Neurosymbolic Artificial Intelligence
XX(X):1–30
©The Author(s) 2026
neurosymbolic-ai-journal.com

Incorporating Token Usage into Prompting Strategy Evaluation

Anonymous Author(s)

Abstract

In recent years, large language models have demonstrated remarkable performance across diverse tasks. However, their task effectiveness is heavily dependent on the prompting strategy used to elicit output, which can vary widely in both performance and token usage. While task performance is often used to determine prompting strategy success, we argue that efficiency—balancing performance and token usage—can be a more practical metric for real-world utility. This is especially relevant to the growing debate of whether LLMs with, say, chain-of-thought, counts as a neuro-symbolic pipeline, and where the boundary should be drawn with explicit symbolic solvers and their effectiveness vs linguistic reasoning strategies. To enable this, we propose Big-O_{tok}, a theoretical framework for describing the token usage growth of prompting strategies, and analyze Token Cost, an empirical measure of tokens per performance. We apply these to several common prompting strategies to demonstrate their utility and observe that increased token usage leads to drastically diminishing performance returns. This finding has become increasingly relevant with the rise of reasoning-oriented models and multi-hop agentic workflows, where token budgets can grow by orders of magnitude. Our results validate the Big-O_{tok} and Token Cost analyses and reinforce the need for efficiency-aware evaluations.

Keywords

Large Language Models, Prompting Strategies, Token Efficiency, Chain-of-Thought, Evaluation

1 Introduction

Large language models (LLMs) are primarily interacted with through natural language prompts. The composition of a prompt exercises significant, often unexpected, influence over the generated output. This has sparked research into "prompt engineering," the

[0] Anonymous Institution

study of prompt design to extract maximum performance from LLMs (White et al. 2023). There are many ways to approach prompt engineering; in this paper, we focus on formalized **prompting strategies**, the overarching paradigms of prompt design (e.g., providing examples of question-answer pairs (Brown et al. 2020)).

As prompting strategies have developed alongside LLMs, benchmark accuracy has emerged as the primary metric for success. New prompting strategies are often tested alongside prior ones on a selection of benchmarks and LLMs, using the gain in accuracy over existing strategies as validation of the new approach. Token usage, if included, is often analyzed post hoc, indicating that it was only a secondary consideration during development. Optimization for performance without regard for token usage can lead to inefficient prompting strategies. Our purpose in this work is to demonstrate another, more holistic approach to prompting strategy evaluation and analysis by (1) proposing Big-O_{tok}, a framework for comparing theoretical token usage between distinct prompting strategies and (2) introducing Token Cost (TC), a simple empirical metric to quantify prompting strategy efficiency.

The need for such efficiency-aware metrics has only grown since this line of work was initiated. Reasoning-oriented models such as OpenAI's o1 and o3 families and DeepSeek-R1 (DeepSeek-AI et al. 2025) routinely generate tens of thousands of reasoning tokens per query, and agentic workflows that chain multiple LLM calls compound this further. In this setting, the cost of prompting strategies is no longer a secondary concern but a first-order design constraint. Meanwhile, there is also a growing debate of whether LLMs with, say, chain-of-thought, counts as a neuro-symbolic pipeline, and where the boundary should be drawn with explicit symbolic solvers and their effectiveness vs linguistic reasoning strategies (Valmeekam et al. 2025).

To achieve these goals, we approach prompting strategy efficiency on two fronts: theoretical and empirical. For our theoretical analysis, we derive Big-O_{tok} token complexities for a selection of prompting strategies, similar to time complexity analyses common in software engineering. We substantiate our Big-O_{tok} analyses by evaluating our selection of prompting strategies against common benchmarks using multiple models. We analyze the results of those experiments in terms of TC, to compare how performance and token usage interact. For our experiments, we observe that, while there is performance improvement to be gained from more complex, token-hungry prompting strategies, increasing token usage results in drastically diminishing performance returns.

2 Related Work

Although LLM efficiency has been an active area of research for years (see Wan et al. (2024)), underwhelming emphasis has been placed on the efficiency of prompting strategies. Techniques have emerged to reduce token usage, such as frameworks that dynamically manage token usage at inference time—e.g., FrugalGPT (Chen et al. 2023) and LLMLingua (Jiang et al. 2023)—and prompt compression—e.g., Mu et al. (2023). These approaches seek to enable efficient LLM usage in spite of inefficient prompting strategies, whereas this work promotes a focus on prompting strategy efficiency.

Some work, such as Sivarajkumar et al. (2024); Kim et al. (2023); Chu et al. (2024), has been done to evaluate prompting strategies in specific domains, but token usage statistics are often not included. Prompting strategy evaluation is largely left to new prompting strategy proposals (e.g., CoT (Wei et al. 2022)) or benchmarks (e.g., BBH (Suzgun et al. 2023)). These evaluations tend to prioritize benchmark accuracy without significant consideration of token usage. In this work, we demonstrate the relevance of an increased focus on token efficiency and how to proactively incorporate it into prompting strategy analysis.

Some metrics pertaining to cost and efficiency have emerged but tend to be tailored to specific use cases. Wang et al. (2024) proposes a budget-limited evaluation framework to compare prompting strategies and explores why certain prompting strategies may not scale performance with increased compute budget, using raw token and query counts as cost metrics. Wan et al. (2025) uses the weighted average of piecewise accuracy and cost functions to quantify efficiency specifically for self-consistent methods. We propose TC as a simple, independent metric of prompting strategy efficiency and Big-O_{tok} as an analysis that can be used to compare arbitrary prompting strategies without execution.

Some recent prompting strategies, such as Constrained-CoT (Nayab et al. 2024), Concise-CoT (Renze and Guven 2024), and Algorithm-of-Thoughts (Sel et al. 2024), are designed as optimizations over extant strategies with token usage reduction as a primary motivation. Our hope for this work is that it will enable future prompting strategy development and analysis that similarly prioritizes efficiency.

Since this work was first developed, the question of token efficiency in LLM reasoning has attracted considerable further attention. Han et al. (2025) propose token-budget-aware reasoning frameworks that dynamically allocate token budgets based on problem complexity, showing that chain-of-thought reasoning can be substantially compressed with minimal accuracy loss. Wilhelm et al. (2025) extend the efficiency discussion beyond monetary cost to energy consumption, showing that strategies like majority voting and CoT can increase energy usage substantially relative to zero-shot baselines. More recently, budget-aware scaling has been studied where both token consumption and tool-call counts must be jointly managed (Liu et al. 2025). These developments reinforce the central premise of our work: that token usage deserves systematic treatment in prompting strategy evaluation, not merely post-hoc acknowledgement.

3 Methodology

We explore the importance of token usage both theoretically and empirically. Due to the popularity of LLMs, there exists an infeasible number of possible evaluation combinations[1]. To focus the scope of this paper on token usage, we restrict the number of prompting strategies, benchmarks, and models we use. We discuss our selection processes in Sections 3.1 and 3.2.

[1]Prompting strategies: 40+ (Vatsal and Dubey 2024; Chu et al. 2024); Benchmarks: 130+ (Gao et al. 2023); Open-source, benchmarked LLMs: 3200+, as of January, 2025 (Fourrier et al. 2024).

3.1 Theoretical Analysis

Table 1. Big-O_{tok} token complexities for each prompting strategy.

Prompting Strategy	Big-O_{tok}	Variables
Vanilla IO	$O(1)$	
Zeroshot CoT (Kojima et al. 2022)	$O(1)$	
Vanilla Fewshot (Brown et al. 2020)	$O(k)$	k: k-shot exemplars
Fewshot CoT (Wei et al. 2022)	$O(k)$	k: k-shot exemplars
CoT-SC (Wang et al. 2023b)	$O(pk)$	k: k-shot exemplars; p: sampled chains

We categorize prompting strategies into three broad groups: **(1) linguistic prompt engineering**, which relies on specific phrasing techniques—e.g., Plan-and-Solve (Wang et al. 2023a) or Zeroshot CoT (Kojima et al. 2022); **(2) in-context learning**, which consists of providing examples of task-response pairs before providing the task to the LLM—e.g., Vanilla Fewshot (Brown et al. 2020) or Fewshot CoT (Wei et al. 2022); and **(3) multi-hop**, which is characterized by multiple LLM calls—e.g., Least-To-Most (Zhou et al. 2023), Tree-of-Thought (Yao et al. 2023), or CoT Self-Consistency[2] (Wang et al. 2023b). These three prompting strategy categories roughly correspond to the following Big-O_{tok} complexity classes, respectively: **(1)** constant—e.g., the consistent overhead of "Think step by step" (Wei et al. 2022); **(2)** linear—e.g., the number of fewshot exemplars; and **(3)** polynomial or higher—e.g., the number of multi-hop steps times the number of exemplars. These Big-O_{tok} complexity classes are reflected in Table 1.

To ensure our investigation represents all three categorizations, we select prompting strategies from each. Namely, we choose Vanilla IO (i.e., simply providing the benchmark question) as a baseline; Zeroshot CoT to represent **(1)**; Vanilla Fewshot and Fewshot CoT for **(2)**; and CoT-SC for **(3)**. These strategies are widely adopted, tend to build on each other without significant changes to prompt design, and demonstrate an organic evolution of prompting strategies over several years (Chu et al. 2024).

The purpose of Big-O_{tok} is to provide an objective representation of the theoretical token usage growth rate of a given prompting strategy, enabling direct comparison with the Big-O_{tok} of other prompting strategies. Big-O_{tok} is based on Big-O notation (Knuth 1976) and thus we rely on the terminology and definitions associated with it.

Big-O_{tok} describes the asymptotic growth of token usage as a function of variables[3] in the prompting strategy (e.g., the number of fewshot examples). It is derived analogously to Big-O time complexity: by considering how token usage increases as prompting strategy variables approach infinity. The variables with the highest growth rate dominate the other terms (e.g., constants, lower-order variables, and scalars), which can then

[2] Abbreviated as CoT-SC$_n$, where n is the number of sampled chains.

[3] We treat the initial input that the prompting strategy modifies (e.g., a benchmark question) to be constant and exclude it for simplicity. Similarly, we treat additive adjustments to the input or output (e.g., CoT's "Think step by step" (Wei et al. 2022)) as constants.

be omitted. Big-O_{tok} token complexity can often be derived from a natural language description of a prompting strategy. We provide sample derivations for the Big-O_{tok} functions from Table 1 in Section 4.

3.2 Empirical Analysis

We test our selection of prompting strategies against three common benchmarks using three LLMs. To perform the empirical evaluations, we leverage LM Evaluation Harness (Biderman et al. 2024; Gao et al. 2023), a framework aimed at increasing the reproducibility of LLM evaluations.

We base our selection of models on recency, popularity, and size. We do not use commercial models due to budget constraints[4]. We select Llama 3.1 8B Instruct (Dubey et al. 2024), Qwen 2.5 14B Instruct, and Qwen 2.5 32B Instruct (Qwen et al. 2025). This selection provides coverage of various sizes of smaller models (each approximately doubling the parameter count of the prior) and diversity of origin, to ensure multiple approaches to data collection, training, and alignment are represented[5].

For benchmarks, we select BBH (Suzgun et al. 2023), GSM8K (Cobbe et al. 2021), and MMLU (Hendrycks et al. 2021). This represents a diverse group of general-purpose benchmarks based on typical accuracy ranges[6] and response type[7]. For fewshot prompting strategies, we use 3 exemplars for BBH, 8 for GSM8K, and 4 for MMLU[8].

4 Deriving Big-O_{tok} Token Complexities

We provide an expanded token complexity analysis for each prompting strategy examined above in Table 2. We define the minimally viable IO pair (MVIO) for a given benchmark question to be the full text of the question and the minimum amount of text to convey the answer (e.g., Question: “How many days are there in a week?”; Answer: “7”). All other prompting strategies that incur additive adjustments to the input or output (e.g., the natural language thinking induced by CoT’s “Think step by step” (Wei et al. 2022)) are treated as constant overheads on top of the MVIO, which are represented by Greek letters.

Table 3 shows theoretical token usage ratios based on Big-O_{tok}. The expected token usage ratios used in Table 4 are based on these, averaged across the three benchmarks.

We include examples of Big-O_{tok} derivations for each prompting strategy examined above in Figure 1.

[4]See Section 7.3 for cost estimates for commercial APIs.
[5]We choose two Qwen 2.5 models to facilitate a comparison between model size, found in Section 8.2.
[6]GSM8K: 80-95%; BBH: 50-87%; MMLU: 70-92% (Fourrier et al. 2024; Dubey et al. 2024; Qwen et al. 2025).
[7]GSM8K: free response number; BBH: free response text; MMLU: multiple choice.
[8]These numbers are based on the availability of CoT examples in LM Evaluation Harness and closely reflect the number of examples suggested in CoT-SC (Wang et al. 2023b).

Table 2. The theoretical token complexities for various prompting strategies. Greek letters represent the overhead associated with an IO pair (assumed constant per prompting strategy) and Roman letters represent variables.

Prompting Strategy	Big-O_{tok}	Token Complexity	Variables	Values[†]
MVIO	$O(1)$	1		
Vanilla IO	$O(1)$	$1+\psi$		
Zeroshot CoT (Kojima et al. 2022)	$O(1)$	$1+\alpha$		
Vanilla Fewshot (Brown et al. 2020)	$O(k)$	$1+k$	k: k-shot exemplars	$k = 0, 1, 10-100$
Fewshot CoT (Wei et al. 2022)	$O(k)$	$1+\alpha+k+k\alpha$	k: k-shot exemplars	$k=8$
CoT-SC (Wang et al. 2023b)	$O(pk)$	$p(1+\alpha+k+k\alpha)$	k: k-shot exemplars; p: sampled chains	$k=4-8$; $p=40$

[†] Values suggested in the papers that originally introduced the prompting strategy.

Table 3. The token usage growth rate over MVIO per prompting strategy, derived from Table 1 using the variables used in our experiments.

Prompting Strategy	Token Usage Growth Rate		
	BBH	GSM8K	MMLU
Vanilla IO	1	1	1
CoT Zeroshot	1	1	1
Vanilla Fewshot	3	8	4
Fewshot CoT	3	8	4
CoT-SC_5	15	40	20
CoT-SC_{10}	30	80	40

5 Interpreting Token Cost

Although we include a thorough example of using TC for prompting strategy analysis, we seek to provide an expanded discussion on its interpretation here. For brevity, we stated above that, generally, low TC can be thought of as more efficient and high TC as less efficient. In most instances, this will hold true; however, there are some plausible edge cases that deserve consideration.

5.1 Average TC

One such edge case is exploiting extremely low token usages to achieve low average TC. For example, consider a multiple-choice benchmark with four options: (A), (B), (C), and (D). Assuming uniform distribution across the possible answers, a prompting strategy that would yield 25% accuracy (assuming the LLM could consistently produce the correct output) might be: "Output (B)." At approximately 4 combined input and outputs tokens, such a prompting strategy would achieve an average TC of 0.16 $\frac{t}{p}$, a value much lower than any of our observed values in Table 6. This demonstrates the need to test prompting strategies for generalizability.

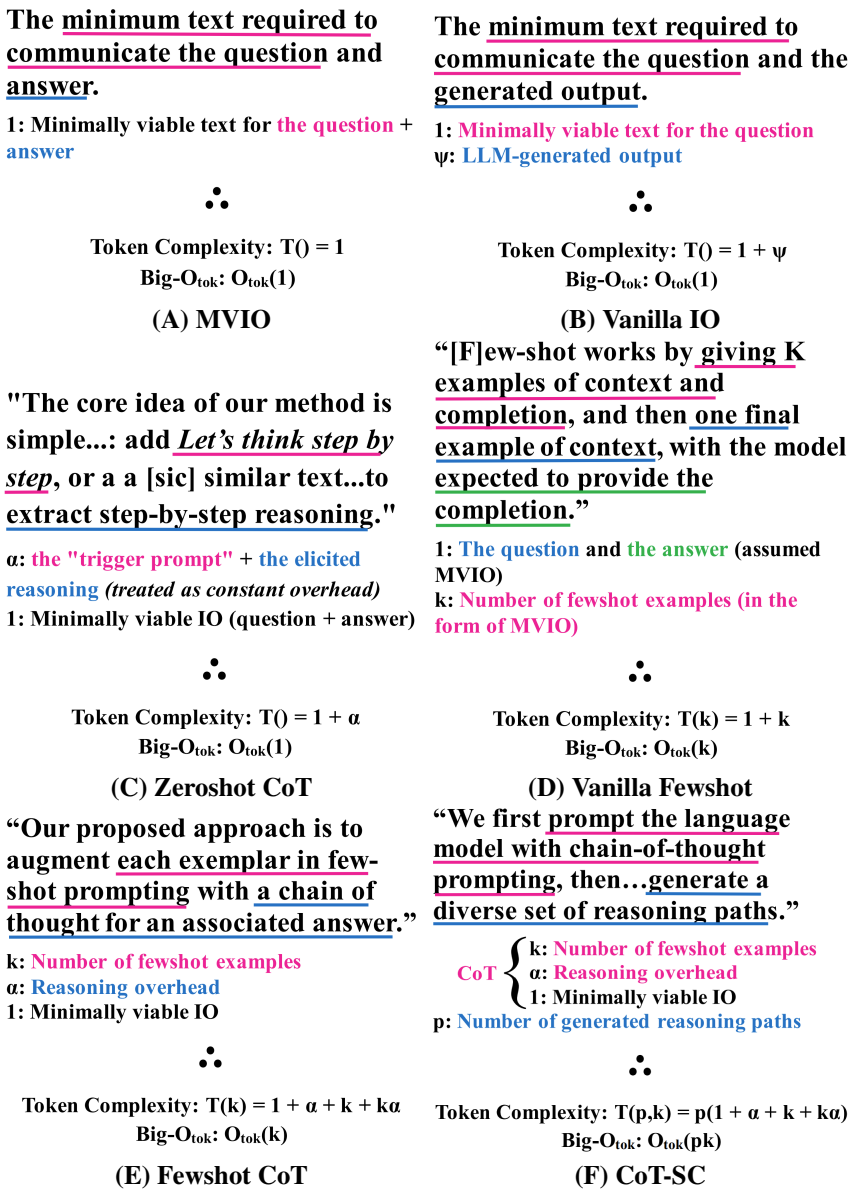


Figure 1. Sample derivations of Big-O_{tok}. The textual descriptions in each figure are drawn from the following sources: (C) Kojima et al. (2022); (D) Brown et al. (2020); (E) Wei et al. (2022); (F) Wang et al. (2023b). Note that for (D), the fewshot examples are equivalent to the MVIO and we make the assumption that the LLM follows that pattern.

5.2 Marginal TC

For marginal TC, there are additional edge cases to consider. The formula for calculating marginal TC,

$$\frac{num\ tokens_2 - num\ tokens_1}{accuracy_2 - accuracy_1}$$

where $num\ tokens_2 >= num\ tokens_1$, allows for negative values. Despite being "low," a negative marginal TC value indicates extreme inefficiency since the prompting strategy will have consumed more tokens to achieve lower accuracy.

In the unlikely event that $num\ tokens_2 == num\ tokens_1$, marginal TC will not provide useful information. The more efficient prompting strategy would, in that case, be the one that attained higher accuracy. Similarly, if $accuracy_2 == accuracy_1$, marginal TC cannot be calculated, but the prompting strategy that consumed fewer tokens can be considered more efficient.

6 Results

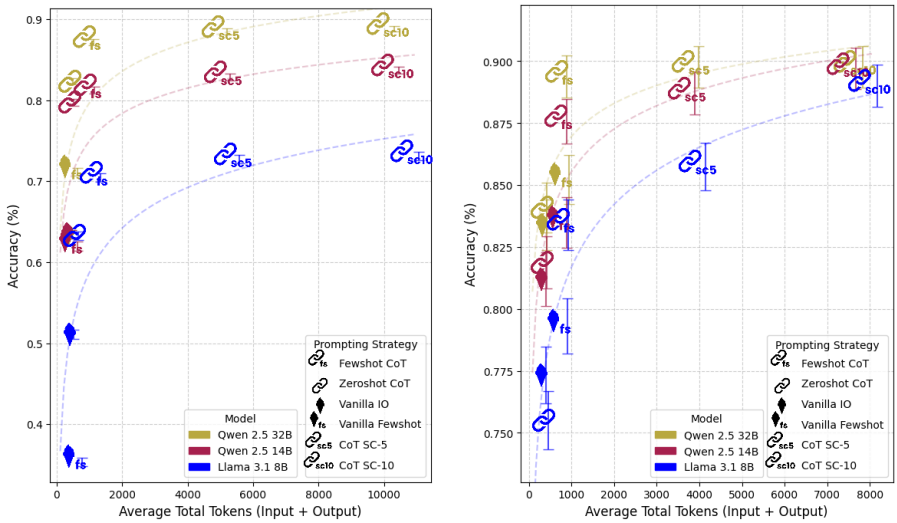
6.1 Big-O_{tok}

Table 4. Theoretical and observed token usage ratios between prompting strategies, averaged over the three benchmarks. Values are formatted as <theoretical>; <observed> and derived by $\frac{num\ tokens_2}{num\ tokens_1}$, where $num\ tokens_2 > num\ tokens_1$.

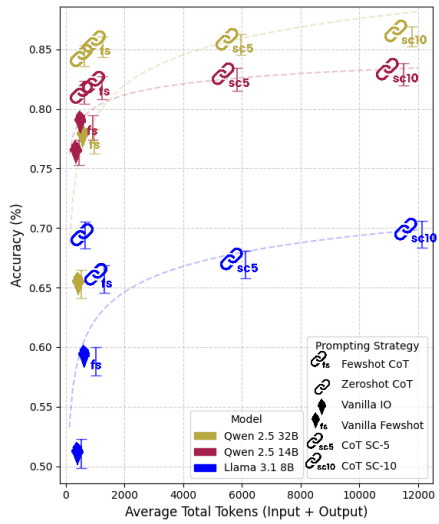
Column: Higher Token Usage (Numerator) **Row:** Lower Token Usage (Denominator)	CoT-SC$_{10}$	CoT-SC$_5$	Fewshot CoT	Vanilla Fewshot	Zeroshot CoT
Vanilla IO	50; 29.3	25; 14.6	5; 3.0	5; 2.2	1; 1.3
Zeroshot CoT	50; 23.4	25; 11.7	5; 2.4	5; 1.7	
Vanilla Fewshot	10; 13.5	5; 6.8	1; 1.4		
Fewshot CoT	10; 9.9	5; 5.0			
CoT-SC$_5$	2; 2.0				

To substantiate our Big-O$_{tok}$ analyses, we use the observed token usages from our experiments to calculate the relative token usage ratios between prompting strategies. We derive theoretical estimates of those ratios from our Big-O$_{tok}$ functions by substituting in the values from our experiments for the variables in Big-O$_{tok}$ (e.g., $p = 5$ and $\bar{k} = 5$ for CoT-SC$_5$). The results of that comparison are found in Table 4. We expect noise in the observed token usage due to: inherent token usage (e.g., chat templates); the relatively low values of prompting strategy variables (e.g., $k = 3$ fewshot exemplars for BBH); and the unpredictability of LLM output. However, while the observed and theoretical factors are not perfect matches, our findings do correctly align with the Big-O$_{tok}$ token complexity classes discussed in Section 3.1 (constant, linear, and polynomial). In other words,

$$T_{\text{O}_{\text{tok}}(1),\text{b,m}}() < T_{\text{O}_{\text{tok}}(\text{k}),\text{b,m}}(k) < T_{\text{O}_{\text{tok}}(\text{pk}),\text{b,m}}(p,k)$$



(A) BBH **(B) GSM8K**



(C) MMLU

Figure 2. Accuracy vs. token usage plots with standard error bars for various prompting strategies, models, and benchmarks. The trend lines reflect the rapid growth of TC for these strategies.

where $T_{\mathrm{O_{tok}},\mathrm{b},\mathrm{m}}(x)$ is the observed token usage for each prompting strategy in the $\mathrm{O_{tok}}$ complexity class, benchmark b, model m, and variable x in our experiments.

6.2 Token Cost

To quantify token efficiency, we discuss the results from our experiments in terms of Token Cost (TC). We define TC as the number of tokens[9] per percentage point of accuracy (expressed as $\frac{t}{p}$). The inverse of TC can be thought of as token efficiency; thus, relatively high TC is less efficient while lower TC is more efficient. We use this metric, $\frac{t}{p}$, rather than the inverse, $\frac{p}{t}$, because we find it to be more intuitive in the context of prompting strategies. We include an expanded discussion on interpreting TC, including edge cases, in Section 5.

The results of the experiments outlined in Section 3.2 are found in Figure 2. Across all benchmarks and models, our empirical results follow consistent trend lines (of the form $y = \log(\log(x))$), reflecting the diminishing accuracy returns from increased token usage. In other words, it requires significantly more tokens to realize accuracy gains as token usage increases. To discuss this trend in terms of TC, we explore both *average* and *marginal* TC[10]. Average TC is simply the token usage divided by the accuracy for a given prompting strategy ($\frac{num\ tokens_{obs}}{accuracy_{obs}}$). Marginal TC is the change in token usage to realize the change in accuracy between two prompting strategies. In other words, for $num\ tokens_2 >= num\ tokens_1$,

$$\frac{num\ tokens_2 - num\ tokens_1}{accuracy_2 - accuracy_1}$$

Both average and marginal TC can be thought of as the slope between two points (one being the origin, for average TC), which represents the expected cost, in tokens, of adding one point of accuracy.

Across all experiments, the average TC for the prompting strategy with the lowest accuracy is 5.0 $\frac{t}{p}$, while that of the highest performing prompting strategy is 119.4 $\frac{t}{p}$, a more than 20x increase in average TC and, inversely, 20x *decrease* in efficiency. This is reflected in the plots in Figure 2 in that even the worst performing prompting strategies still attain relatively high accuracy and more complex ones make only small gains over that for vastly more token usage. Using accuracy as the sole metric ignores the drastic decrease in efficiency that can result from increased token usage.

In Figure 2, we observe an initially steep curve, indicating that tokens are traded relatively efficiently for accuracy, followed by a plateau, where tokens are traded more inefficiently for accuracy. To compare the marginal TC along this curve, we select Vanilla IO, Fewshot CoT, and CoT-SC$_{10}$ which tend to lie on the extremes of the trend lines. Across all experiments, the marginal TC between Vanilla IO and Fewshot CoT is 65.3 $\frac{t}{p}$ while the marginal TC between Fewshot CoT and CoT-SC is 6701.8 $\frac{t}{p}$, a decrease in efficiency of more than two orders of magnitude. In a high-stakes scenario where accuracy is paramount, pursuing performance gains at a rate of 6701.8 $\frac{t}{p}$ may be an acceptable cost. For many real-world scenarios, increasing accuracy at a rate of 65.3 $\frac{t}{p}$ might be more reasonable. TC provides an intuitive way to compare the tradeoff between

[9] Token counts are estimated by $\frac{\text{num characters}}{4}$.
[10] We use the average of ratios when discussing each to give equal weighting to every observation.

Table 5. The percentage of total IO pairs removed from token usage statistics. Pairs were removed if the output was empty or the length of the output was more than 4 standard deviations from the mean.

Model	Benchmark	Percentage of Total IO Pairs Excluded					
		Vanilla IO	Vanilla Fewshot	Zeroshot CoT	Fewshot CoT	CoT-SC$_5$	CoT-SC$_{10}$
Llama 3.1 8B Instruct	BBH	2.53	8.59	2.90	5.04	4.09	4.00
	GSM8K	0.68	1.14	0.76	0.91	1.27	1.28
	MMLU	1.18	1.76	1.89	2.35	1.89	1.93
Qwen 2.5 14B Instruct	BBH	0.29	0.12	0.23	0.49	0.24	0.19
	GSM8K	0.30	0.45	0.23	0.23	0.36	0.09
	MMLU	0.07	0.65	0.20	0.46	0.20	0.38
Qwen 2.5 32B Instruct	BBH	0.22	0.09	0.26	0.29	0.16	0.10
	GSM8K	0.00	0.00	0.08	0.38	0.18	0.13
	MMLU	0.20	0.46	0.07	0.07	0.10	0.11

token usage and performance and allows for more informed prompting strategy and variable selection.

We perform an ablation study, found in Section 8.1, on the number of fewshot exemplars. We observe that incrementally increasing the number of fewshot examples follows a similar trend of diminishing returns as token usage increases.

All information for reproducing our results, as well as our verbatim results, are detailed in Sections 7.4 and 7.5.

7 Empirical Evaluation Details

7.1 Detailed Results

We provide detailed results from our experiments in Table 6. Note that the results presented for each prompting strategy, model, and benchmark combination are from a single execution with the number of samples noted in Table 9.

We removed empty outputs and outputs more than four standard deviations from the mean (e.g., instances where the LLM generated a looping output) from our token usage statistics. Such erroneous outputs were surprisingly common for Llama 3.1 8B Instruct. Details of the number of outputs removed from consideration are detailed in Table 5.

7.2 Additional Observations

While our experiments were simple, we made a number of interesting observations that may warrant further analysis in future work.

		Strategy	Avg. Tokens			Acc.*	Std. Error	Average TC
			In	Out	Total			
Llama 3.1 8B Instruct	BBH	Vanilla IO	172	221	393	51.1	0.56	7.70
		Vanilla 3-shot	420	226	646	35.4	0.54	18.28
		Zeroshot CoT	178	351	530	63.2	0.55	8.38
		3-shot CoT	876	335	1212	70.5	0.51	17.20
		CoT-SC$_5$	4378	1089	5468	72.7	0.50	75.19
		CoT-SC$_{10}$	8758	2177	10935	73.1	0.50	149.58
	GSM8K	Vanilla IO	122	162	284	77.3	1.15	3.68
		Vanilla 8-shot	639	147	787	79.3	1.12	9.93
		Zeroshot CoT	126	203	330	75.5	1.18	4.37
		8-shot CoT	654	145	800	83.4	1.02	9.60
		CoT-SC$_5$	3275	751	4026	85.7	0.96	46.96
		CoT-SC$_{10}$	6550	1499	8050	89.0	0.86	90.44
	MMLU	Vanilla IO	201	201	402	51.1	1.24	7.88
		Vanilla 4-shot	711	208	920	58.8	1.19	15.65
		Zeroshot CoT	207	342	550	69.4	1.13	7.93
		4-shot CoT	1008	182	1191	65.7	1.16	18.13
		CoT-SC$_5$	5037	955	5993	66.9	1.16	89.53
		CoT-SC$_{10}$	10076	1929	12006	69.5	1.12	172.76
Qwen 2.5 14B Instruct	BBH	Vanilla IO	134	197	332	63.5	0.51	5.24
		Vanilla 3-shot	379	143	523	62.0	0.55	8.44
		Zeroshot CoT	140	254	395	79.7	0.42	4.96
		3-shot CoT	830	195	1026	81.2	0.43	12.64
		CoT-SC$_5$	4155	1011	5166	82.8	0.41	62.37
		CoT-SC$_{10}$	8310	2028	10339	83.7	0.40	123.52
	GSM8K	Vanilla IO	101	180	281	81.2	1.08	3.47
		Vanilla 8-shot	618	150	768	83.5	1.02	9.21
		Zeroshot CoT	104	194	299	81.9	1.06	3.65
		8-shot CoT	633	125	759	87.6	0.91	8.67
		CoT-SC$_5$	3168	602	3770	88.7	0.87	42.51
		CoT-SC$_{10}$	6337	1221	7559	89.7	0.84	84.28
	MMLU	Vanilla IO	162	162	325	76.4	1.05	4.26
		Vanilla 4-shot	673	130	804	78.4	1.03	10.25
		Zeroshot CoT	169	347	516	81.4	0.97	6.35
		4-shot CoT	965	163	1129	81.8	0.96	13.81
		CoT-SC$_5$	4829	883	5713	82.5	0.95	69.26
		CoT-SC$_{10}$	9650	1743	11394	82.9	0.94	137.47
Qwen 2.5 32B Instruct	BBH	Vanilla IO	134	201	336	63.4	0.50	5.30
		Vanilla 3-shot	379	144	523	71.2	0.49	7.36
		Zeroshot CoT	141	242	383	82.3	0.39	4.66
		3-shot CoT	830	182	1012	87.2	0.37	11.62
		CoT-SC$_5$	4153	938	5092	88.5	0.36	57.56
		CoT-SC$_{10}$	8308	1883	10191	88.8	0.35	114.75
	GSM8K	Vanilla IO	101	197	298	83.4	1.02	3.58
		Vanilla 8-shot	618	192	810	85.2	0.98	9.51
		Zeroshot CoT	104	199	304	84.1	1.01	3.62
		8-shot CoT	633	135	769	89.4	0.85	8.60
		CoT-SC$_5$	3168	687	3856	89.8	0.83	42.96
		CoT-SC$_{10}$	6337	1373	7711	89.8	0.83	85.90
	MMLU	Vanilla IO	162	249	412	65.3	1.18	6.32
		Vanilla 4-shot	671	194	865	77.3	1.05	11.21
		Zeroshot CoT	169	357	526	84.5	0.89	6.23
		4-shot CoT	966	186	1153	85.2	0.87	13.53
		CoT-SC$_5$	4827	1013	5840	85.4	0.87	68.37
		CoT-SC$_{10}$	9652	2030	11682	86.1	0.86	135.70

* Accuracy as a percentage.

Table 6. Detailed results from the empirical evaluation described in Section 3.2.

Model	$\text{Price}_{\text{input}}$ *	$\text{Price}_{\text{output}}$ *	Cost† (US$)		
			BBH	GSM8K	MMLU
GPT-4o	2.5	10.0	488.25	72.53	121.57
GPT-4o-mini	0.15	0.6	29.29	4.35	7.29
Claude 3.5 Sonnet	3.0	15.0	662.89	97.81	163.16
Claude 3.5 Haiku	0.8	4.0	176.77	26.08	43.51

* Prices are in $\frac{\text{US\$}}{\text{1M Tokens}}$.

† Cost to run all prompting strategies on the given benchmark.

Table 7. Cost estimates for recreating the empirical evaluation with common commercial models.

The initial motivation behind fewshot prompting strategies was to define a pattern that the LLM would then follow in its answer (Brown et al. 2020; Wei et al. 2022). For Fewshot CoT, this pattern included a reasoning chain that led to the right answer (Wei et al. 2022). Now that LLMs are more capable and often aligned with human preferences after training, their default response to a question is to explain their reasoning before providing a response. This results in high output token usage even for Vanilla IO and Zeroshot CoT. Interestingly, the reasoning chains (averaged, per benchmark) that the LLMs produced for Zeroshot CoT were longer in every instance than the ones produced for Fewshot CoT, in some cases more than twice as long. Nonetheless, Fewshot CoT yielded accuracy improvements for nearly every benchmark. This supports the idea that the in-context learning of correct reasoning chains does positively influence the correctness of the generated reasoning chain, even if the LLM's default output is constrained.

That trend does not apply to Vanilla IO and Vanilla Fewshot, however. While Vanilla Fewshot outperforms Vanilla IO in almost every benchmark, the output token usage is less in most instances. This is likely caused by the pattern that is matched during Vanilla Fewshot, where the example outputs are the minimum number of tokens to convey the answer (e.g., for multiple-choice: "(A)").

We also observed how the quality of the fewshot reasoning chains provided as a part of Fewshot CoT affected performance. While Fewshot CoT yielded modest accuracy gains over Zeroshot CoT for BBH (4.6%) and GSM8K (6.3%), it actually registered a 0.9% accuracy *loss* on MMLU. This prompted an investigation into the CoT fewshot exemplars included in LM Evaluation Harness (Gao et al. 2023). The reasoning chains were significantly shorter than for the other two benchmarks. Interestingly, recent work on using concise reasoning chains, such as Constrained-CoT (Nayab et al. 2024) and Concise-CoT (Renze and Guven 2024), has demonstrated performance improvements from shorter chains of thought. A potentially insightful future work could explore how the form and content of intentionally concise chains of thought influence LLM performance to explain this discrepancy.

Model*	Max Context Length	Max Gen. Tokens	Temperature
meta-llama/Llama-3.1-8B-Instruct	128000	16384	0.0 (0.5 for CoT-SC)
Qwen/Qwen2.5-14B-Instruct	128000	8192	0.0 (0.5 for CoT-SC)
Qwen/Qwen2.5-32B-Instruct	128000	8192	0.0 (0.5 for CoT-SC)

* Models were sourced from `https://huggingface.co/models`.

Table 8. Model configurations used for the empirical evaluation. Additional hyperparameters are found in the accompanying Supplementary Materials.

7.3 Cost Estimates for Commercial Models

The cost estimates found in Table 7 are derived from the pricing pages for Anthropic[11] and OpenAI[12], accessed on January 31, 2025. We do not include the effects of prompt caching.

7.4 Reproducibility

All results[13], as produced by LM Evaluation Harness, (e.g., LLM inputs and outputs, hyperparameters, runtimes, model configurations, etc.) are found at the following URL: [link redacted].

All code used to run the evaluations for this paper is found at the following GitHub repository: [link redacted]. Although we used LM Evaluation Harness, we link our fork[14] as significant bug fixes had to be made to get the framework to function as expected. Despite the bugs we encountered, we encourage others to support this open-source project that promotes reproducible results for LLM projects.

7.5 Configurations

7.5.1 Models We include details of model configurations in Table 8. All models were sourced from Hugging Face[15]. Where required, the authors complied with the necessary terms and conditions for gated models.

7.5.2 Benchmarks We include the benchmark configurations used for our experiments in Table 9. The underlying datasets for BBH, GSM8K, and MMLU were sourced from Hugging Face[16]. While the fewshot examples for BBH were drawn from the same split used for evaluation, care was taken to ensure that the fewshot examples did not overlap with the target question.

We noted above that we selected general-purpose LLM benchmarks for our experiments but recognize that GSM8K could be seen as targeted towards the math domain. While it is true that GSM8K is composed of questions that require basic math, the reasoning capabilities and basic world knowledge it probes are generally applicable.

[11] `https://anthropic.com/pricing#anthropic-api`
[12] `https://openai.com/api/pricing/`
[13] Provided under the CC-BY-4.0 license.
[14] Under the same MIT license as LM Evaluation Harness.
[15] `https://huggingface.co/models`
[16] `https://huggingface.co/datasets`

Table 9. Benchmark configurations used for the empirical evaluation.

Benchmark	Split	Fewshot Split	# Samples
BBH	test	test*	6511
GSM8K	test	train	1319
MMLU	validation	dev	1531

* Sampled fewshot examples did not overlap with the current benchmark question.

The focus of the benchmark is "properly interpreting a question and reasoning through the steps to solve it" (Cobbe et al. 2021), not to evaluate advanced math skills. We believe this justifies GSM8K's inclusion as a general-purpose benchmark.

We rely on the steps taken by the authors in creating BBH (Suzgun et al. 2023; Srivastava et al. 2023), GSM8K (Cobbe et al. 2021), and MMLU (Hendrycks et al. 2021) to ensure ethical dataset creation, including the mitigation of bias, offensive content, and personally identifying information. We refer readers to those papers for additional information on the breadth of representation in those benchmarks, such as demographic groups.

7.5.3 Prompts To limit the inadvertent effects of linguistic prompt engineering (see Salinas and Morstatter (2024)), we avoided making adjustments to the prompts as they were constructed by default in LM Evaluation Harness, for consistency. As such, we did not include output formatting instructions and used the default prompt template for each model. We excluded output formatting instructions for two reasons: (1) to not interfere with the LLM's natural output for the benchmark question and (2) to allow the benefits of implicit formatting for fewshot strategies. We include an example of a raw text input and output pair for Qwen 2.5 14B on MMLU using Zeroshot CoT below:

Input

```
<|im_start|>system\nThe following are multiple choice
    questions (with answers) about high school psychology.\
    n\n<|im_end|>\n<|im_start|>user\nQuestion: Why are
    psychoanalysts sometimes interested in talking with a
    patient about dreams?\n(A) Psychoanalysts discovered
    that neurotransmitter abnormalities sometimes cause
    dream disturbances.; (B) Since all people are striving
    for self-actualization, psychoanalysts look at the
    cognitive obstacles in dreams.; (C) Nonproductive
    counterconditioning behaviors are sometimes revealed in
     dreams.; (D) Some psychoanalysts believe that dream
    symbols represent unconscious conflicts.\nAnswer: Let's
     think step by step.<|im_end|>\n<|im_start|>assistant\n
```

Output

```
To determine the correct answer, let's analyze each option
    in the context of psychoanalytic theory:\n\n(A)
    Psychoanalysts discovered that neurotransmitter
    abnormalities sometimes cause dream disturbances.\n-
    This statement is more aligned with biological
    psychology rather than psychoanalysis. Psychoanalysis
    does not focus on neurotransmitters or biological
    abnormalities as the primary cause of dream
    disturbances.\n\n(B) Since all people are striving for
    self-actualization, psychoanalysts look at the
    cognitive obstacles in dreams.\n- This statement aligns
     more with humanistic psychology, particularly the work
     of Carl Rogers and Abraham Maslow, who emphasized self
    -actualization. Psychoanalysis, as developed by Freud,
    does not focus on self-actualization or cognitive
    obstacles in dreams.\n\n(C) Nonproductive
    counterconditioning behaviors are sometimes revealed in
     dreams.\n- This statement is more aligned with
    behaviorism and the concept of counterconditioning,
    which is not a primary focus of psychoanalysis.
    Psychoanalysis does not typically discuss
    counterconditioning in the context of dreams.\n\n(D)
    Some psychoanalysts believe that dream symbols
    represent unconscious conflicts.\n- This statement
    aligns with the psychoanalytic perspective,
    particularly Freud's theory of dreams. Freud believed
    that dreams are a manifestation of unconscious desires
    and conflicts, often symbolized in dream content.\n\
    nGiven this analysis, the correct answer is:\n\n(D)
    Some psychoanalysts believe that dream symbols
    represent unconscious conflicts.
```

7.5.4 Compute Experiments were run on A100 40GB, A100 80GB, and H100 GPUs, as availability permitted. Configurations were limited to 1 or 2 GPUs per experiment. The experiments were run in tensor parallel or data parallel configurations, depending on the size of the model and the number of GPUs used. The exact configurations per experiment, as well as exact wall times, are detailed in the reproducibility materials referenced in Section 7.4. The approximate number of GPU hours across all GPUs was 95.

7.6 Licensing

We use a number of open-source artifacts in this work. We list the licenses for each in Table 10. We verify that our usage was in accordance with the projects' licenses.

Table 10. Licenses for the artifacts used in this work.

Artifact	License	Notes
Llama 3.1 8B Instruct	Llama 3.1 Community License	Copyright © Meta Platforms, Inc.
Qwen 2.5 14B Instruct	Qwen LICENSE AGREEMENT	Version: September 19, 2024
Qwen 2.5 32B Instruct	Qwen LICENSE AGREEMENT	Version: September 19, 2024
LM Evaluation Harness	MIT License	Copyright (c) 2020 EleutherAI
BBH	MIT License	Copyright (c) 2022 suzgunmirac
GSM8K	MIT License	Copyright (c) 2021 OpenAI
MMLU	MIT License	Copyright (c) 2020 Dan Hendrycks

Table 11. Average marginal TCs ($\frac{\Delta tokens}{\Delta accuracy}$) calculated between 0 and 3 exemplars and 3 and 8 exemplars for the ablation study on the number of fewshot exemplars.

Fewshot Range	Marginal TC ($\frac{t}{p}$)	
	Vanilla Fewshot	Fewshot CoT
0-3	117.2	30.5
3-8	1621.5	553.8

8 Additional Studies

8.1 Ablation Study on the Number of Fewshot Exemplars

We present the results from our ablation study on the number of fewshot examples in Figure 3, with detailed results in Table 12. For this experiment, we used the Vanilla Fewshot and Fewshot CoT prompting strategies with the number of exemplars ranging from 0 to 8. As can be seen in Figure 3, the results are noisier but there is a clear trend of diminishing returns as the number of fewshot exemplars increases. To demonstrate this in a way that mitigates the noise, we examine the results piecewise, comparing the average marginal TC between 0 and 3 exemplars and 3 and 8 exemplars. Those values are found in Table 11, for concision. The marginal TC between 0 and 3 fewshot exemplars is, for both Vanilla Fewshot and Fewshot CoT, over an order of magnitude less than between 3 and 8. This indicates a significant decrease in efficiency as token usage increases, which corroborates our observations in Section 6.

8.2 Model Size

We included two models from the Qwen 2.5 family to facilitate a discussion on the impact of model size on our observed trends. Selecting two models from the same family ensures that potentially confounding variables, such as differences in training, data collection, and alignment, are presumably kept the same. We present the results from our experiments in Figure 4[17].

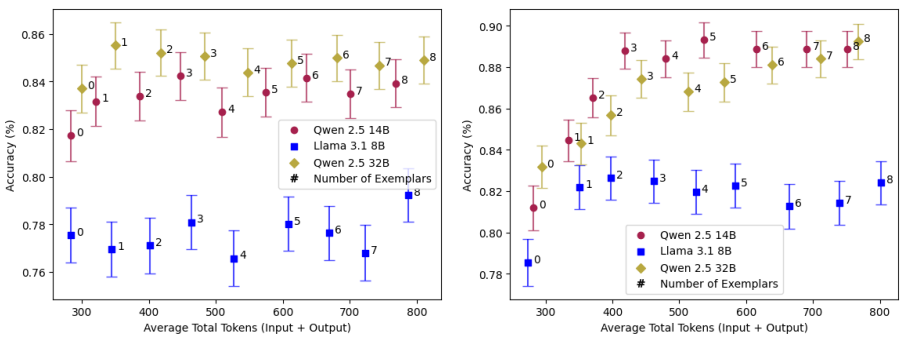
We observe that the trend towards diminishing accuracy returns for increased token usage is consistent between the 14B and 32B models. As expected, the 32B model

[17]These plots are identical to those from Figure 2 but with Llama 3.1 8B excluded, for ease of comparison.

Table 12. Detailed results from the ablation study on the number of fewshot exemplars. Token counts represent the total token usage (input and output).

	# Exemplars	Llama 3.1 8B Instruct			Qwen 2.5 14B Instruct			Qwen 2.5 32B Instruct		
		Tokens	**Acc.***	**SE†**	**Tokens**	**Acc.***	**SE†**	**Tokens**	**Acc.***	**SE†**
Vanilla Fewshot	0	283.8	77.6	1.15	283.0	81.7	1.06	299.4	83.7	1.02
	1	344.5	77.0	1.16	321.1	83.2	1.03	349.6	85.5	0.97
	2	401.6	77.1	1.16	386.6	83.4	1.02	418.0	85.2	0.98
	3	463.4	78.1	1.14	447.1	84.2	1.00	483.1	85.1	0.98
	4	526.3	76.6	1.17	508.8	82.7	1.04	547.9	84.4	1.00
	5	607.9	78.0	1.14	574.9	83.5	1.02	613.2	84.8	0.99
	6	668.8	77.6	1.15	635.1	84.2	1.01	680.4	85.0	0.98
	7	722.6	76.8	1.16	700.9	83.5	1.02	743.7	84.7	0.99
	8	787.3	79.2	1.12	768.9	83.9	1.01	810.1	84.9	0.99
Fewshot CoT	0	273.2	78.5	1.13	281.7	81.2	1.08	294.2	83.2	1.03
	1	351.3	82.2	1.05	334.1	84.5	1.00	353.7	84.3	1.00
	2	398.1	82.6	1.04	370.5	86.5	0.94	397.7	85.7	0.97
	3	462.6	82.5	1.05	418.9	88.8	0.87	443.1	87.4	0.91
	4	524.8	82.0	1.06	479.6	88.4	0.88	514.2	86.8	0.93
	5	584.3	82.3	1.05	537.4	89.3	0.85	567.8	87.3	0.92
	6	664.6	81.3	1.07	615.6	88.9	0.87	638.9	88.1	0.89
	7	739.2	81.4	1.07	691.0	88.9	0.87	711.0	88.4	0.88
	8	801.3	82.4	1.05	751.9	88.9	0.87	768.3	89.2	0.85

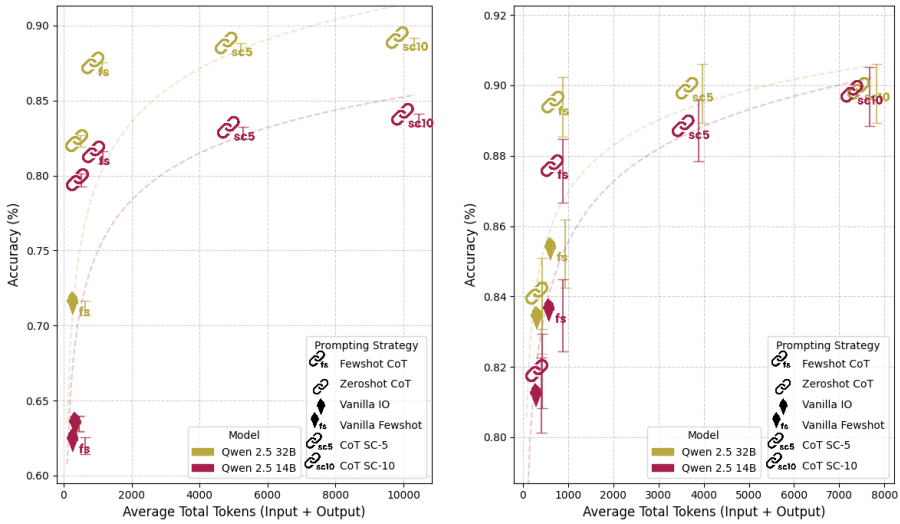
* Accuracy as a percentage.
† Standard error.



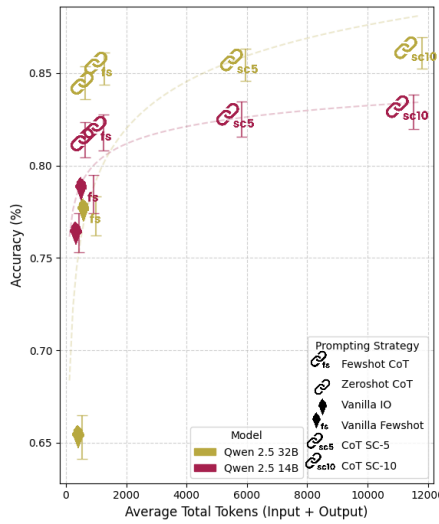
(A) Vanilla Fewshot **(B) Fewshot CoT**

Figure 3. Accuracy and total token usage for the ablation study on the number of fewshot exemplars on the GSM8K benchmark. Standard error bars are included.

generally outperforms the 14B model. However, we note some instances for prompting strategies that consume fewer tokens, such as Vanilla IO and Fewshot, where the 14B model outperforms the larger one. This suggests that larger models may be able to use additional tokens more effectively, as reflected in the consistently higher accuracy for Fewshot CoT and CoT-SC, but struggle to perform better than smaller models when fewer tokens are provided as context. This provides a promising route for future work.



(A) BBH **(B) GSM8K**



(C) MMLU

Figure 4. Accuracy and total token usage information for Qwen 2.5 14B and Qwen 2.5 32B from the empirical evaluation. The trend lines reflect the rapid growth of TC for these prompting strategies.

9 Limitations

To focus the contribution of this work, we make a number of thoughtful concessions, such as the models, benchmarks, and prompting strategies we use. We explicitly justify

the most relevant limitations alongside the material they qualify—such as in Sections 3.1 and 3.2—and note other minor limitations here to further demonstrate the purpose and scope of this work. Despite the limitations, we maintain that the core premise of this work—demonstrating how to incorporate token usage into prompting strategy evaluation—remains broadly applicable beyond our experiments.

Prompting Strategies. In Section 1, we narrow the scope of this work to a subset of the broader prompt engineering landscape: formalized prompting strategies. As detailed in Sections 7.5 and 8, we strive to control for factors extraneous to the minimal instantiation of each prompting strategy. We do so to isolate the effects of token usage dictated by the prompting strategies and the resultant benchmark performance to demonstrate how a significant area of research has become prone to inefficiencies due to a lack of relevant metrics. Although we focus on formalized prompting strategies, Big-O_{tok} and TC are useful metrics for other aspects of prompt engineering as well, such as linguistic and language choices.

As noted in Sections 1 and 2, the focus of this work is on incorporating token usage into prompting strategy evaluation and analysis. We do not aim to solve issues of prompting strategy efficiency but instead provide methods for quantifying it, both for researchers and practitioners. To maintain that scope, we do not explore nor propose specific methods of optimizing prompts and instead focus on introducing insightful metrics and demonstrating their utility in practice.

To do so, we approximate tokens as $\frac{\text{num characters}}{4}$. The reason we use this approximation instead of the actual token counts is because tokenizer- and model-specific idiosyncrasies can result in numerous valid tokenizations of the same text. For our general-purpose benchmarks in English, it is a well-established approximation[18]. However, we recognize that this approximation may not apply to all models, languages, and domains.

The results we present here represent a single thread of prompting strategy evolution; we expect the results to follow a predictable pattern (e.g., diminishing accuracy returns) because there are no drastic changes to the principles underlying our selection of prompting strategies. It is likely that fundamentally different prompting strategies, such as Least-to-Most (Zhou et al. 2023) or Algorithm of Thoughts (Sel et al. 2024), would not follow our observed trend lines and thus we do not make any generalized claims based on our observations.

Models. Depending on the data collection and processing methods used by LLM creators, benchmark data leakage could influence our empirical results, as demonstrated by Mirzadeh et al. (2025). LLMs that were trained on data that included BBH question-answer pairs, for example, could be influenced by prompting strategies to a lesser degree than those that were not. We use models from multiple sources in conjunction with multiple benchmarks to mitigate the potential effects of data leakage.

[18]Common commercial LLM providers also suggest this estimate despite models trained in different years, across multiple languages, and on diverse domains (e.g., Google: `https://ai.google.dev/gemini-api/docs/tokens`; OpenAI: `https://platform.openai.com/tokenizer`).

In this work, we exclusively consider autoregressive, text-to-text ("traditional") LLMs because most prompting strategies are optimized for them. We recognize, however, that multimodal and, more recently, reasoning LLMs have become increasingly relevant (DeepSeek-AI et al. 2025; Yin et al. 2024; Caffagni et al. 2024). We exclude them from our investigation here for a number of reasons: (1) prompt engineering specific to such models is a nascent field and distinct from prompt engineering for traditional LLMs[19] Wu et al. (2024); (2) due to the recency of reasoning models, there are very few (especially open-source) models available; and (3) the inclusion of multimodal benchmarks and the use of reasoning models would drastically increase the compute required to undertake a similar study. Nevertheless, we maintain that the efficiency-aware metrics explored here remain relevant to such models since they function simply on tokenized inputs and outputs. We see prompting strategies designed for multimodal and, particularly, reasoning LLMs as a significant avenue for future research and are hopeful that Big-O_{tok} and TC will be incorporated into their development.

We note that the reasoning model landscape has evolved rapidly since this work was first conducted. Models such as OpenAI's o1 and o3, DeepSeek-R1 (DeepSeek-AI et al. 2025), and their distilled variants are now widely deployed and routinely produce reasoning traces spanning thousands to tens of thousands of tokens per query. This development makes the efficiency arguments presented here, if anything, more urgent: the diminishing returns we observe for traditional prompting strategies are likely amplified when reasoning token budgets grow by an additional order of magnitude. Extending Big-O_{tok} and TC to formally characterize reasoning model token usage remains an important direction for future work.

While we believe it a fair comparison that lends itself to real-world deployment, we recognize that running our selection of prompting strategies with relatively small LLMs on a subset of benchmarks does not fully reflect the performance of the strategies under all conditions. It is very likely, for example, that the CoT prompting strategy (Wei et al. 2022) would be leveraged better by Llama 3.1 405B than by Llama 3.1 8B, due to the former exhibiting superior reasoning capabilities (Dubey et al. 2024). The strength of an LLM may magnify the disparity between a specific prompting strategy and others.

Benchmarks. Similarly, while we attempt to cover a breadth of domains in our selection of benchmarks, this selection may fail to highlight the strengths of some prompting strategies over others in certain domains. Our purpose is not to rank prompting strategies but to analytically explore the tradeoffs between token usage and benchmark accuracy. We recognize, however, that our selection of benchmarks may fail to cover the strength of a particular prompting strategy entirely, which may paint it in a worse light than it deserves. To mitigate this point, we focus on generalist prompting strategies and benchmarks and detail our selection processes in Sections 3.1 and 3.2.

[19]For reasoning models, some commercial providers even advise against the use of established prompting strategies (see `https://platform.openai.com/docs/guides/reasoning-best-practices`).

A common issue in LLM benchmarking is reliable answer extraction. Often, regular expressions are used, which are not robust to the unpredictable output formats an LLM may generate. We rely largely on the extraction methods from LM Evaluation Harness but observe certain inaccuracies in answer extraction. This is an open problem in LLM research (Yu et al. 2025). For this project, we rely on consistency in answer extraction methods between experiments but recognize that certain correct answers may be marked incorrectly.

Big-O_{tok}. A minor limitation of Big-O_{tok} is a lack of differentiation between input and output tokens. It is inherently more expensive for an LLM to generate output tokens than to process input tokens due to its autoregressive nature, a fact that is reflected in the pricing structure of common commercial APIs[20]. We considered that differentiating between input and output tokens would have introduced excessive complexity to Big-O_{tok}, particularly since it does not serve as a precise measure. We instead consider combined token usage (input **and** output) to provide a holistic view of token consumption.

Another limitation of Big-O_{tok} is the decreased range of values for prompting strategy variables. While variables in traditional Big-O analyses often span many orders of magnitude, variables in prompting strategies tend to be lower (typically $<= 100$) (Brown et al. 2020; Wang et al. 2023b). As noted in Section 6, the relatively low values for those variables could result in extraneous factors (e.g., chat templates, model idiosyncrasies, etc.) limiting their impact on overall token usage. However, we observe that, even for extremely low values (e.g., $k = 3$ fewshot examples for BBH), the token usages from each experiment align with the expected Big-O_{tok} token complexity classes. Although not a precise measure, Big-O_{tok} can still provide useful insights for expected prompting strategy token usage, even for small values.

10 Conclusion

Token usage represents a significant, yet often underrepresented, component of prompting strategy evaluation. To facilitate the comparison of token usage between distinct prompting strategies, we present Big-O_{tok} token complexity and substantiate it empirically by comparing predicted token usages to those derived from our experiments. We analyze our experiments in terms of Token Cost and use it to demonstrate the tradeoffs between token usage and performance. Our analyses demonstrate the importance of including token usage in prompting strategy evaluation and validate Big-O_{tok} and Token Cost as viable means of doing so.

Looking ahead, the proliferation of reasoning models and agentic systems that consume vastly more tokens per query than the strategies studied here makes the case for efficiency-aware evaluation even more pressing. We hope that Big-O_{tok} and TC will serve as useful building blocks for this broader effort.

[20] E.g., `https://openai.com/api/pricing/`

Broader Impact Statement

LLM usage incurs real-world monetary and environmental costs (Schwartz et al. 2020; Dhar 2020; Wu et al. 2022). This work promotes the consideration of token usage in prompting strategy development and evaluation to increase the long-term efficiency of LLM inference.

Use of AI

There was limited use of AI in the research and writing of this work. For writing, ChatGPT[21] was used to rephrase several sentences (<5) and to help debug LaTeX and Kubernetes errors. GitHub Copilot[22] was used to help generate some of the plots. Some grammatical suggestions from Writefull's Overleaf integration[23] were considered and included. All AI outputs were thoroughly reviewed by the authors prior to inclusion.

References

Biderman S, Schoelkopf H, Sutawika L, Gao L, Tow J, Abbasi B, Aji AF, Ammanamanchi PS, Black S, Clive J, DiPofi A, Etxaniz J, Fattori B, Forde JZ, Foster C, Hsu J, Jaiswal M, Lee WY, Li H, Lovering C, Muennighoff N, Pavlick E, Phang J, Skowron A, Tan S, Tang X, Wang KA, Winata GI, Yvon F and Zou A (2024) Lessons from the trenches on reproducible evaluation of language models. URL https://arxiv.org/abs/2405.14782.

Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler D, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I and Amodei D (2020) Language models are few-shot learners. In: Larochelle H, Ranzato M, Hadsell R, Balcan M and Lin H (eds.) *Advances in Neural Information Processing Systems*, volume 33. Curran Associates, Inc., pp. 1877–1901. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Caffagni D, Cocchi F, Barsellotti L, Moratelli N, Sarto S, Baraldi L, Baraldi L, Cornia M and Cucchiara R (2024) The revolution of multimodal large language models: A survey. In: Ku LW, Martins A and Srikumar V (eds.) *Findings of the Association for Computational Linguistics: ACL 2024*. Bangkok, Thailand: Association for Computational Linguistics, pp. 13590–13618. DOI:10.18653/v1/2024.findings-acl.807. URL https://aclanthology.org/2024.findings-acl.807/.

Chen L, Zaharia M and Zou J (2023) Frugalgpt: How to use large language models while reducing cost and improving performance. URL https://arxiv.org/abs/2305.05176.

Chu Z, Chen J, Chen Q, Yu W, He T, Wang H, Peng W, Liu M, Qin B and Liu T (2024) Navigate through enigmatic labyrinth a survey of chain of thought reasoning: Advances, frontiers

[21] https://chatgpt.com/
[22] https://github.com/features/copilot
[23] https://www.overleaf.com/learn/how-to/Writefull_integration

and future. In: Ku LW, Martins A and Srikumar V (eds.) *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Bangkok, Thailand: Association for Computational Linguistics, pp. 1173–1203. DOI:10.18653/v1/2024.acl-long.65. URL https://aclanthology.org/2024.acl-long.65.

Cobbe K, Kosaraju V, Bavarian M, Hilton J, Nakano R, Hesse C and Schulman J (2021) Training verifiers to solve math word problems.

DeepSeek-AI, Guo D, Yang D, Zhang H, Song J, Zhang R, Xu R, Zhu Q, Ma S, Wang P, Bi X, Zhang X, Yu X, Wu Y, Wu ZF, Gou Z, Shao Z, Li Z, Gao Z, Liu A, Xue B, Wang B, Wu B, Feng B, Lu C, Zhao C, Deng C, Zhang C, Ruan C, Dai D, Chen D, Ji D, Li E, Lin F, Dai F, Luo F, Hao G, Chen G, Li G, Zhang H, Bao H, Xu H, Wang H, Ding H, Xin H, Gao H, Qu H, Li H, Guo J, Li J, Wang J, Chen J, Yuan J, Qiu J, Li J, Cai JL, Ni J, Liang J, Chen J, Dong K, Hu K, Gao K, Guan K, Huang K, Yu K, Wang L, Zhang L, Zhao L, Wang L, Zhang L, Xu L, Xia L, Zhang M, Zhang M, Tang M, Li M, Wang M, Li M, Tian N, Huang P, Zhang P, Wang Q, Chen Q, Du Q, Ge R, Zhang R, Pan R, Wang R, Chen RJ, Jin RL, Chen R, Lu S, Zhou S, Chen S, Ye S, Wang S, Yu S, Zhou S, Pan S, Li SS, Zhou S, Wu S, Ye S, Yun T, Pei T, Sun T, Wang T, Zeng W, Zhao W, Liu W, Liang W, Gao W, Yu W, Zhang W, Xiao WL, An W, Liu X, Wang X, Chen X, Nie X, Cheng X, Liu X, Xie X, Liu X, Yang X, Li X, Su X, Lin X, Li XQ, Jin X, Shen X, Chen X, Sun X, Wang X, Song X, Zhou X, Wang X, Shan X, Li YK, Wang YQ, Wei YX, Zhang Y, Xu Y, Li Y, Zhao Y, Sun Y, Wang Y, Yu Y, Zhang Y, Shi Y, Xiong Y, He Y, Piao Y, Wang Y, Tan Y, Ma Y, Liu Y, Guo Y, Ou Y, Wang Y, Gong Y, Zou Y, He Y, Xiong Y, Luo Y, You Y, Liu Y, Zhou Y, Zhu YX, Xu Y, Huang Y, Li Y, Zheng Y, Zhu Y, Ma Y, Tang Y, Zha Y, Yan Y, Ren ZZ, Ren Z, Sha Z, Fu Z, Xu Z, Xie Z, Zhang Z, Hao Z, Ma Z, Yan Z, Wu Z, Gu Z, Zhu Z, Liu Z, Li Z, Xie Z, Song Z, Pan Z, Huang Z, Xu Z, Zhang Z and Zhang Z (2025) Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. URL https://arxiv.org/abs/2501.12948.

Dhar P (2020) The carbon impact of artificial intelligence.

Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, Letman A, Mathur A, Schelten A, Yang A, Fan A, Goyal A, Hartshorn A, Yang A, Mitra A, Sravankumar A, Korenev A, Hinsvark A, Rao A, Zhang A, Rodriguez A, Gregerson A, Spataru A, Roziere B, Biron B, Tang B, Chern B, Caucheteux C, Nayak C, Bi C, Marra C, McConnell C, Keller C, Touret C, Wu C, Wong C, Ferrer CC, Nikolaidis C, Allonsius D, Song D, Pintz D, Livshits D, Esiobu D, Choudhary D, Mahajan D, Garcia-Olano D, Perino D, Hupkes D, Lakomkin E, AlBadawy E, Lobanova E, Dinan E, Smith EM, Radenovic F, Zhang F, Synnaeve G, Lee G, Anderson GL, Nail G, Mialon G, Pang G, Cucurell G, Nguyen H, Korevaar H, Xu H, Touvron H, Zarov I, Ibarra IA, Kloumann I, Misra I, Evtimov I, Copet J, Lee J, Geffert J, Vranes J, Park J, Mahadeokar J, Shah J, van der Linde J, Billock J, Hong J, Lee J, Fu J, Chi J, Huang J, Liu J, Wang J, Yu J, Bitton J, Spisak J, Park J, Rocca J, Johnstun J, Saxe J, Jia J, Alwala KV, Upasani K, Plawiak K, Li K, Heafield K, Stone K, El-Arini K, Iyer K, Malik K, Chiu K, Bhalla K, Rantala-Yeary L, van der Maaten L, Chen L, Tan L, Jenkins L, Martin L, Madaan L, Malo L, Blecher L, Landzaat L, de Oliveira L, Muzzi M, Pasupuleti M, Singh M, Paluri M, Kardas M, Oldham M, Rita M, Pavlova M, Kambadur M, Lewis M, Si M, Singh MK, Hassan M, Goyal N, Torabi N, Bashlykov N, Bogoychev N, Chatterji N, Duchenne O, Çelebi O, Alrassy P, Zhang P, Li P, Vasic P, Weng P, Bhargava P, Dubal P, Krishnan P, Koura PS, Xu P, He Q, Dong Q, Srinivasan

R, Ganapathy R, Calderer R, Cabral RS, Stojnic R, Raileanu R, Girdhar R, Patel R, Sauvestre R, Polidoro R, Sumbaly R, Taylor R, Silva R, Hou R, Wang R, Hosseini S, Chennabasappa S, Singh S, Bell S, Kim SS, Edunov S, Nie S, Narang S, Raparthy S, Shen S, Wan S, Bhosale S, Zhang S, Vandenhende S, Batra S, Whitman S, Sootla S, Collot S, Gururangan S, Borodinsky S, Herman T, Fowler T, Sheasha T, Georgiou T, Scialom T, Speckbacher T, Mihaylov T, Xiao T, Karn U, Goswami V, Gupta V, Ramanathan V, Kerkez V, Gonguet V, Do V, Vogeti V, Petrovic V, Chu W, Xiong W, Fu W, Meers W, Martinet X, Wang X, Tan XE, Xie X, Jia X, Wang X, Goldschlag Y, Gaur Y, Babaei Y, Wen Y, Song Y, Zhang Y, Li Y, Mao Y, Coudert ZD, Yan Z, Chen Z, Papakipos Z, Singh A, Grattafiori A, Jain A, Kelsey A, Shajnfeld A, Gangidi A, Victoria A, Goldstand A, Menon A, Sharma A, Boesenberg A, Vaughan A, Baevski A, Feinstein A, Kallet A, Sangani A, Yunus A, Lupu A, Alvarado A, Caples A, Gu A, Ho A, Poulton A, Ryan A, Ramchandani A, Franco A, Saraf A, Chowdhury A, Gabriel A, Bharambe A, Eisenman A, Yazdan A, James B, Maurer B, Leonhardi B, Huang B, Loyd B, Paola BD, Paranjape B, Liu B, Wu B, Ni B, Hancock B, Wasti B, Spence B, Stojkovic B, Gamido B, Montalvo B, Parker C, Burton C, Mejia C, Wang C, Kim C, Zhou C, Hu C, Chu CH, Cai C, Tindal C, Feichtenhofer C, Civin D, Beaty D, Kreymer D, Li D, Wyatt D, Adkins D, Xu D, Testuggine D, David D, Parikh D, Liskovich D, Foss D, Wang D, Le D, Holland D, Dowling E, Jamil E, Montgomery E, Presani E, Hahn E, Wood E, Brinkman E, Arcaute E, Dunbar E, Smothers E, Sun F, Kreuk F, Tian F, Ozgenel F, Caggioni F, Guzmán F, Kanayet F, Seide F, Florez GM, Schwarz G, Badeer G, Swee G, Halpern G, Thattai G, Herman G, Sizov G, Guangyi, Zhang, Lakshminarayanan G, Shojanazeri H, Zou H, Wang H, Zha H, Habeeb H, Rudolph H, Suk H, Aspegren H, Goldman H, Damlaj I, Molybog I, Tufanov I, Veliche IE, Gat I, Weissman J, Geboski J, Kohli J, Asher J, Gaya JB, Marcus J, Tang J, Chan J, Zhen J, Reizenstein J, Teboul J, Zhong J, Jin J, Yang J, Cummings J, Carvill J, Shepard J, McPhie J, Torres J, Ginsburg J, Wang J, Wu K, U KH, Saxena K, Prasad K, Khandelwal K, Zand K, Matosich K, Veeraraghavan K, Michelena K, Li K, Huang K, Chawla K, Lakhotia K, Huang K, Chen L, Garg L, A L, Silva L, Bell L, Zhang L, Guo L, Yu L, Moshkovich L, Wehrstedt L, Khabsa M, Avalani M, Bhatt M, Tsimpoukelli M, Mankus M, Hasson M, Lennie M, Reso M, Groshev M, Naumov M, Lathi M, Keneally M, Seltzer ML, Valko M, Restrepo M, Patel M, Vyatskov M, Samvelyan M, Clark M, Macey M, Wang M, Hermoso MJ, Metanat M, Rastegari M, Bansal M, Santhanam N, Parks N, White N, Bawa N, Singhal N, Egebo N, Usunier N, Laptev NP, Dong N, Zhang N, Cheng N, Chernoguz O, Hart O, Salpekar O, Kalinli O, Kent P, Parekh P, Saab P, Balaji P, Rittner P, Bontrager P, Roux P, Dollar P, Zvyagina P, Ratanchandani P, Yuvraj P, Liang Q, Alao R, Rodriguez R, Ayub R, Murthy R, Nayani R, Mitra R, Li R, Hogan R, Battey R, Wang R, Maheswari R, Howes R, Rinott R, Bondu SJ, Datta S, Chugh S, Hunt S, Dhillon S, Sidorov S, Pan S, Verma S, Yamamoto S, Ramaswamy S, Lindsay S, Lindsay S, Feng S, Lin S, Zha SC, Shankar S, Zhang S, Zhang S, Wang S, Agarwal S, Sajuyigbe S, Chintala S, Max S, Chen S, Kehoe S, Satterfield S, Govindaprasad S, Gupta S, Cho S, Virk S, Subramanian S, Choudhury S, Goldman S, Remez T, Glaser T, Best T, Kohler T, Robinson T, Li T, Zhang T, Matthews T, Chou T, Shaked T, Vontimitta V, Ajayi V, Montanez V, Mohan V, Kumar VS, Mangla V, Albiero V, Ionescu V, Poenaru V, Mihailescu VT, Ivanov V, Li W, Wang W, Jiang W, Bouaziz W, Constable W, Tang X, Wang X, Wu X, Wang X, Xia X, Wu X, Gao X, Chen Y, Hu Y, Jia Y, Qi Y, Li Y, Zhang Y, Zhang Y, Adi Y, Nam Y, Yu, Wang, Hao Y, Qian Y, He Y, Rait Z, DeVito Z, Rosnbrick Z, Wen Z, Yang Z and Zhao Z (2024) The llama 3

herd of models. URL https://arxiv.org/abs/2407.21783.

Fourrier C, Habib N, Lozovskaya A, Szafer K and Wolf T (2024) Open llm leaderboard v2. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard.

Gao L, Tow J, Abbasi B, Biderman S, Black S, DiPofi A, Foster C, Golding L, Hsu J, Le Noac'h A, Li H, McDonell K, Muennighoff N, Ociepa C, Phang J, Reynolds L, Schoelkopf H, Skowron A, Sutawika L, Tang E, Thite A, Wang B, Wang K and Zou A (2023) A framework for few-shot language model evaluation. DOI:10.5281/zenodo.10256836. URL https://zenodo.org/records/10256836.

Han T, Wang Z, Fang C, Zhao S, Ma S and Chen Z (2025) Token-budget-aware llm reasoning. In: *Findings of the Association for Computational Linguistics: ACL 2025*. pp. 24842–24855.

Hendrycks D, Burns C, Basart S, Zou A, Mazeika M, Song D and Steinhardt J (2021) Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)* .

Jiang H, Wu Q, Lin CY, Yang Y and Qiu L (2023) LLMLingua: Compressing prompts for accelerated inference of large language models. In: Bouamor H, Pino J and Bali K (eds.) *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, pp. 13358–13376. DOI:10.18653/v1/2023.emnlp-main.825. URL https://aclanthology.org/2023.emnlp-main.825/.

Kim J, Lee S, Hun Han S, Park S, Lee J, Jeong K and Kang P (2023) Which is better? exploring prompting strategy for LLM-based metrics. In: Deutsch D, Dror R, Eger S, Gao Y, Leiter C, Opitz J and Rücklé A (eds.) *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*. Bali, Indonesia: Association for Computational Linguistics, pp. 164–183. DOI:10.18653/v1/2023.eval4nlp-1.14. URL https://aclanthology.org/2023.eval4nlp-1.14/.

Knuth DE (1976) Big omicron and big omega and big theta. *SIGACT News* 8(2): 18–24. DOI:10.1145/1008328.1008329. URL https://doi.org/10.1145/1008328.1008329.

Kojima T, Gu SS, Reid M, Matsuo Y and Iwasawa Y (2022) Large language models are zero-shot reasoners. In: Oh AH, Agarwal A, Belgrave D and Cho K (eds.) *Advances in Neural Information Processing Systems*. URL https://openreview.net/forum?id=e2TBb5y0yFf.

Liu T, Wang Z, Miao J, Hsu I, Yan J, Chen J, Han R, Xu F, Chen Y, Jiang K et al. (2025) Budget-aware tool-use enables effective agent scaling. *arXiv preprint arXiv:2511.17006* .

Mirzadeh SI, Alizadeh K, Shahrokhi H, Tuzel O, Bengio S and Farajtabar M (2025) GSM-symbolic: Understanding the limitations of mathematical reasoning in large language models. In: *The Thirteenth International Conference on Learning Representations*. URL https://openreview.net/forum?id=AjXkRZIvjB.

Mu J, Li XL and Goodman N (2023) Learning to compress prompts with gist tokens. In: *Thirty-seventh Conference on Neural Information Processing Systems*. URL https://openreview.net/forum?id=2DtxPCL3T5.

Nayab S, Rossolini G, Buttazzo G, Manes N and Giacomelli F (2024) Concise thoughts: Impact of output length on llm reasoning and cost. URL https://arxiv.org/abs/2407.

19825.

Qwen, :, Yang A, Yang B, Zhang B, Hui B, Zheng B, Yu B, Li C, Liu D, Huang F, Wei H, Lin H, Yang J, Tu J, Zhang J, Yang J, Yang J, Zhou J, Lin J, Dang K, Lu K, Bao K, Yang K, Yu L, Li M, Xue M, Zhang P, Zhu Q, Men R, Lin R, Li T, Tang T, Xia T, Ren X, Ren X, Fan Y, Su Y, Zhang Y, Wan Y, Liu Y, Cui Z, Zhang Z and Qiu Z (2025) Qwen2.5 technical report. URL https://arxiv.org/abs/2412.15115.

Renze M and Guven E (2024) The benefits of a concise chain of thought on problem-solving in large language models. In: *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*. IEEE, pp. 476–483.

Salinas A and Morstatter F (2024) The butterfly effect of altering prompts: How small changes and jailbreaks affect large language model performance. In: Ku LW, Martins A and Srikumar V (eds.) *Findings of the Association for Computational Linguistics: ACL 2024*. Bangkok, Thailand: Association for Computational Linguistics, pp. 4629–4651. DOI:10.18653/v1/2024.findings-acl.275. URL https://aclanthology.org/2024.findings-acl.275/.

Schwartz R, Dodge J, Smith NA and Etzioni O (2020) Green ai. *Commun. ACM* 63(12): 54–63. DOI:10.1145/3381831. URL https://doi.org/10.1145/3381831.

Sel B, Tawaha A, Khattar V, Jia R and Jin M (2024) Algorithm of thoughts: Enhancing exploration of ideas in large language models. In: Salakhutdinov R, Kolter Z, Heller K, Weller A, Oliver N, Scarlett J and Berkenkamp F (eds.) *Proceedings of the 41st International Conference on Machine Learning*, *Proceedings of Machine Learning Research*, volume 235. PMLR, pp. 44136–44189. URL https://proceedings.mlr.press/v235/sel24a.html.

Sivarajkumar S, Kelley M, Samolyk-Mazzanti A, Visweswaran S and Wang Y (2024) An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: Algorithm development and validation study. *JMIR Med Inform* 12: e55318. DOI:10.2196/55318. URL https://doi.org/10.2196/55318.

Srivastava A, Rastogi A, Rao A, Shoeb AAM, Abid A, Fisch A, Brown AR, Santoro A, Gupta A, Garriga-Alonso A, Kluska A, Lewkowycz A, Agarwal A, Power A, Ray A, Warstadt A, Kocurek AW, Safaya A, Tazarv A, Xiang A, Parrish A, Nie A, Hussain A, Askell A, Dsouza A, Slone A, Rahane A, Iyer AS, Andreassen AJ, Madotto A, Santilli A, Stuhlmüller A, Dai AM, La A, Lampinen AK, Zou A, Jiang A, Chen A, Vuong A, Gupta A, Gottardi A, Norelli A, Venkatesh A, Gholamidavoodi A, Tabassum A, Menezes A, Kirubarajan A, Mullokandov A, Sabharwal A, Herrick A, Efrat A, Erdem A, Karakaş A, Roberts BR, Loe BS, Zoph B, Bojanowski B, Özyurt B, Hedayatnia B, Neyshabur B, Inden B, Stein B, Ekmekci B, Lin BY, Howald B, Orinion B, Diao C, Dour C, Stinson C, Argueta C, Ferri C, Singh C, Rathkopf C, Meng C, Baral C, Wu C, Callison-Burch C, Waites C, Voigt C, Manning CD, Potts C, Ramirez C, Rivera CE, Siro C, Raffel C, Ashcraft C, Garbacea C, Sileo D, Garrette D, Hendrycks D, Kilman D, Roth D, Freeman CD, Khashabi D, Levy D, González DM, Perszyk D, Hernandez D, Chen D, Ippolito D, Gilboa D, Dohan D, Drakard D, Jurgens D, Datta D, Ganguli D, Emelin D, Kleyko D, Yuret D, Chen D, Tam D, Hupkes D, Misra D, Buzan D, Mollo DC, Yang D, Lee DH, Schrader D, Shutova E, Cubuk ED, Segal E, Hagerman E, Barnes E, Donoway E, Pavlick E, Rodolà E, Lam E, Chu E, Tang E, Erdem E, Chang E, Chi EA, Dyer E, Jerzak E, Kim E, Manyasi EE, Zheltonozhskii E, Xia F, Siar F, Martínez-Plumed F, Happé F, Chollet F,

Rong F, Mishra G, Winata GI, de Melo G, Kruszewski G, Parascandolo G, Mariani G, Wang GX, Jaimovitch-Lopez G, Betz G, Gur-Ari G, Galijasevic H, Kim H, Rashkin H, Hajishirzi H, Mehta H, Bogar H, Shevlin HFA, Schuetze H, Yakura H, Zhang H, Wong HM, Ng I, Noble I, Jumelet J, Geissinger J, Kernion J, Hilton J, Lee J, Fisac JF, Simon JB, Koppel J, Zheng J, Zou J, Kocon J, Thompson J, Wingfield J, Kaplan J, Radom J, Sohl-Dickstein J, Phang J, Wei J, Yosinski J, Novikova J, Bosscher J, Marsh J, Kim J, Taal J, Engel J, Alabi J, Xu J, Song J, Tang J, Waweru J, Burden J, Miller J, Balis JU, Batchelder J, Berant J, Frohberg J, Rozen J, Hernandez-Orallo J, Boudeman J, Guerr J, Jones J, Tenenbaum JB, Rule JS, Chua J, Kanclerz K, Livescu K, Krauth K, Gopalakrishnan K, Ignatyeva K, Markert K, Dhole K, Gimpel K, Omondi K, Mathewson KW, Chiafullo K, Shkaruta K, Shridhar K, McDonell K, Richardson K, Reynolds L, Gao L, Zhang L, Dugan L, Qin L, Contreras-Ochando L, Morency LP, Moschella L, Lam L, Noble L, Schmidt L, He L, Oliveros-Colón L, Metz L, Senel LK, Bosma M, Sap M, Hoeve MT, Farooqi M, Faruqui M, Mazeika M, Baturan M, Marelli M, Maru M, Ramirez-Quintana MJ, Tolkiehn M, Giulianelli M, Lewis M, Potthast M, Leavitt ML, Hagen M, Schubert M, Baitemirova MO, Arnaud M, McElrath M, Yee MA, Cohen M, Gu M, Ivanitskiy M, Starritt M, Strube M, Swędrowski M, Bevilacqua M, Yasunaga M, Kale M, Cain M, Xu M, Suzgun M, Walker M, Tiwari M, Bansal M, Aminnaseri M, Geva M, Gheini M, T MV, Peng N, Chi NA, Lee N, Krakover NGA, Cameron N, Roberts N, Doiron N, Martinez N, Nangia N, Deckers N, Muennighoff N, Keskar NS, Iyer NS, Constant N, Fiedel N, Wen N, Zhang O, Agha O, Elbaghdadi O, Levy O, Evans O, Casares PAM, Doshi P, Fung P, Liang PP, Vicol P, Alipoormolabashi P, Liao P, Liang P, Chang PW, Eckersley P, Htut PM, Hwang P, Miłkowski P, Patil P, Pezeshkpour P, Oli P, Mei Q, Lyu Q, Chen Q, Banjade R, Rudolph RE, Gabriel R, Habacker R, Risco R, Millière R, Garg R, Barnes R, Saurous RA, Arakawa R, Raymaekers R, Frank R, Sikand R, Novak R, Sitelew R, Bras RL, Liu R, Jacobs R, Zhang R, Salakhutdinov R, Chi RA, Lee SR, Stovall R, Teehan R, Yang R, Singh S, Mohammad SM, Anand S, Dillavou S, Shleifer S, Wiseman S, Gruetter S, Bowman SR, Schoenholz SS, Han S, Kwatra S, Rous SA, Ghazarian S, Ghosh S, Casey S, Bischoff S, Gehrmann S, Schuster S, Sadeghi S, Hamdan S, Zhou S, Srivastava S, Shi S, Singh S, Asaadi S, Gu SS, Pachchigar S, Toshniwal S, Upadhyay S, Debnath SS, Shakeri S, Thormeyer S, Melzi S, Reddy S, Makini SP, Lee SH, Torene S, Hatwar S, Dehaene S, Divic S, Ermon S, Biderman S, Lin S, Prasad S, Piantadosi S, Shieber S, Misherghi S, Kiritchenko S, Mishra S, Linzen T, Schuster T, Li T, Yu T, Ali T, Hashimoto T, Wu TL, Desbordes T, Rothschild T, Phan T, Wang T, Nkinyili T, Schick T, Kornev T, Tunduny T, Gerstenberg T, Chang T, Neeraj T, Khot T, Shultz T, Shaham U, Misra V, Demberg V, Nyamai V, Raunak V, Ramasesh VV, vinay uday prabhu, Padmakumar V, Srikumar V, Fedus W, Saunders W, Zhang W, Vossen W, Ren X, Tong X, Zhao X, Wu X, Shen X, Yaghoobzadeh Y, Lakretz Y, Song Y, Bahri Y, Choi Y, Yang Y, Hao S, Chen Y, Belinkov Y, Hou Y, Hou Y, Bai Y, Seid Z, Zhao Z, Wang Z, Wang ZJ, Wang Z and Wu Z (2023) Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research* URL `https://openreview.net/forum?id=uyTL5Bvosj`. Featured Certification.

Suzgun M, Scales N, Schärli N, Gehrmann S, Tay Y, Chung HW, Chowdhery A, Le Q, Chi E, Zhou D and Wei J (2023) Challenging BIG-bench tasks and whether chain-of-thought can solve them. In: Rogers A, Boyd-Graber J and Okazaki N (eds.) *Findings of the Association*

for Computational Linguistics: ACL 2023. Toronto, Canada: Association for Computational Linguistics, pp. 13003–13051. DOI:10.18653/v1/2023.findings-acl.824. URL https://aclanthology.org/2023.findings-acl.824/.

Valmeekam K, Palod V, Stechly K, Gundawar A and Kambhampati S (2025) Beyond semantics: The unreasonable effectiveness of reasonless intermediate tokens. *arXiv preprint arXiv:2505.13775* .

Vatsal S and Dubey H (2024) A survey of prompt engineering methods in large language models for different nlp tasks. URL https://arxiv.org/abs/2407.12994.

Wan G, Wu Y, Chen J and Li S (2025) Reasoning aware self-consistency: Leveraging reasoning paths for efficient LLM sampling. In: Chiruzzo L, Ritter A and Wang L (eds.) *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Albuquerque, New Mexico: Association for Computational Linguistics. ISBN 979-8-89176-189-6, pp. 3613–3635. DOI:10.18653/v1/2025.naacl-long.184. URL https://aclanthology.org/2025.naacl-long.184/.

Wan Z, Wang X, Liu C, Alam S, Zheng Y, Liu J, Qu Z, Yan S, Zhu Y, Zhang Q, Chowdhury M and Zhang M (2024) Efficient large language models: A survey. *Transactions on Machine Learning Research* URL https://openreview.net/forum?id=bsCCJHbO8A. Survey Certification.

Wang J, Jain S, Zhang D, Ray B, Kumar V and Athiwaratkun B (2024) Reasoning in token economies: Budget-aware evaluation of LLM reasoning strategies. In: Al-Onaizan Y, Bansal M and Chen YN (eds.) *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Miami, Florida, USA: Association for Computational Linguistics, pp. 19916–19939. DOI:10.18653/v1/2024.emnlp-main.1112. URL https://aclanthology.org/2024.emnlp-main.1112/.

Wang L, Xu W, Lan Y, Hu Z, Lan Y, Lee RKW and Lim EP (2023a) Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In: Rogers A, Boyd-Graber J and Okazaki N (eds.) *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, pp. 2609–2634. DOI:10.18653/v1/2023.acl-long.147. URL https://aclanthology.org/2023.acl-long.147.

Wang X, Wei J, Schuurmans D, Le QV, Chi EH, Narang S, Chowdhery A and Zhou D (2023b) Self-consistency improves chain of thought reasoning in language models. In: *The Eleventh International Conference on Learning Representations*. URL https://openreview.net/forum?id=1PL1NIMMrw.

Wei J, Wang X, Schuurmans D, Bosma M, brian ichter, Xia F, Chi EH, Le QV and Zhou D (2022) Chain of thought prompting elicits reasoning in large language models. In: Oh AH, Agarwal A, Belgrave D and Cho K (eds.) *Advances in Neural Information Processing Systems*. URL https://openreview.net/forum?id=_VjQlMeSB_J.

White J, Fu Q, Hays S, Sandborn M, Olea C, Gilbert H, Elnashar A, Spencer-Smith J and Schmidt DC (2023) A prompt pattern catalog to enhance prompt engineering with chatgpt. In: *Proceedings of the 30th Conference on Pattern Languages of Programs*, PLoP '23. USA: The Hillside Group. ISBN 9781941652190.

Wilhelm P, Wittkopp T and Kao O (2025) Beyond test-time compute strategies: Advocating energy-per-token in llm inference. In: *Proceedings of the 5th Workshop on Machine Learning and Systems*. pp. 208–215.

Wu CJ, Raghavendra R, Gupta U, Acun B, Ardalani N, Maeng K, Chang G, Aga F, Huang J, Bai C et al. (2022) Sustainable ai: Environmental implications, challenges and opportunities. *Proceedings of Machine Learning and Systems* 4: 795–813.

Wu J, Zhang Z, Xia Y, Li X, Xia Z, Chang A, Yu T, Kim S, Rossi RA, Zhang R, Mitra S, Metaxas DN, Yao L, Shang J and McAuley J (2024) Visual prompting in multimodal large language models: A survey. URL https://arxiv.org/abs/2409.15310.

Yao S, Yu D, Zhao J, Shafran I, Griffiths TL, Cao Y and Narasimhan KR (2023) Tree of thoughts: Deliberate problem solving with large language models. In: *Thirty-seventh Conference on Neural Information Processing Systems*. URL https://openreview.net/forum?id=5Xc1ecxO1h.

Yin S, Fu C, Zhao S, Li K, Sun X, Xu T and Chen E (2024) A survey on multimodal large language models. *National Science Review* 11(12): nwae403. DOI:10.1093/nsr/nwae403. URL https://doi.org/10.1093/nsr/nwae403.

Yu Q, Zheng Z, Song S, Li Z, Xiong F, Tang B and Chen D (2025) xfinder: Large language models as automated evaluators for reliable evaluation. In: *The Thirteenth International Conference on Learning Representations*. URL https://openreview.net/forum?id=7UqQJUKaLM.

Zhou D, Schärli N, Hou L, Wei J, Scales N, Wang X, Schuurmans D, Cui C, Bousquet O, Le QV and Chi EH (2023) Least-to-most prompting enables complex reasoning in large language models. In: *The Eleventh International Conference on Learning Representations*. URL https://openreview.net/forum?id=WZH7099tgfM.