in LLM outputs. Some approaches (Chen et al., 2024; Luo et al., 2025; Team et al., 2025) focus on how o1-like models overthink and propose methods for pruning extraneous reasoning steps. Another approach, TokenSkip (Xia et al., 2025), first identifies and discards low-importance tokens from self-generated CoT outputs, then fine-tunes the model on these compressed examples to enable efficient inference. In contrast, our method elicits concise reasoning more simply yet effectively by training on a self-generated dataset of shorter rationales, using best-of-N sampling and few-shot conditioning.

B Methodological Details

B.1 Self-Generation of Few-Shot Exemplars (FS-Self)

We devise a systematic approach to obtain few-shot exemplars using the default prompt (Appendix F.1) of concise reasoning from the target model itself, motivated by BoN sampling, through a two-phase process: **sampling** and **selection**.

Sampling Phase In the sampling phase, 128 questions were randomly selected from the training set, and for each selected question, 128 diverse reasoning paths were generated with a sampling temperature of $T = 0.7$. This process created a diverse set of candidate paths for both the GSM8K and MATH datasets.

Selection Phase The selection process was executed through the following steps:

1. Sorting candidate paths by token count in ascending order.
2. Filtering only correct samples by comparing oracle labels with the models' answers using the parsing code (Section C.2).
3. Sequential validation using GPT-4o with the following specialized prompt was performed until eight valid and concise examples were obtained. To ensure question uniqueness, alternative solutions from previously selected questions were excluded.

You are a math reasoning validator. Given a math problem, its answer, and a proposed rationale: First, determine if the problem requires step-by-step reasoning to solve.

Then, validate if the rationale contains explicit step-by-step reasoning that correctly solves the problem.

Question: [question here]

Expected Answer: [label here]

Proposed Rationale: [rationale here]

Respond with ONLY ACCEPT or REJECT:

ACCEPT - if the problem requires step-by-step reasoning AND the rationale contains explicit step-by-step reasoning that correctly leads to the answer.

REJECT - if the problem is too simple (can be solved in one step).

REJECT - if the rationale lacks explicit reasoning steps or contains incorrect logic.

REJECT - if the rationale only states the answer.

Dataset-Specific Selection Criteria For GSM8K, eight valid examples were selected using the above selection phase. An example is shown in Table 22. For MATH, to avoid selecting only very easy questions, the selection was distributed across different categories: two examples from algebra (the largest category) and one example from each of the other categories (geometry, intermediate algebra, prealgebra, number theory, counting and probability, and precalculus). An example is shown in Table 23.

B.2 Generation of Few-Shot Exemplars From GPT-4o (FS-GPT4o)

To obtain few-shot exemplars of concise reasoning from GPT-4o, we use our zero-shot prompt **Hand Crafted 3** (Appendix F.1) to generate short reasoning paths. For GSM8K, we randomly sample eight questions from the training set. For MATH, we randomly select two questions from the largest category, Algebra and one from each of the other categories (Geometry, Intermediate Algebra, Pre-algebra, Number Theory, Counting & Probability, Precalculus), all from the training set. We manually filter low-quality outputs to ensure the quality of few-shot examples. However, we found that GPT-4o reliably followed the zero-shot prompt in the majority of cases. Generated examples of GPT-4o are in Tables 20 and 21.

C Experimental Details

C.1 Datasets

A summary of the datasets used in our experiments, along with their original licenses, is provided in Table 4. Both datasets are in English and focus on mathematical reasoning tasks. GSM8K contains grade school math word problems requiring multi-step reasoning, while MATH covers more advanced problems across various categories (algebra, geometry, precalculus, etc.) with different difficulty levels. Both datasets use straightforward language and standard mathematical notation. The original train/test splits are used for fine-tuning and evaluation. Results for MATH are reported on the MATH-500 subset, following the previous work (Lightman et al., 2024).

Dataset	Train	Test	License	Reference
GSM8K	7,473	1,319	MIT	Cobbe et al. (2021)
MATH	7,500	500	MIT	Hendrycks et al. (2021)

Table 4: Dataset statistics for GSM8K and MATH.

C.2 Answer Parsing

For accuracy, we extract and verify the final numerical answer using Python-based parsing code. We build previous implementations for GSM8K (Kojima et al., 2022) and MATH (Lai et al., 2024), with additional rules to ensure consistent parsing across all models. Our implementation additionally handles mathematical expressions by standardizing numerical formats, removing units and mathematical notation, and cleaning model-specific formatting to ensure consistent evaluation across models.

During our experiments, we observed that different models require different parsing approaches to extract answers correctly for the MATH dataset. We implemented a model-specific parsing strategy where Gemma-2-2b and Qwen2.5-3B models utilize an enhanced parser, while all other models (including Llama, DeepSeekMath, and Qwen2.5-Math) use our standard parser. This adjustment led to more accurate model performance evaluation. All experiments use the appropriate parser for each model to ensure consistent evaluation.

C.3 Justification For Length Metric

We define our length metric for model evaluation as the average number of *output* tokens in model responses across the entire evaluation set. For clear comparison, we do not include input tokens because (1) the number of input tokens do not differ significantly between zero-shot prompting and direct prompting on fine-tuned models, but more importantly (2) the real-world wall-clock latency incurred by each output token is significantly greater than that of each input token. The greater latency of output tokens stems from the sequential nature of their processing (the decode stage). Input tokens, in contrast, are processed in parallel (the pre-fill stage). The decode stage is heavily memory-bound, requiring parameter fetching for every forward pass, which significantly limits compute utilization (Agrawal et al., 2024). This difference in processing is also reflected in the pricing structures of commercial LLM API providers, where input tokens are cheaper than output tokens³.

³<https://openai.com/api/pricing/>

C.4 Generation and Fine-Tuning

We implemented all model fine-tuning using the HuggingFace Trainer library and conducted generation using vLLM for optimized inference. For evaluation, we use greedy decoding, while for distribution analysis and training data generation we use temperature sampling with $T=0.7$, following Cobbe et al. (2021); Wang et al. (2023). We generate up to 512 output tokens on the GSM8K dataset and up to 1024 output tokens on the MATH dataset, following the token limits established in prior works (Dubey et al., 2024; Ren et al., 2024). For fine-tuning, we use a batch size of 16 and train for one epoch with a learning rate of $1e-5$. The fine-tuning phase requires at maximum 469 training steps, making it computationally modest compared to the generation phase, as demonstrated in Table 5. For comparison with the iterative fine-tuning baseline MR (De Sabbata et al., 2024), we limit to 4 iterations as the authors observe performance generally plateaus after iteration 4. All experiments are conducted using bfloat16 precision for both generation and fine-tuning. We conduct our experiments on 8 NVIDIA H100 GPUs. Our total usage for main experiments is approximately 1,000 H100 GPU hours.

D Additional results

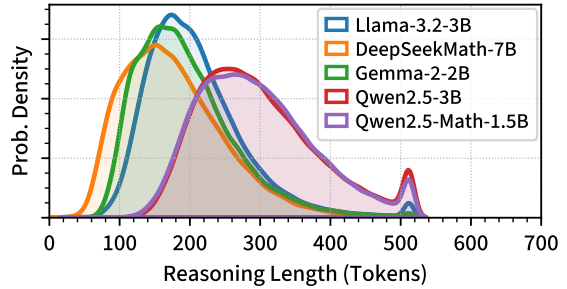
D.1 Full Analysis on Reasoning Length Distribution (Section 2.1)

We analyze the distribution of output lengths across different model families using three metrics. First, we examine the absolute output token count distribution, which directly reflects the computational cost during inference. Second, we analyze the raw character count distribution to verify our findings are consistent across different tokenization schemes. Third, we study the normalized output token count, where each solution’s output length is divided by the mean output length of correct solutions for that specific question, allowing us to identify relatively concise solutions.

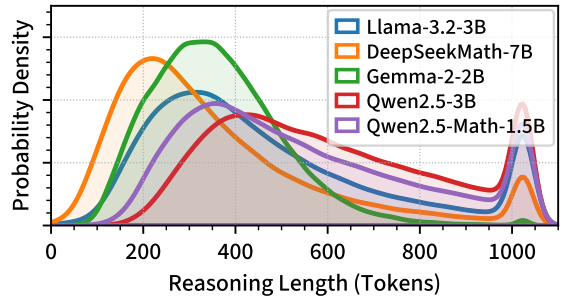
Figure 8 reveals significant variations in how different model families approach mathematical reasoning. On GSM8K, DeepSeekMath-7B demonstrates notably concise reasoning with a mean output length of 179 tokens, while Qwen2.5-3B uses substantially more tokens, averaging 298 tokens per solution. This pattern persists in the MATH dataset, though with generally longer output lengths due to the increased complexity of the problems.

Approach	Time
Default zero-shot generation (16 paths)	1h 0m 38s
GPT4o few-shot generation (16 paths)	1h 29m 54s
Model training	2m 24s

Table 5: Generation and training time comparison on a single H100 GPU for GSM8K using Llama-3.2-3B.



(a) GSM8K



(b) MATH

Figure 8: Distribution of **absolute output token counts** across model families on GSM8K and MATH datasets. The distributions show significant variation in output token usage patterns between different models, with some models consistently using more tokens than others for correct solutions.

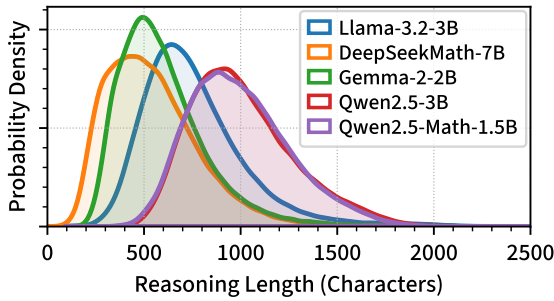
The presence of a small peak at 512 tokens in some distributions is an artifact of the maximum output length limit used during generation, following standard practices in the field.

Figure 9 confirms that our observations about output token usage patterns are not merely artifacts of different tokenization schemes. The relative patterns between models remain consistent, suggesting that the observed differences in output length are fundamental to how each model approaches problem-solving, rather than being an artifact of tokenization choices.

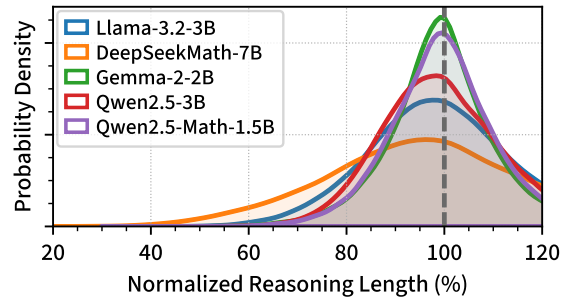
The character-level distributions also provide additional insight into the actual verbosity of solutions, as they measure the raw text length without the influence of tokenization strategies.

Method	Avg. Relative		Ll-3B		Ge-2B		Qw-3B		QM-1.5B		DM-7B	
	Acc	Len	Acc	Len	Acc	Len	Acc	Len	Acc	Len	Acc	Len
Default (Pang et al., 2024)	100.0	100.0	44.4	511.8	24.0	376.0	56.8	605.5	69.8	551.4	37.0	357.3
Be Concise (Renze and Guven, 2024)	102.7	92.7	47.2	477.5	22.2	307.1	60.2	541.9	68.0	550.2	41.2	353.8
Est. Budget (Han et al., 2024)	80.7	132.7	45.6	907.4	16.2	255.1	20.4	125.7	70.0	1070.0	36.0	726.3
Fixed Budget (Nayab et al., 2024)	88.1	70.8	42.6	413.1	22.0	193.6	33.0	170.3	69.6	536.7	35.2	344.8
Hand Crafted 1 (ours)	100.5	86.6	46.8	481.1	20.2	241.6	56.8	446.3	68.0	548.8	42.8	362.3
Hand Crafted 2 (ours)	101.6	85.3	46.6	445.9	24.0	275.7	54.8	393.8	68.8	543.7	40.0	365.2
Hand Crafted 3 (ours)	84.9	80.1	24.4	283.0	22.0	365.7	46.0	272.9	67.6	552.3	37.0	367.7
Hand Crafted 4 (ours)	63.9	45.9	13.2	65.7	13.0	88.7	16.2	9.4	69.4	535.5	39.8	337.9

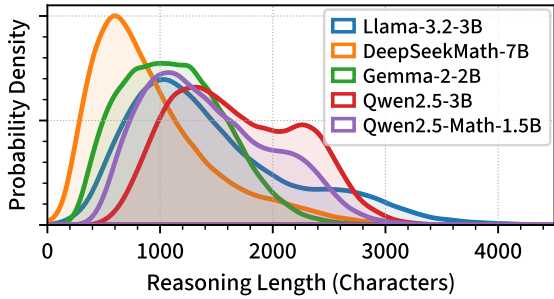
Table 6: **Inconsistency of zero-shot prompting methods on MATH.** Zero-shot prompting methods for concise reasoning on MATH. Relative accuracy (%) and length (%) compared to default prompting are shown, averaged over five main models: Llama-3.2-3B, Gemma-2-2B, Qwen2.5-3B, Qwen2.5-Math-1.5B, and DeepSeekMath-7B. Absolute accuracy (%) and length (tokens) are reported for each individual model. Values in red indicate **significant relative accuracy drop** greater than 5%. Values in gray indicate **negligible relative length reduction** less than 5%.



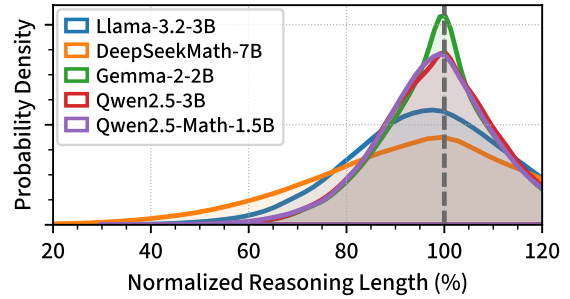
(a) GSM8K



(a) GSM8K



(b) MATH



(b) MATH

Figure 9: Distribution of **character counts** across model families. The character-level analysis confirms that the observed patterns in output token usage are consistent when measured in raw UTF-8 characters, suggesting that the findings are not artifacts of different tokenization schemes.

Figure 10: Distribution of **normalized output token counts**, where each solution’s output length is divided by the mean output length of correct solutions for that question. The distributions reveal that models can produce valid solutions using significantly fewer tokens than their typical output length, though this capability varies across model families.

Figure 10 shows a particularly interesting phenomenon. All models demonstrate the ability to solve problems with significantly fewer output tokens than their typical output length, though this capability varies across model families. The nor-

malization process, which divides each solution’s output length by the mean output length of correct solutions for that specific question, reveals that some models are more flexible in producing concise solutions than others.

Dataset (# samples)	Model	Base (sec)	Ours (sec)	↓%
GSM8K (1319)	LI-3B	16	10	37.50
GSM8K (1319)	QM-1.5B	17	8	52.94
MATH (500)	LI-3B	17	13	23.53
MATH (500)	QM-1.5B	13	11	15.38

Table 7: Wall-clock time comparison between baseline zero-shot prompting and our method (FS-GPT4o-BoN) using vLLM on a single H100 GPU for two models: the general-purpose Llama-3.2-3B and the task-specific Qwen2.5-Math-1.5B.

The presence of significant probability mass below 100% in the normalized distributions highlights an intriguing aspect of model behavior: while models may tend toward verbose outputs, they are not constrained to this verbosity and can successfully solve problems with more compact reasoning in some cases.

D.2 Full Analysis on Efficacy of Zero-Shot Prompting Methods (Section 2.2)

Table 6 shows evaluation results for zero-shot prompting methods on MATH. Prompting methods exhibit similar limitations as those on the GSM8K dataset.

D.3 Impact of Question-Wise Selection

We compare the performance of FS-GPT4o-BoN when selecting the shortest samples per question versus selecting the shortest samples across all questions. Both methods are tested without augmentation. The average relative accuracy across all models and datasets is 99.03% with question-wise selection and 96.03% without it, while the average relative length is 70.54% and 64.33%, respectively. Although selecting the shortest samples across all questions reduces relative length, it comes at the cost of accuracy. Notably, the accuracy of Gemma-2-2B drops significantly from 90.53% with question-wise selection to 75.81% without it. Therefore, we adopt question-wise selection to maintain question diversity and accuracy.

D.4 Full Result Tables on Main Methods

We present full results on the accuracy and reasoning length elicited by main methods for GSM8K in Table 13 and MATH in Table 14. We show absolute accuracy and output token length for individual models and show average over relative accuracy and relative length for aggregates across models.

Dataset (# samples)	Model	Base (MB)	Ours (MB)	↓%
GSM8K (1319)	LI-3B	17873.76	16756.52	6.25
GSM8K (1319)	QM-1.5B	6250.06	6044.87	3.28
MATH (500)	LI-3B	34739.72	33871.16	2.50
MATH (500)	QM-1.5B	13019.13	12342.20	5.20

Table 8: Peak memory usage comparison between baseline zero-shot prompting and our method (FS-GPT4o-BoN) using HuggingFace with batch size 128 on a single H100 GPU for two models: the general-purpose Llama-3.2-3B and the task-specific Qwen2.5-Math-1.5B.

D.5 Impact of Augmentation on Training Data Length

Table 12 demonstrates the impact of augmentation on training data length. Augmentation effectively expands the dataset by leveraging samples from zero-shot BoN. This is achieved by incorporating previously unsolvable long solutions, while also replacing solutions that can be solved with shorter solutions with zero-shot prompting. Consequently, the average solution length remains relatively stable, while the number of correct samples for fine-tuning increases, leading to improved model accuracy.

D.6 Efficiency Analysis: Wall-Clock Time and Memory Usage

We conduct additional efficiency measurements to quantify the practical benefits of our method (FS-GPT4o-BoN) beyond token count reduction.

Wall-Clock Time Savings We measured wall-clock generation times using vLLM, a highly optimized implementation, on a single H100 GPU. Our results confirm that models trained with our method (FS-GPT4o-BoN) achieve significant wall-clock time reductions ranging from 15.38% to 52.94% compared to the original models using baseline zero-shot prompting. The results are presented in Table 7.

Memory Usage Reduction We also measured memory usage to assess efficiency gains. Since vLLM automatically utilizes all available memory, we compared memory usage under fixed batch-size decoding using HuggingFace, a popular LLM framework. Running inference with a batch size of 128 on a single H100 GPU, we found that peak memory usage (measured via `torch.cuda.max_memory_allocated`) was reduced by 2.50% to 6.25% after fine-tuning the model with our method. These results are shown in Table 8.

Train	Test	Rel Acc (%)	Rel Len (%)
GSM8K	In-Domain (GSM8K)	97.00	64.25
	Out-of-Domain (MATH)	99.21	88.34
MATH	In-Domain (MATH)	102.56	76.30
	Out-of-Domain (GSM8K)	98.53	90.36

Table 9: Cross-task evaluation results using FS-GPT4o-BoN method. Models are fine-tuned on one dataset and evaluated on another, averaged across five main models. Relative metrics are computed against the baseline performance on each respective evaluation dataset.

D.7 Cross-Task Generalization Analysis

We conducted cross-task evaluation to assess how well our method generalizes beyond task-specific fine-tuning. We fine-tuned models on one dataset and evaluated them on another across all five main models (Llama-3.2-3B, Gemma-2-2B, Qwen2.5-3B, Qwen2.5-Math-1.5B, DeepSeekMath-7B).

Our results show that reasoning length is reduced even under out-of-domain (OOD) evaluation, with only minor drops in accuracy as shown in Table 9. For example, when training on GSM8K and testing on MATH, we observed a 0.79% relative drop in accuracy alongside an 11.66% reduction in output length. Conversely, when training on MATH and testing on GSM8K, we found a 1.47% relative drop in accuracy with a 9.64% reduction in output length. These results suggest that our method generalizes reasonably well across different types of math problems.

D.8 Evaluation of Broader Reasoning Domains

To demonstrate that our method is applicable beyond mathematical domains, we conducted additional experiments on diverse subject areas that benefit from chain-of-thought reasoning. We evaluated our method on the business, chemistry, and physics categories from MMLU-Pro (Wang et al., 2024a), which have been shown to benefit significantly from CoT reasoning compared to direct answering (Sprague et al., 2024).

We trained and evaluated three general-purpose models (Llama-3.2-3B, Qwen2.5-3B, Gemma-2-2B) on these MMLU-Pro categories using a random 7:3 train-test split. We compared our FS-GPT4o-BoN method against naive BoN and baseline zero-shot prompting, following the same methodology described in Section 3.

Our results show that reasoning length is reduced while accuracy is increased compared to

Domain	FS-GPT4o-BoN		Naive BoN	
	Rel Acc (%)	Rel Len (%)	Rel Acc (%)	Rel Len (%)
Business	102.78	67.38	113.03	92.48
Chemistry	116.33	78.25	119.48	92.99
Physics	130.43	73.90	124.79	91.02
Average	116.51	73.18	119.10	92.16

Table 10: Evaluation results on broader reasoning domains from MMLU-Pro. Results are averaged across three general-purpose models (Llama-3.2-3B, Qwen2.5-3B, Gemma-2-2B) using a random 7:3 train-test split. Relative metrics are computed against baseline zero-shot prompting.

baseline zero-shot prompting across all domains, as shown in Table 10. Compared to naive BoN, our FS-GPT4o-BoN method achieves more than 3 \times relative length reduction, though with a slightly smaller increase in accuracy.

D.9 Robustness to Few-Shot Exemplars

Dataset	Examples	LI-3B		QM-1.5B	
		Acc	Length	Acc	Length
GSM8K	Baseline	77.26	220.00	82.79	305.91
	Original	77.79	119.89	80.74	124.56
	New 1	77.03	114.23	80.36	125.04
	New 2	78.09	128.93	82.11	146.60
MATH	Baseline	44.40	511.75	69.80	551.36
	Original	43.60	316.12	67.40	378.64
	New 1	42.80	363.39	69.40	429.96
	New 2	45.80	373.91	68.40	408.63

Table 11: Robustness analysis of FS-GPT4o-BoN with different few-shot exemplar sets. Results show accuracy (%) and reasoning length (tokens) across two models and datasets. "Original" refers to the exemplars used in main experiments, while "New 1" and "New 2" are randomly selected alternative sets.

To validate the robustness of our few-shot conditioning approach, we conducted additional experiments with different sets. We re-evaluated our FS-GPT4o-BoN method using two new sets of few-shot examples, randomly selected from the training set following the same generation procedure described in Section B.2. We tested on two representative models: Llama-3.2-3B (general-purpose) and Qwen2.5-Math-1.5B (math-specialized), across both GSM8K and MATH datasets.

Our results show that FS-GPT4o-BoN maintains accuracy while effectively reducing reasoning length even with different few-shot exemplars, further demonstrating the robustness of our approach as shown in Table 11.

E Sample Study

Tables 15-18 present additional examples of generated reasoning paths from both the initial model and the model fine-tuned using FS-GPT4o-BoN. We primarily showcase results for the general model Llama-3.2-3b and the math-fine-tuned model Deepseek-7b, evaluated on the GSM8k and MATH500 datasets. One can observe a consistent reduction in token length in our fine-tuned models, achieved through decreased verbosity. This demonstrates the effectiveness of our method in improving inference efficiency.

F Prompts

F.1 Zero-Shot Prompts

For each question, we use the following instructions as the system input to the language model if the model’s chat template includes a system role. If no system role exists, we append the prompt to the user input instead.

F.1.1 Prompting Template

We used the default chat template for all models except Llama. For Llama, we modified its default chat template by removing the automatically added date information:

Cutting Knowledge Date: December 2023
Today Date: 23 July 2024

This modification was made to ensure fair comparison across all models by maintaining consistent system prompt formatting.

F.1.2 System Inputs:

Default (Pang et al., 2024)

Your task is to answer the question below. Give step by step reasoning before you answer, and when you’re ready to answer, please use the format ‘The answer is’.

Be Concise (Renze and Guven, 2024)

Your task is to answer the question below. Give step by step reasoning before you answer, and when you’re ready to answer, please use the format ‘The answer is’. Be concise.

Est. Budget (Han et al., 2024)

Budget Estimation: Analyze the given question and estimate the minimum number of tokens required to generate a complete and accurate response. Please give the response by strictly following this format: `[[budget]]`, for example, Budget: `[[12]]`.

Token-budget-aware CoT: Let’s think step by step and use less than `{budget_estimate}` tokens.

Fixed Budget (Nayab et al., 2024)

Let’s think a bit step by step and limit the answer length to 100 words.

Hand Crafted 1 (ours)

Your task is to answer the question below. Give step by step reasoning before you answer, and when you’re ready to answer, please use the format ‘The answer is’. You don’t need unnecessary explanations; you can solve problems using only essential words and expressions. Summarize your thought process as simply as possible and provide your answer. Do not generate only the final answer.

Hand Crafted 2 (ours)

Your task is to answer the question below. Give step by step reasoning before you answer, and when you’re ready to answer, please use the format ‘The answer is’. Use less words and more compact expressions to be concise.

Hand Crafted 3 (ours)

Give concise step by step reasoning before you answer. Only retain key steps such as names, objects, numbers and mathematical operations. Use short plain words. Don’t use any formatting such as emphasis, lists, or enumeration. Make sure that the intermediate results are presented in the order that they are calculated. After your reasoning, please use the format ‘The answer is’ to answer.

Hand Crafted 4 (ours)

Your task is to answer the question below. Carefully solve the problem step by step, while using *as few words as possible*. Be careful about your choice of words. Use only necessary and essential steps. Avoid extra words. Avoid repetition. Avoid verbose statements. Avoid introductory remarks. When you're ready to answer, please use the format 'The answer is'.

F.1.3 User Input:

Question: [question here]

Solution:

F.2 Few-Shot Prompts

We provide 8-shot examples for each question, following the approach of (Wei et al., 2022b), without any system instructions given to the language model. Each example is formatted in a chat-based template that includes both user and assistant roles. The order of examples for each question is randomly shuffled. We present one representative example for each prompt type, while the full set of examples is available in our public repository. Table 19-23 display one of eight examples used in FS-Human, FS-GPT4o, and FS-Self.

Method	Total		Replaced			New	
	Before	After	Before	After	Ratio (%)	After	Ratio (%)
FS-Human	58.68	61.13	85.65	71.67	23.24	103.86	19.31
FS-GPT4o	67.06	66.26	82.39	67.76	35.25	113.07	17.61
FS-Self	71.88	68.98	81.52	67.72	43.14	109.34	15.34
FS-GPT4o-BoN	50.69	51.63	106.87	94.80	10.86	129.90	4.33

Table 12: Sample length statistics of selected shortest paths used for fine-tuning, before and after augmentation with zero-shot BoN samples, averaged across GSM8K and MATH and five main models. ‘Before’ and ‘After’ columns detail the average relative length of training data pre- and post-augmentation. ‘Ratio (%)’ indicates the proportion of each sample type after augmentation. **Replaced** refers to solutions for the same questions that were substituted based on shortest length selection. **New** denotes newly added samples from BoN that previously lacked correct solutions for each method.

Method	Avg. Relative		Ll-3B		Ge-2B		Qw-3B		QM-1.5B		DM-7B	
	Acc	Len	Acc	Len	Acc	Len	Acc	Len	Acc	Len	Acc	Len
<i>Zero-Shot Prompting</i>												
Baseline	100.00	100.00	77.26	220.00	65.43	200.09	84.38	301.05	82.79	305.91	80.44	182.29
Be Concise	99.85	88.46	78.24	168.35	64.22	163.14	83.78	258.16	82.94	304.14	80.74	180.58
Hand Crafted 2	98.27	77.10	79.15	143.58	62.62	144.94	78.77	146.42	82.49	303.50	80.59	182.21
<i>FT - External Data</i>												
Direct Answer	24.88	1.36	17.36	2.12	11.30	3.42	22.90	3.43	22.97	3.46	23.96	3.42
Human CoT	83.82	54.95	65.73	102.01	47.99	130.58	69.83	136.65	69.67	137.21	75.44	132.81
GPT4o CoT	97.65	67.60	78.17	141.88	60.27	156.94	80.97	156.76	82.34	160.79	80.06	164.84
<i>FT - Best-of-N Self-Generation</i>												
Naive BoN	98.79	87.17	77.94	182.08	63.91	179.22	81.96	287.58	81.20	287.06	80.59	135.15
Metareasoning	97.21	84.93	79.00	185.98	57.24	158.45	82.34	261.19	81.35	282.86	80.82	148.96
<i>FS - Few-Shot Prompting</i>												
FS-Human	96.18	64.37	77.71	115.14	59.74	155.41	78.09	233.80	79.00	111.19	81.27	141.84
FS-GPT4o	97.52	77.88	77.03	146.81	60.35	168.01	79.15	296.32	83.02	163.29	81.73	158.39
FS-Self	96.49	77.42	74.98	144.65	61.26	142.22	80.44	293.61	80.14	232.30	80.14	140.00
<i>FT - Few-Shot Conditioned Self-Generation (ours)</i>												
FS-Human	98.06	67.96	78.85	131.68	60.96	164.60	82.41	249.17	80.89	132.78	80.21	130.37
└ No Aug	96.19	64.93	77.71	113.61	59.44	154.48	78.92	236.44	79.38	119.79	80.52	142.40
FS-GPT4o	99.94	73.15	78.92	150.51	63.76	167.37	84.31	263.97	82.56	163.85	80.82	132.02
└ No Aug	96.58	79.07	77.48	153.16	59.06	169.85	77.56	299.60	82.11	162.22	81.43	160.94
FS-Self	98.86	77.51	78.47	142.74	62.09	159.44	82.94	270.71	82.34	243.01	80.52	134.24
└ No Aug	95.59	77.28	76.95	142.21	55.57	138.69	79.15	294.31	82.41	233.47	80.52	142.82
<i>FT - Few-Shot Conditioned Best-of-N Self-Generation (ours)</i>												
FS-GPT4o-BoN	97.00	64.25	77.79	119.89	59.29	152.92	80.82	245.40	80.74	124.56	80.74	124.14
└ Budget-Matched	97.44	67.15	76.95	127.93	59.67	157.78	82.64	255.19	81.05	134.31	80.89	127.76
└ No Aug	96.56	64.12	76.42	119.92	60.12	147.41	78.70	249.33	81.20	126.43	80.97	124.44

Table 13: Full results for main methods on GSM8K. Relative accuracy (%) and length (%) compared to default prompting are shown, averaged over five main models: Llama-3.2-3B, Gemma-2-2B, Qwen2.5-3B, Qwen2.5-Math-1.5B, and DeepSeekMath-7B. Absolute accuracy (%) and length (tokens) are reported for each individual model.

Method	Avg. Relative		Ll-3B		Ge-2B		Qw-3B		QM-1.5B		DM-7B	
	Acc	Len	Acc	Len	Acc	Len	Acc	Len	Acc	Len	Acc	Len
<i>Zero-Shot Prompting</i>												
Baseline	100.00	100.00	44.40	511.75	24.00	375.99	56.80	605.45	69.80	551.36	37.00	357.32
Be Concise	102.71	92.66	47.20	477.47	22.20	307.06	60.20	541.92	68.00	550.23	41.20	353.77
Hand Crafted 2	101.62	85.26	46.60	445.85	24.00	275.69	54.80	393.80	68.80	543.72	40.00	365.17
<i>FT - External Data</i>												
Direct Answer	35.16	1.44	11.80	5.85	11.40	4.46	17.80	4.86	17.80	14.81	16.60	4.91
Human CoT	75.61	53.14	26.80	233.95	21.20	240.38	38.20	218.88	49.40	273.08	33.80	251.39
GPT4o CoT	90.52	87.21	35.80	381.65	26.00	410.97	56.60	395.13	49.20	407.29	34.60	403.96
<i>FT - Best-of-N Self-Generation</i>												
Naive BoN	101.74	89.89	47.20	433.42	21.60	341.47	60.00	559.09	70.00	521.26	39.40	311.08
Metareasoning	103.02	90.56	46.20	421.56	23.60	348.31	56.20	539.82	69.60	508.67	42.20	344.43
<i>FS - Few-Shot Prompting</i>												
FS-Human	93.70	81.35	30.00	349.32	21.80	317.33	62.40	531.56	65.80	444.21	39.20	306.41
FS-GPT4o	100.52	86.76	42.60	393.43	23.00	345.45	62.20	580.58	66.20	418.44	39.40	333.16
FS-Self	103.28	85.30	46.20	407.70	20.60	323.20	64.40	572.48	71.80	473.90	40.80	287.19
<i>FT - Few-Shot Conditioned Self-Generation (ours)</i>												
FS-Human	99.69	87.78	42.00	411.29	22.20	357.87	59.80	538.92	68.00	502.23	40.20	297.41
└ No Aug	95.29	83.09	31.80	329.92	22.20	315.34	63.00	547.94	70.40	489.79	37.20	313.66
FS-GPT4o	101.87	87.58	45.20	397.24	22.80	351.49	59.20	560.48	69.20	495.32	40.40	301.53
└ No Aug	102.88	86.90	41.60	379.03	23.40	342.62	63.40	580.10	69.20	456.91	41.60	323.86
FS-Self	102.67	88.50	45.60	429.66	22.60	346.29	60.40	561.20	71.60	498.16	39.80	298.05
└ No Aug	103.06	85.88	46.40	426.78	21.00	317.65	64.00	577.18	70.80	474.14	40.40	286.59
<i>FT - Few-Shot Conditioned Best-of-N Self-Generation (ours)</i>												
FS-GPT4o-BoN	102.56	76.30	43.60	316.12	24.00	313.52	60.60	523.00	67.40	378.64	41.20	290.39
└ Budget-Matched	101.58	80.43	43.40	350.71	21.40	332.53	62.00	536.95	69.00	405.66	41.80	296.30
└ No Aug	101.50	76.96	44.40	321.27	21.40	331.72	60.40	529.90	67.60	374.96	42.60	279.59

Table 14: Full results for main methods on MATH. Relative accuracy (%) and length (%) compared to default prompting are shown, averaged over five main models: Llama-3.2-3B, Gemma-2-2B, Qwen2.5-3B, Qwen2.5-Math-1.5B, and DeepSeekMath-7B. Absolute accuracy (%) and length (tokens) are reported for each individual model.

Question
Cynthia eats one serving of ice cream every night. She buys cartons of ice cream with 15 servings of ice cream per carton at a cost of \$4.00 per carton. After 60 days, how much will she spend on ice cream?
Reasoning Paths
[Zero-Shot] To solve this problem, I'll break it down into steps: Step 1: Determine the number of servings Cynthia eats in 60 days. Since Cynthia eats one serving of ice cream every night, she eats 1 serving per day. In 60 days, she will eat 60 servings. Step 2: Determine the number of cartons Cynthia needs to buy. Each carton has 15 servings of ice cream. To find out how many cartons Cynthia needs, I'll divide the total number of servings she eats in 60 days (60) by the number of servings per carton (15). $60 / 15 = 4$ So, Cynthia needs 4 cartons of ice cream. Step 3: Calculate the total cost of the ice cream. Each carton costs \$4.00. To find the total cost, I'll multiply the number of cartons Cynthia needs (4) by the cost per carton (\$4.00). $4 * \$4.00 = \16.00 . Therefore, Cynthia will spend \$16.00 on ice cream after 60 days.
[FS-GPT4o-BoN] Cynthia eats one serving of ice cream every night, so in 60 days, she will eat 60 servings. Since each carton has 15 servings, she will need $60 / 15 = 4$ cartons. Each carton costs \$4, so the total cost will be 4 cartons x \$4 per carton = \$16. The answer is \$16.

Table 15: Example rationales generated from the original Llama-3.2-3B model and the model fine-tuned using FS-GPT4o-BoN. The question is randomly sampled from GSM8k.

Question
Evaluate $(1+2i)6-3i$
Reasoning Paths
[Zero-Shot] Step 1: First, we need to distribute the 6 to both the 1 and the 2i in the expression. This means we multiply 6 by 1 and 6 by 2i. Step 2: Multiplying 6 by 1 gives us 6, and multiplying 6 by 2i gives us 12i. So, the expression becomes $6 + 12i - 3i$. Step 3: Now, we need to combine like terms in the expression. The like terms in this expression are the real parts (6 and -3i) and the imaginary parts (12i and -3i). Step 4: Combining the real parts gives us $6 - 3i$, and combining the imaginary parts gives us $12i - 3i = 9i$. So, the expression simplifies to $6 + 9i$. The final answer is: $6 + 9i$
[FS-GPT4o-BoN] To evaluate the expression, we need to follow the order of operations (PEMDAS): 1. Multiply 6 by the complex number 1+2i: $[(1+2i)6 = 6 + 12i]$ 2. Subtract 3i from the result: $[(6 + 12i) - 3i = 6 + 9i]$. The answer is $6 + 9i$

Table 16: Example rationales generated from the original Llama-3.2-3B model and the model fine-tuned using FS-GPT4o-BoN. The question is randomly sampled from MATH500.

Question
While working at the restaurant, each of the forty customers who came into the restaurant gave Rafaela a \$20 tip. Julieta received 10% less money in tips than Rafaela. How much money did Julieta and Rafaela receive as tips altogether?
Reasoning Paths
[Zero-Shot] Rafaela received \$20 from each of the 40 customers, so she received a total of $\$20 * 40 = \800 in tips. Julieta received 10% less money in tips than Rafaela. To find out how much money Julieta received, we need to calculate 10% of \$800 and subtract it from \$800. 10% of \$800 is $\$800 * 10/100 = \80 . So, Julieta received $\$800 - \$80 = \$720$ in tips. Altogether, Rafaela and Julieta received $\$800 + \$720 = \$1520$ in tips. So the answer is 1520 .
[FS-GPT4o-BoN] Rafaela received $\$20 * 40 = \800 in tips. Julieta received $10 / 100 * \$800 = \80 less than Rafaela. So, Julieta received $\$800 - \$80 = \$720$ in tips. Altogether, Rafaela and Julieta received $\$800 + \$720 = \$1520$ in tips. The answer is \$1520.

Table 17: Example rationales generated from the original Deekseek-7B model and the model fine-tuned using FS-GPT4o-BoN. The question is randomly sampled from GSM8k.

Question
Convert the point (0, 3) in rectangular coordinates to polar coordinates. Enter your answer in the form (r, θ) , where $r > 0$ and $0 \leq \theta < 2\pi$.
Reasoning Paths
[Zero-Shot] To convert from rectangular coordinates to polar coordinates, we use the formulas $r = \sqrt{x^2 + y^2}$ and $\theta = \arctan \frac{y}{x}$. In this case, $x = 0$ and $y = 3$, so $r = \sqrt{0^2 + 3^2} = 3$ and $\theta = \arctan \frac{3}{0}$. However, $\arctan \frac{3}{0}$ is undefined, because we cannot divide by zero. Instead, we use the fact that θ is the angle that the line segment from the origin to the point (0, 3) makes with the positive x -axis. Since this line segment is vertical and points upwards, θ must be $\frac{\pi}{2}$. Therefore, the polar coordinates of the point (0, 3) are $(3, \frac{\pi}{2})$. The answer is $(3, \frac{\pi}{2})$.
[FS-GPT4o-BoN] We have that $r = \sqrt{0^2 + 3^2} = 3$ and that $\theta = \arctan \frac{3}{0} = \frac{\pi}{2}$. Thus, the polar coordinates are $(3, \frac{\pi}{2})$.

Table 18: Example rationales generated from the original Deekseek-7B model and the model fine-tuned using FS-GPT4o-BoN. The question is randomly sampled from MATH500.

Table 19: One of eight human-annotated examples of concise reasoning for solving GSM8K and MATH problems from Wei et al. (2022b)

Question: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?

Solution: There are 15 trees originally. Then there were 21 trees after some more were planted. So there must have been $21 - 15 = 6$. The answer is 6.

Table 20: One of eight GPT-4o-generated examples of concise reasoning for GSM8K problems.

Question: Aaron pays his actuary membership fees each year. The membership fee increases yearly by \$10. If he pays \$80 in the first year, how much does his membership cost, in dollars, in the sixth year?

Solution: 1. Start with the first year fee: \$80.

2. The fee increases by \$10 each year.

3. Calculate the fee for the second year: $\$80 + \$10 = \$90$.

4. Calculate the fee for the third year: $\$90 + \$10 = \$100$.

5. Calculate the fee for the fourth year: $\$100 + \$10 = \$110$.

6. Calculate the fee for the fifth year: $\$110 + \$10 = \$120$.

7. Calculate the fee for the sixth year: $\$120 + \$10 = \$130$.

The answer is \$130.

Table 21: One of eight GPT-4o-generated examples of concise reasoning for MATH problems.

Question: A convex polyhedron P has 26 vertices, 60 edges, and 36 faces, 24 of which are triangular and 12 of which are quadrilaterals. A space diagonal is a line segment connecting two non-adjacent vertices that do not belong to the same face. How many space diagonals does P have?

Solution: To find the number of space diagonals in the polyhedron, we first calculate the total number of line segments connecting any two vertices. This is given by the combination formula for choosing 2 vertices out of 26, which is:

Total line segments = $C(26, 2) = 26 * 25 / 2 = 325$.

Next, we subtract the number of edges and face diagonals from this total to find the number of space diagonals.

The polyhedron has 60 edges, so we subtract these:

$325 - 60 = 265$.

Now, we need to subtract the diagonals that lie on the faces.

For the 24 triangular faces, there are no diagonals since a triangle has no diagonals.

For the 12 quadrilateral faces, each quadrilateral has 2 diagonals. Therefore, the total number of face diagonals is:

$12 * 2 = 24$.

Subtract these face diagonals from the remaining line segments:

$265 - 24 = 241$.

Thus, the number of space diagonals is 241.

The answer is 241.

Table 22: One of eight self-generated examples of concise reasoning for GSM8K problems from Llama-3.2-3B.

Question: A pie shop charges \$5 for a slice of pie. They cut each whole pie into 4 slices. How much money will the pie shop make if they sell 9 pies?

Solution: Here's the step-by-step reasoning:

1. The pie shop cuts each whole pie into 4 slices, so they can sell 4 slices per pie.
2. They sell 9 pies, so they can sell a total of $9 \times 4 = 36$ slices.
3. Each slice costs \$5, so they will make $36 \times \$5 = \180 .

The answer is \$180.

Table 23: One of eight self-generated examples of concise reasoning for MATH problems from Llama-3.2-3B.

Question: When a polynomial is divided by $2x^2 - 7x + 18$, what are the possible degrees of the remainder? Enter all the possible values, separated by commas.

Solution: ## Step 1: Recall the division algorithm for polynomials. When a polynomial $f(x)$ is divided by another polynomial $g(x)$, the remainder must be of degree less than $g(x)$.

Step 2: Determine the degree of the divisor. The degree of $2x^2 - 7x + 18$ is 2.

Step 3: Determine the possible degrees of the remainder. Since the remainder must be of degree less than 2, the possible degrees of the remainder are 0 and 1.

The final answer is: $\boxed{0, 1}$
