LESS IS MORE: HIGHER INFORMATION DENSITY INCREASES PERFORMANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

In order for LLMs to target larger and more complex tasks, they are given increasing amounts of information in context meant to help them solve those tasks. Prior work has shown that giving LLMs irrelevant information, such as random facts or whitespace, degrades LLM performance on tasks because attention is diluted by attending to irrelevant information. However, complex tasks often involve the LLM ingesting information that could be used to solve the task (i.e., relevant information) of varying degrees of relevance, rather than purely irrelevant information. In this paper, we investigate the role of relevant information density, or the amount of relevant information in context given to an LLM. We find that denser representations of relevant information perform better, even when controlling for context length by appending irrelevant information. We show this is consistent with LLMs attending to more than the minimum possible information they need, causing distraction, which we show with higher reasoning token lengths. We construct benchmark tasks which allow us to precisely control the amount and density of relevant information in context and find that, across 16 models, performance degrades on average 30.1 percentage points on all tasks due to the difference in information density, despite the model seeing the same relevant information. This degradation is accompanied by a $4.98\times$ increase in reasoning length. Our results show that managing context to maximize information density can improve model performance.

1 INTRODUCTION

Users of LLMs provide a task they want completed and add information into context that they believe will help the LLM complete the task. Tasks such as web search (Nakano et al., 2021), deep research (Shao et al., 2024), retrieval-augmented generation (Lewis et al., 2020), agentic coding (Yang et al., 2024), and tasks that involve chain-of-thought (Wei et al., 2022) are becoming increasingly complex and put more and more information in context. However, simply passing LLMs more information has risks: due to context rot (Hong et al., 2025) and attention dispersion (Veličković et al., 2025), filling context with information that is not useful to the task (i.e., irrelevant information, such as white space) harms LLM performance (Shi et al., 2023; Liu et al., 2024; Du et al., 2025; Hong et al., 2025).

But information in context is not always completely irrelevant. We define relevant information as a set of tokens in context that, under some reasoning chain, are useful for completing a task. In LLMs, the goal of the attention mechanism is to figure out how relevant different information in context is. Therefore, LLMs do not classify information as relevant or irrelevant but instead weigh the relevance of information (Figure 1). In classical machine learning models, such as Bayes classifiers (Kohavi & John, 1997) and Bayes Networks (Tsamardinos & Aliferis, 2003), different levels of relevant information impact the performance of the model. Although prior work has shown that attention to irrelevant information harms performance (Jiang et al., 2025), they do not explore how LLM performance is affected by different levels of relevant information.

In this work, we study how different levels of relevant information impact LLM performance. Specifically, we introduce a suite of benchmarks called *SmolBench* that allows us to control the density and relevance of information in context through changing the representation of information (Section 3). By construction, all information that differs between conditions in a SmolBench bench-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

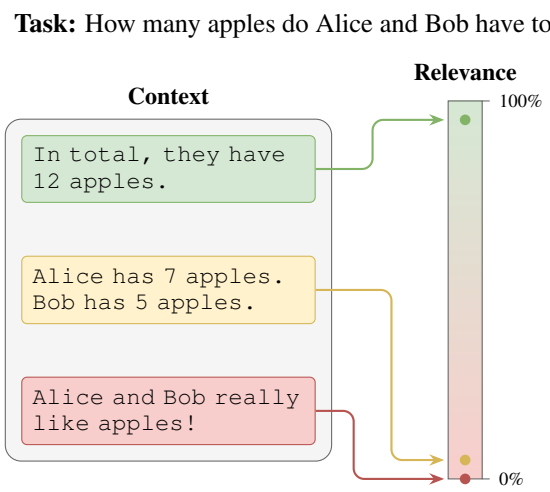


Figure 1: Attention weighs how relevant certain tokens are to generating a response relative to other tokens in context. Information is not simply relevant or irrelevant, but instead has levels of relevance depending on how useful they are to reasoning towards the solution.

mark is relevant. We use SmolBench to show that less dense representations of relevant information and representations that include information with lower relevance can harm LLM performance for both deductive [1] and inductive [2] reasoning.

Across 16 models, lower information density causes LLM performance to decrease because the LLM struggles to distinguish between relevant information, leading to long and error-prone reasoning chains. When the information is represented in a manner where each token is not highly relevant (i.e., low information density), the LLM attends to each token less. Following the resulting reasoning chains leads to lower performance on tasks. On the other hand, when information is highly relevant, the LLM attends to each relevant token more, allowing them to follow shorter reasoning chains. This attention behavior occurs even if those tokens are buried within irrelevant information, showing that the effect is by the relevance of the information, rather than the context length. Based on this property, we show that low density relevant information can cause LLMs to pursue worse reasoning chains.

In summary, we make the following contributions:

1. We show greater information density causes LLM performance to increase, as lower information density causes LLMs to go on unoptimal reasoning chains.
2. We introduce SmolBench, a suite of benchmarks that tests LLM performance on different information relevance and density for various reasoning tasks by changing the representation of information.
3. We demonstrate that the effect of information density is distinct from the effect of context length, as longer context lengths alone have a smaller effect than low density.

2 Background

Modern LLMs decide how relevant the information in context is to the task through the attention mechanism (Vaswani et al., 2017). The attention mechanism processes information in context in discrete units called tokens, which are parts of context that are encoded by vectors of length d_{model}. After a tokenizer converts the context into a matrix of tokens (T) of dimensions $d_{\text{length}} \times d_{\text{model}}$, those vectors are then multiplied by the query (W^Q), key (W^K), and value (W^V) weight matrices, which have the dimensions $d_{\text{model}} \times d_k$, to obtain the query ($Q$), key ($K$), and value ($V$) matrices of

[1] Where different representations do not change what is true of the underlying information (Shanahan, 2023).
[2] Where different representations change what is true of the underlying information (Henderson, 2024).

dimensions $d_{\text{length}} \times d_k$,:

$$Q = TW^Q \tag{1}$$

$$K = TW^K \tag{2}$$

$$V = TW^V \tag{3}$$

The relevance of each token to each other token is scored by taking the dot product of the query and key vectors of every token against every other token. This is encoded by the score matrix S, of dimensions $d_{\text{length}} \times d_{\text{length}}$:

$$S = QK^\top \tag{4}$$

To prevent extreme values in matrix S from the length of the dimension projected out in Equation 4, the matrix is then divided by $\sqrt{d_k}$ to normalize the variance of the values in S to 1. Then each row is normalized to a probability distribution via softmax, which results in the attention weight matrix α , which encodes how much a given token is relevant to every other token:

$$\alpha = \text{softmax}\left(\frac{S}{\sqrt{d_k}}\right) \tag{5}$$

We then use the attention weights to scale the information encoded by each token's vector by multiplying α by V to create an attention matrix A:

$$A = \alpha V \tag{6}$$

A encodes the information necessary to generate future tokens as augmented by attention weights, meaning the information encoded in A is based on how relevant each token is to generating a response to the task. As attention weights in α are between 0 and 1, it demonstrates that tokens, and therefore the information they encode, have varying levels of relevance to the task.

2.1 Context Compression

In context compression, tokens can be removed from context without significant loss in performance (Jiang et al., 2023; Mu et al., 2023; Eyuboglu et al., 2026). In fact, we can explicitly compress the bottom 90% of tokens with the lowest attention weights without significantly losing performance (Zweiger et al., 2026). This demonstrates empirically that not all tokens are equally relevant to the task, and that attention weights correspond to how relevant tokens are to the task.

Compression also gives us reason to believe that lower density relevant information decreases LLM performance. We define information density as the number of tokens it takes to represent the same relevant information. Compression aims to preserve the same relevant information but represent it using fewer tokens, thereby increasing information density. Certain compression methods have empirically shown that they increase LLM performance (Chevalier et al., 2023; Jiang et al., 2024; Cai et al., 2025; Yu & Chai, 2025; Kang et al., 2026), which implies that higher information density results in better LLM performance. However, these works fail to decouple the relationship between information density and reduced context length.

2.2 Context Management Benchmarks

SmolBench differs from LongBench (Bai et al., 2024) and other prior context management benchmarks (Hsieh et al., 2024; Kuratov et al., 2024; Trivedi et al., 2024; Wang et al., 2024; Adams et al., 2025; Zhou et al., 2026), as we guarantee that all information in context that differs between conditions is relevant. This allows us to vary context length and information density rather than relying solely on introducing irrelevant information to increase context length. Specifically, prior benchmarks were formulated as needle-in-a-haystack (NIAH) tasks, where relevant information was buried within irrelevant information. While reasoning may also be tested, the importance of those benchmarks was to measure the ability of the LLM to attend to the right information, and therefore identify information relevant to the ask, which is different from what we aim to measure: if LLMs have already identified information relevant to the task, does a more compact representation of relevant information (i.e., higher information density) improve LLM performance?

Task: Prove phi(x).

High Information Density

```
Facts
alpha(x)
beta(x)
Lemmas
alpha(x) ^ beta(x) ->
phi(x)
```

Low Information Density

```
Facts
alpha(x)
beta(x)
Lemmas
alpha(x) ^ beta(x) ->
phi(x)
alpha(x) ^ beta(x) ->
gamma(x)
alpha(x) ^ gamma(x) ->
delta(x)
beta(x) ^ epsilon(x) ->
zeta(x)
alpha(x) ^ zeta(x) ->
eta(x)
alpha(x) ^ eta(x) ->
phi(x)
```

(a) We use Horn clauses as our deduction task, where the lemma Horn clause that proves the target statement and the proof of the lemma are both entirely relevant information, but the former has higher information density than the latter.

Task: How many times does "Fizz" appear in one period?

High Information Density

```
"Fizz" if i % 3 == 0
"Buzz" if i % 5 == 0
```

Low Information Density

```
One Period:
i = 1: ""
i = 2: ""
i = 3: "Fizz"
i = 4: ""
i = 5: "Buzz"
i = 6: "Fizz"
i = 7: ""
i = 8: ""
i = 9: "Fizz"
i = 10: "Buzz"
i = 11: ""
i = 12: "Fizz"
i = 13: ""
i = 14: ""
i = 15: "FizzBuzz"
```

(b) We use FizzBuzz as our induction task, where the rules of FizzBuzz and the enumeration of FizzBuzz contain identical, relevant information, but the rules have higher information density than the enumeration.

Figure 2: Illustrations of our deduction and induction tasks.

3 SmolBench: Creating Distracting Relevant Information

Relevant information is any information that, for some task, context, and reasoning, allows the given task to be completed. For information to be relevant, there must exist a valid reasoning chain using that information which influences the task output. However, there may exist *more dense* representations and *less dense* representations of relevant information. For example, while theorems in mathematics can be proven directly from base axioms and thus all base axioms for a given proof are relevant information, typical proofs are made using intermediate lemmas. These intermediate lemmas are often shorter than the full proof from the axioms and thus a *more dense* representation of relevant information. In order to vary the density of relevant information that we put into context for an LLM without adding irrelevant information, we propose a suite of synthetic reasoning benchmarks of scalable difficulty which we call SmolBench.

To ensure that the context in SmolBench only includes relevant information, we formulate tasks for which the set of relevant information is known *a priori*: for every statement, we construct a reasoning chain that includes the given information to come to the desired result. To vary information density, we deterministically generate various representations of the same information that result in different lengths. For example, in one suite we constructed a set of axioms and a set of target lemmas, and then prove each lemma from the axioms. Then, we generate different representations of the information necessary to prove the target lemma by providing either previous intermediary lemmas or proofs from the axioms of previous intermediary lemmas. All information remains relevant, as any proof of an intermediary lemma is itself reasoning necessary to complete the task.

We measure whether higher information density affects LLM performance in SmolBench with two benchmarks: SmolBench-Deduction and SmolBench-Induction, capturing the impact of relevant information density for both forms of reasoning. Our split along deductive and inductive reasoning tasks mirrors that in BABILong (Kuratov et al., 2024), but we differ in that we use the split in deduction and induction to control for how we vary information density instead of determining what distracting but ultimately irrelevant information we can add to context. We leverage the fact that in deduction, changing the representation of a proven statement does not change what is true about unproven statements (Section 3.1), while in induction, changing the representation of observed data can change what is true about unseen data (Section 3.2).

3.1 SMOLBENCH-DEDUCTION

Similar to SimpleLogic (Zhang et al., 2023), we construct a benchmark using random propositional Horn rules with variable depth depending on the minimum number of proof steps. Each task is a randomly generated theory of unary Horn rules tasked with showing a proof that satisfies a goal from the list of rules. As the dense representation of relevant information, we provide a lemma library within which contains the lemmas required for the shortest proof to the goal. To add additional relevant information, we extend the lemma library with logical relationships from axioms to each lemma (Figure 2a). To account for models differing in deductive ability, SmolBench-Deduction is parameterized by the minimum number of proof steps.

3.2 SMOLBENCH-INDUCTION

We construct SmolBench-Induction using a setup similar to the FizzBuzz coding task (Ghory, 2007), using a pattern of repeating values. Each task is to count the number of occurrences of a different symbol (e.g., "Fizz" or "Buzz") before the pattern repeats (i.e., one *period*). As the dense representation of relevant information, we provide the rules generating each FizzBuzz sequence which together can be used to calculate the period of the sequence and the number of times a specific symbol occurs in a period. As the less dense version of relevant information, we provide the full enumeration of the FizzBuzz period, which can be used to count the symbols specifically, or back solve for the rules and reason from the rules (Figure 2b). To account for models differing in inductive ability, SmolBench-Induction is parameterizable by the number of FizzBuzz rules.

4 RESULTS

To ensure that our hypothesis is not an artifact of any specific LLM, we evaluate against a wide variety of parameter weights, model families, and training pipelines. Specifically, we evaluate against the Gemma4 family (Team, 2026), Nemotron family (NVIDIA, 2025a;b; 2026), Qwen 3.5 family (Qwen Team, 2026), DeepSeek V3.1 and V4 family (DeepSeek-AI, 2024; 2026), GLM 4.X family (Team et al., 2025), and Ministral 3 Reasoning-2512 family (Liu et al., 2026). We vary parameter weights to validate the effects of information density hold across model size. We evaluate across model families to ensure the effects of information density hold across model architectures. Finally, we chose the most popular families of models from various frontier labs to account for the various training pipelines available. To control for the fact that the different LLMs have different context lengths which affects depth of reasoning that is possible, we chose to limit the context length for all LLMs to the minimum natively supported context length across all LLMs: 131,072 tokens. Outputs which exceed this context limit are treated as failures.

We demonstrate that in both cases of induction and deduction, higher information density outperforms lower information density. We also demonstrate that the LLM accuracy drop caused by low information density is distinct from the effect of the LLM accuracy drop caused by context length because when we control for context length by adding irrelevant information, denser representations still outperform less dense representations of information.

4.1 SETUP: SMOLBENCH-DEDUCTION

Each task is a randomly generated theory of unary Horn rules over a single constant. A theory has a chain of m lemmas $\ell_1, \ldots, \ell_m$ whose composition derives the goal from the given facts, so the shortest proof has exactly m steps, one per lemma in the chain. To make the lemma library a search problem rather than a lookup, we add $5m$ open alternatives: rules that share a head with a chain lemma but whose premises are never satisfied. For every chain lemma we also generate a depth-two binary derivation tree from axioms; the axioms are additional Horn rules that together prove the lemma in three steps. The lemma library is shuffled with a fixed per-theory order so that position carries no information.

From one theory we produce three representations. *High density* contains the facts, the goal, and the lemma library (about 80 tokens per chain lemma). *Low density* adds the derivation tree of every chain lemma, which is entirely relevant: each tree is a valid proof of its lemma and the trees alone are enough to prove the goal. *High density + irrelevant* appends lorem ipsum text to the high-density

Model	m	Model	m	Model	m
Gemma4 E2B-it	6	Nemotron3 Nano-4B	2	Qwen3.5 27B	64
Gemma4 12B-it	32	Nemotron3 Nano-30B	6	Qwen3.5 122B	48
Gemma4 31B-it	64	Nemotron3 Super-120B	20	Qwen3.5 397B	64
Deepseek V4-Flash	64	GLM-4.7-Flash	10	Ministral3-2512 3B	1
Deepseek V3.1	64	GLM-4.7	48	Ministral3-2512 8B	1
				Ministral3-2512 14B	2

Table 1: Chain length m per model.

context until it matches the low-density token count, so the two differ only in whether the added tokens are relevant. The model is asked for a proof as a list of `derive` steps; a proof passes if every step applies a rule present in its context to premises already established and the goal is reached through any valid proof sequence.

Models differ in their deductive reasoning ability, so we calibrate m per model on the high-density context alone, on theories disjoint from the evaluation set: we search a ladder $m \in \{1, 2, 3, 4, 6, 8, 10, 12, 16, 20, 24, 32, 48, 64, 96, 192\}$ so that the high density representation passes more than 50% of the time. Ministral 3B and 8B fall below 50% at $m = 1$ and are evaluated at $m = 1$. We then evaluate all 3 contexts on 100 fresh theories with three samples per theory and context. Sampling occurred with the model's thinking mode enabled and a 131,072-token context; a response that reaches the budget without a proof counts as a failure.

4.2 Setup: SmolBench-Induction

Each task is a periodic sequence built from k rules, each pairing a period with a symbol. Every position p in the period carries the symbols of every rule whose period divides p, and the sequence repeats after the least common multiple of the periods of each rule, so the sequence period is equal to the least common multiple of the rule periods. We evaluate with $k = 9$ rules whose periods have a least common multiple of 2520, which is the largest period that keeps every LLM's context within the 131,072-token window. The LLM is asked once for each of the 9 symbols how many times that symbol occurs in one period of the sequence, and an answer is correct if the count is exact. *High density* states the 9 rules which allows the LLMs to directly compute the period and the counts. *Low density* lists all 2520 positions with their symbols, which contains the same information: the counts can be read off directly, or the rules recovered and the counts computed. *High density + irrelevant* pads the rules with white space to the token count of the listing. Symbols are drawn at random so that no memorized sequence helps we generate 30 symbol sets and sample each with a corresponding LLM seed, reporting accuracy per LLM over the 270 queries.

4.3 Analysis: LLM Accuracy

Across both the deduction results (Table 2) and induction results (Table 3), we can see that irrelevant information has a smaller effect on LLM accuracy than the effect of low information density. For deduction and induction, high information density appended with irrelevant information generally performs similarly to contexts with only high information density, with a 10% average degradation for deduction and a 7.9% average degradation for induction.

We also have cases where irrelevant information increases LLM accuracy for both inductive and deductive reasoning. For deduction, Deepseek V4-Flash and Ministral3-2512 14B unexpectedly perform better with irrelevant information in context. For induction, Gemma4 E2B, Nemotron3 Nano-4B, and Deepseek V4-Flash perform better with irrelevant information in context. The likely explanation is inter-run variance and the models being unusually robust to irrelevant information, as for all cases besides deduction's Deepseek V4-Flash results, the high information density results and high information density with irrelevant information results have overlapping 95% error bars.

Additionally, we can see that low information density results in significantly worse performance than high information density. We can see that for deduction, we always perform better with high information density than low density information across all models, with an average accuracy degradation of 40.2%. For induction, we generally perform better with high information density than low infor-

LLM Deduction Accuracy

Model	high density	high density + irrelevant	low density	Δ (high − low)	Δ (irrelevant − low)
Gemma4 E2B-it	$\mathbf{0.913 \pm 0.035}$	0.783 ± 0.053	0.457 ± 0.067	45.7	32.7
Gemma4 12B-it	$\mathbf{0.760 \pm 0.058}$	0.533 ± 0.060	0.237 ± 0.050	52.3	29.7
Gemma4 31B-it	$\mathbf{0.668 \pm 0.068}$	0.575 ± 0.069	0.127 ± 0.039	54.2	44.8
Nemotron3 Nano-4B	$\mathbf{0.813 \pm 0.052}$	0.670 ± 0.065	0.550 ± 0.067	26.3	12.0
Nemotron3 Nano-30B	$\mathbf{0.657 \pm 0.055}$	0.287 ± 0.058	0.043 ± 0.025	61.3	24.3
Nemotron3 Super-120B	$\mathbf{0.643 \pm 0.065}$	0.477 ± 0.058	0.080 ± 0.032	56.3	39.7
Qwen3.5 27B	$\mathbf{0.678 \pm 0.055}$	0.650 ± 0.068	0.150 ± 0.055	52.8	50.0
Qwen3.5 122B	$\mathbf{0.693 \pm 0.073}$	0.665 ± 0.075	0.263 ± 0.078	43.0	40.2
Qwen3.5 397B	$\mathbf{0.695 \pm 0.073}$	0.635 ± 0.080	0.365 ± 0.080	33.0	27.0
Deepseek V4-Flash	0.730 ± 0.053	$\mathbf{0.850 \pm 0.047}$	0.407 ± 0.055	32.3	44.3
Deepseek V3.1	$\mathbf{0.640 \pm 0.058}$	0.417 ± 0.058	0.157 ± 0.045	48.3	26.0
GLM-4.7-Flash	$\mathbf{0.640 \pm 0.058}$	0.493 ± 0.053	0.117 ± 0.035	52.3	37.7
GLM-4.7	$\mathbf{0.740 \pm 0.052}$	0.657 ± 0.058	0.293 ± 0.048	44.7	36.3
Ministral3-2512 3B	$\mathbf{0.333 \pm 0.057}$	0.290 ± 0.058	0.253 ± 0.055	8.0	3.7
Ministral3-2512 8B	$\mathbf{0.540 \pm 0.063}$	0.507 ± 0.070	0.417 ± 0.067	12.3	9.0
Ministral3-2512 14B	0.563 ± 0.072	$\mathbf{0.617 \pm 0.068}$	0.353 ± 0.063	21.0	26.3

Table 2: For deduction, LLMs with high density information outperform low density information, even when context length is accounted for by appending irrelevant information to high density information. The $\pm$ numbers indicate 95% error bars, or up to 2 standard deviations from the mean.

LLM Induction Accuracy

Model	high density	high density + irrelevant	low density	Δ (high − low)	Δ (irrelevant − low)
Gemma4 E2B-it	0.948 ± 0.100	$\mathbf{0.981 \pm 0.042}$	0.115 ± 0.020	83.3	86.6
Gemma4 12B-it	$\mathbf{1.000 \pm 0.000}$	0.993 ± 0.028	0.548 ± 0.162	45.2	44.5
Gemma4 31B-it	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	0.985 ± 0.048	1.5	1.5
Nemotron3 Nano-4B	0.633 ± 0.202	$\mathbf{0.719 \pm 0.186}$	0.396 ± 0.162	23.7	32.3
Nemotron3 Nano-30B	$\mathbf{0.956 \pm 0.069}$	0.937 ± 0.070	0.711 ± 0.150	24.5	22.6
Nemotron3 Super-120B	$\mathbf{1.000 \pm 0.000}$	0.996 ± 0.020	0.841 ± 0.126	15.9	15.5
Qwen3.5 27B	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	0.974 ± 0.056	2.6	2.6
Qwen3.5 122B	$\mathbf{0.996 \pm 0.020}$	0.993 ± 0.028	0.974 ± 0.056	2.2	1.9
Qwen3.5 397B	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	0.974 ± 0.048	2.6	2.6
Deepseek V4-Flash	0.959 ± 0.068	$\mathbf{0.993 \pm 0.028}$	0.963 ± 0.067	-0.4	3.0
Deepseek V3.1	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	0.993 ± 0.028	0.7	0.7
GLM-4.7-Flash	0.289 ± 0.150	0.122 ± 0.144	$\mathbf{0.393 \pm 0.133}$	-10.4	-27.1
GLM-4.7	$\mathbf{0.959 \pm 0.062}$	0.789 ± 0.236	0.889 ± 0.065	7.0	-10.0
Ministral3-2512 3B	$\mathbf{0.663 \pm 0.197}$	0.637 ± 0.214	0.407 ± 0.205	25.6	23.0
Ministral3-2512 8B	$\mathbf{0.789 \pm 0.236}$	0.215 ± 0.198	0.222 ± 0.120	56.7	-0.7
Ministral3-2512 14B	$\mathbf{0.470 \pm 0.181}$	0.026 ± 0.048	0.081 ± 0.092	38.9	-5.5

Table 3: For induction, LLMs with high density information generally outperform low density information, even when context length is accounted for by appending irrelevant information to high density information. The $\pm$ numbers indicate 95% error bars, or up to 2 standard deviations from the mean.

mation density, with an average degradation of 20.0%. Induction on GLM-4.7-Flash and Deepseek V4-Flash are the only cases where low information density performs better. This is likely because the problem was too difficult for GLM-4.7-Flash resulting in large variance in the results, and the problem was too easy for Deepseek V4-Flash, as Deepseek V4-Flash performed near ceiling for the benchmark and the difference is entirely accounted for by variance in the results.

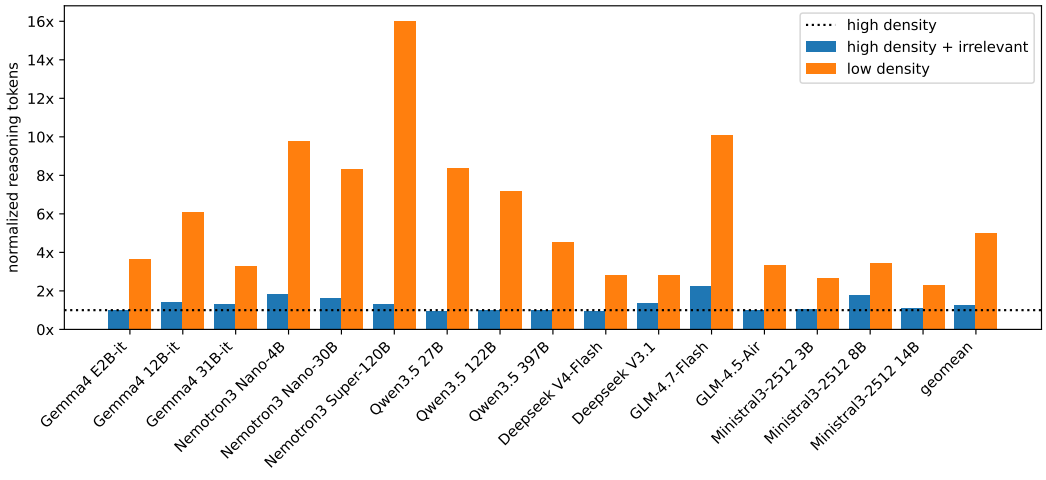


Figure 3: Reasoning length using different representations for each model, normalized to the high density representation. Responses which hit the token budget are included. The low density representation leads to consistently higher reasoning tokens generated for the same task.

Finally, across both deduction and induction, all models perform better with irrelevant information than low density relevant information, with the exception of GLM-4.7-Flash, GLM-4.7, and Ministral-2512 8B and 14B on induction. On average, the LLMs degrade on average 30.2% on deduction from irrelevant information to low density relevant information, while LLMs degrade on average 12.1% on induction from irrelevant information to low density relevant information. The exceptions are a result of the problem being too difficult for the models, resulting in a large variance in the irrelevant information and low density information results. This result shows that across LLMs, low information density has a greater effect on LLM performance than irrelevant information.

4.4 Analysis: Increased Reasoning Length

To demonstrate that the effect is caused by difficulty attending to the right information causing LLMs to go down longer, suboptimal reasoning chains, we collected the length of reasoning LLMs generated before they output a response (Figure 3). We can see that across the board, no matter how accurate the LLM was, the reasoning for irrelevant information is of similar length to that of high information density, but the reasoning length for low density information was always significantly larger. The geomean of reasoning length on irrelevant information is only $1.26\times$ worse, while the geomean for low density relevant information is $4.98\times$ worse.

4.5 Analysis: Different Reasoning Chains

For the deduction benchmark, we can trace the proofs generated by each model to determine whether they followed the provided lemmas or follow a proof of the provided lemmas. We are leveraging the fact that deduction has no implications for unknown information from known information, so we know that the emitted reasoning chains must be traceable to known information. We can see that low density information causes more erroneous proofs, either because the LLM outputted an invalid proof or ran out of context before a proof could be emitted (grey). The new relevant information additionally does get used in proofs (orange), but it has low success rates as the model pursues a worse reasoning chain with the lower density relevant information (Figure 4).

Furthermore, low density information generally causes LLMs to not respond at all, and when they use low density information, they are more likely to respond with the incorrect proof compared to using the high density information. Failed proofs only occur when either an LLM fails to respond or produces a malformed output. If they spend their entire context on reasoning, the LLM could not fully reason towards a solution to the task because they went down longer or invalid reasoning chains. Inaccurate responses indicate that the LLM was unable to correctly reason at all, so a higher rate of incorrect proofs means LLMs are going down invalid reasoning chains when they do decide

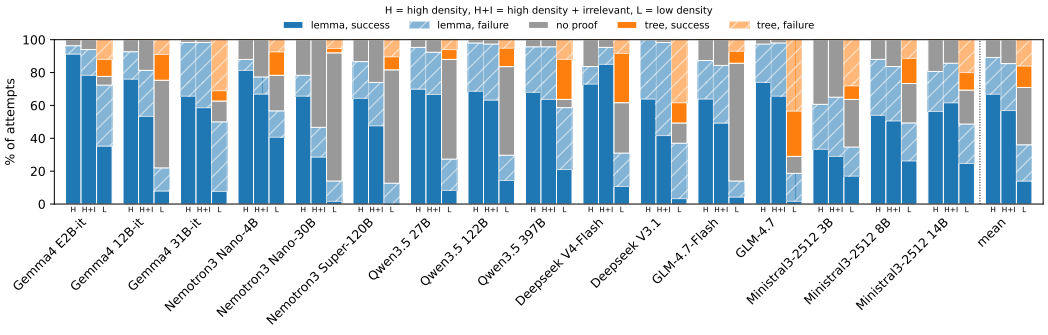


Figure 4: Distribution of results on SmolBench-Deduction, showing whether the model attempted a proof solely from lemmas (blue), used additional tree statements not in the high density representation (orange), or did not produce a proof (grey). For each model, results are shown from left to right as the high density representation, the high density + irrelevant information representation, and the low density representation. Hatched segments indicate failed proofs.

to use the low density information. In aggregate, this demonstrates that low density information causes LLM accuracy to decrease because LLMs are not able to correctly attend to the right relevant information, causing them to go down longer and invalid reasoning chains.

5 Limitations

Although we demonstrate that information density is an important property of context in maximizing LLM accuracy, we do not provide a method of increasing information density in context for applications which add information into context without human input like web search, deep research, RAG, reasoning, and agentic coding.

We also do not investigate the effects of increasing information density versus losing some relevant information. As compression methods can lose information in compression while increasing information density, it could be the case that one could increase information density aggressively at the cost of losing some relevant information, but still see a performance increase due to better attention on the relevant information that survived the compression.

Finally, we do not investigate whether models could be trained to handle lower information densities and which architecture features are more resilient to low information density. Although we demonstrate that low information density is a different problem from irrelevant information, we do not know if the performance degradation is inherent to specific LLM architectures or if it is fixable with a revised training pipeline.

6 Conclusion

Prior work explored how information irrelevant to the task affected LLM performance, which ignores the fact that relevant information has levels of relevance. We show that when relevant information is not represented in a dense manner, LLM performance decreases is consistent with LLMs attending less to each token, causing them to go down longer, suboptimal reasoning chains. This result holds across both inductive and deductive reasoning tasks, showing that for either mode of reasoning lower density information reduces LLM accuracy. It therefore suggests that LLM harnesses and methods should aim to introduce as high density information into context as possible in order to have better and more efficient reasoning.

7 Reproducibility Statement

We empirically validate that information has varying levels of relevance that leads to LLM accuracy decrease via distraction through the benchmarks described in Section 3. We provide the anonymized

code for the benchmark for reviewers to run on AWS Elastic Compute (Services, 2022) at the following link: https://anonymous.4open.science/r/SmolBench-45B7/. We also provide the LLM prompts and responses from the benchmark for reviewers to analyze without AWS Elastic Compute using the analysis code included with the benchmark at the following link: s3://smolbench-public-release. Note that depending on NVIDIA floating point arithmetic differences between hardware (NVIDIA, 2019), LLM response generation may vary.

8 AI USE STATEMENT

In this work, we used generative AI tools to assist in the implementation of methods. Our benchmark methods were manually ideated and starter code created by the authors, with infrastructure to cloud hosting providers and later extensions and test cases to the code being created by generative AI. We manually reviewed the contents and outputs of the test cases and code to ensure alignment with our claimed methods.

We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assist in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, supporting qualitative and thematic data analysis, nor interpreting results. Our synthetic dataset generation algorithms were implemented with generative AI, but no synthetic data was itself generated with generative AI. Assisting with translation and cleaning and reformatting datasets are not applicable to this work.

Finally, we used generative AI in a multi-turn discussion format, starting with a human draft, to improve readability. The final output was a revision using manual review of the result of the multi-turn conversation. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Lisa Adams, Felix Busch, Tianyu Han, Jean-Baptiste Excoffier, Matthieu Ortala, Alexander Löser, Hugo J. W. L. Aerts, Jakob Nikolas Kather, Daniel Truhn, and Keno Bressem. Longhealth: A question answering benchmark with long clinical documents. *Journal of Healthcare Informatics Research*, 9(3):280–296, June 2025. doi: 10.1007/s41666-025-00204-w. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC12290132/.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3119–3137, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.172. URL https://aclanthology.org/2024.acl-long.172/.

Zefan Cai, Wen Xiao, Hanshi Sun, Cheng Luo, Yikai Zhang, Ke Wan, Yucheng Li, Yeyang Zhou, Li-Wen Chang, Jiuxiang Gu, Zhen Dong, Anima Anandkumar, Abedelkadir Asi, and Junjie Hu. R-KV: Redundancy-aware KV cache compression for reasoning models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=2jwAjomEDB.

Alexis Chevalier, Alexander Wettig, Anirudh Ajith, and Danqi Chen. Adapting language models to compress contexts. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=kp1U6wBPXq.

DeepSeek-AI. Deepseek-v3 technical report, 2024. URL https://arxiv.org/abs/2412.19437.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026. URL https://arxiv.org/abs/2606.19348.

Yufeng Du, Minyang Tian, Srikanth Ronanki, Subendhu Rongali, Sravan Babu Bodapati, Aram Galstyan, Azton Wells, Roy Schwartz, Eliu A Huerta, and Hao Peng. Context length alone hurts LLM performance despite perfect retrieval. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 23281–23298, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1264. URL https://aclanthology.org/2025.findings-emnlp.1264/.

Sabri Eyuboglu, Ryan Saul Ehrlich, Simran Arora, Neel Guha, Dylan Zinsley, Emily Ruoyu Liu, Atri Rudra, James Zou, Azalia Mirhoseini, and Christopher Re. Cartridges: Lightweight and general-purpose long context representations via self-study. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=0k5w8O0SNg.

Imran Ghory. Using fizzbuzz to find developers who grok coding, January 2007. URL https://imranontech.com/2007/01/24/using-fizzbuzz-to-find-developers-who-grok-coding/.

Leah Henderson. The Problem of Induction. In Edward N. Zalta and Uri Nodelman (eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2024 edition, 2024. URL https://plato.stanford.edu/entries/induction-problem/.

Kelly Hong, Anton Troynikov, and Jeff Huber. Context rot: How increasing input tokens impacts llm performance. Technical report, Chroma, July 2025. URL https://trychroma.com/research/context-rot.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=kIoBbc76Sy.

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LLMLingua: Compressing prompts for accelerated inference of large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=ADsEdyI32n.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LongLLMLingua: Accelerating and enhancing LLMs in long context scenarios via prompt compression. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1658–1677, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.91. URL https://aclanthology.org/2024.acl-long.91/.

Xinyan Jiang, Hang Ye, Yongxin Zhu, Xiaoying Zheng, Zikang Chen, and Jun Gong. HICD: Hallucination-inducing via attention dispersion for contrastive decoding to mitigate hallucinations in large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 7764–7786, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.405. URL https://aclanthology.org/2025.findings-acl.405/.

Minki Kang, Wei-Ning Chen, Dongge Han, Huseyin A Inan, Lukas Wutschitz, Yanzhi Chen, Robert Sim, and Saravan Rajmohan. ACON: Optimizing context compression for long-horizon LLM agents. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=5EmOOLtH5P.

Ron Kohavi and George H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1):273–324, 1997. ISSN 0004-3702. doi: https://doi.org/10.1016/S0004-3702(97)00043-X. URL https://www.sciencedirect.com/science/article/pii/S000437029700043X. Relevance.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. Babilong: testing the limits of llms with long context reasoning-in-a-haystack. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546. URL https://dl.acm.org/doi/abs/10.5555/3495724.3496517.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, Baptiste Rozière, Baudouin De Monicault, Clémence Lanfranchi, Corentin Barreau, Cyprien Courtot, Daniele Grattarola, Darius Dabert, Diego de las Casas, Elliot Chane-Sane, Faruk Ahmed, Gabrielle Berrada, Gaëtan Ecrepont, Gauthier Guinet, Georgii Novikov, Guillaume Kunsch, Guillaume Lample, Guillaume Martin, Gunshi Gupta, Jan Ludziejewski, Jason Rute, Joachim Studnia, Jonas Amar, Joséphine Delas, Josselin Somerville Roberts, Karmesh Yadav, Khyathi Chandu, Kush Jain, Laurence Aitchison, Laurent Fainsin, Léonard Blier, Lingxiao Zhao, Louis Martin, Lucile Saulnier, Luyu Gao, Maarten Buyl, Margaret Jennings, Marie Pellat, Mark Prins, Mathieu Poirée, Mathilde Guillaumin, Matthieu Dinot, Matthieu Futeral, Maxime Darrin, Maximilian Augustin, Mia Chiquier, Michel Schimpf, Nathan Grinsztajn, Neha Gupta, Nikhil Raghuraman, Olivier Bousquet, Olivier Duchenne, Patricia Wang, Patrick von Platen, Paul Jacob, Paul Wambergue, Paula Kurylowicz, Pavankumar Reddy Muddireddy, Philomène Chagniot, Pierre Stock, Pravesh Agrawal, Quentin Torroba, Romain Sauvestre, Roman Soletskyi, Rupert Menneer, Sagar Vaze, Samuel Barry, Sanchit Gandhi, Siddhant Waghjale, Siddharth Gandhi, Soham Ghosh, Srijan Mishra, Sumukh Aithal, Szymon Antoniak, Teven Le Scao, Théo Cachet, Theo Simon Sorg, Thibaut Lavril, Thiziri Nait

Saada, Thomas Chabal, Thomas Foubert, Thomas Robert, Thomas Wang, Tim Lawson, Tom Bewley, Tom Bewley, Tom Edwards, Umar Jamil, Umberto Tomasini, Valeriia Nemychnikova, Van Phung, Vincent Maladière, Virgile Richard, Wassim Bouaziz, Wen-Ding Li, William Marshall, Xinghui Li, Xinyu Yang, Yassine El Ouahidi, Yihan Wang, Yunhao Tang, and Zaccharie Ramzi. Ministral 3, 2026. URL https://arxiv.org/abs/2601.08584.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

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

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. Webgpt: Browser-assisted question-answering with human feedback. *CoRR*, abs/2112.09332, 2021. URL https://arxiv.org/abs/2112.09332.

NVIDIA. 5.5. floating-point computation — cuda programming guide, 2019. URL https://docs.nvidia.com/cuda/cuda-programming-guide/05-appendices/mathematical-functions.html#notes-on-host-device-computation-accuracy.

NVIDIA. Nvidia nemotron 3: Efficient and open intelligence, 2025a. URL https://arxiv.org/abs/2512.20856. White Paper.

NVIDIA. Nemotron 3 Nano: Open, efficient mixture-of-experts hybrid Mamba-Transformer model for Agentic reasoning, 2025b. URL https://arxiv.org/abs/2512.20848. Technical report.

NVIDIA. Nemotron 3 super: Open, efficient mixture-of-experts hybrid mamba-transformer model for agentic reasoning, 2026. URL https://arxiv.org/abs/2604.12374.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Amazon Web Services. Amazon elastic compute cloud documentation, 2022. URL https://docs.aws.amazon.com/ec2/.

Timothy Shanahan. Deductive and inductive arguments, 2023. URL https://iep.utm.edu/deductive-inductive-arguments/.

Yijia Shao, Yucheng Jiang, Theodore Kanell, Peter Xu, Omar Khattab, and Monica Lam. Assisting in writing Wikipedia-like articles from scratch with large language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6252–6278, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.347. URL https://aclanthology.org/2024.naacl-long.347/.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023. URL https://openreview.net/forum?id=JSZmoN03Op.

Gemma Team. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

GLM Team, Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, Kedong Wang, Lucen Zhong, Mingdao Liu, Rui Lu, Shulin Cao, Xiaohan Zhang, Xuancheng Huang, Yao Wei, Yean Cheng, Yifan An, Yilin Niu, Yuanhao Wen, Yushi Bai, Zhengxiao Du, Zihan Wang, Zilin Zhu, Bohan Zhang, Bosi Wen, Bowen Wu, Bowen Xu, Can Huang, Casey Zhao, Changpeng Cai, Chao Yu, Chen Li, Chendi Ge, Chenghua Huang, Chenhui Zhang, Chenxi Xu, Chenzheng Zhu, Chuang Li, Congfeng Yin, Daoyan Lin, Dayong Yang, Dazhi Jiang, Ding Ai, Erle Zhu, Fei Wang, Gengzheng Pan, Guo Wang, Hailong Sun, Haitao Li, Haiyang Li, Haiyi Hu, Hanyu Zhang, Hao Peng, Hao Tai, Haoke Zhang, Haoran Wang, Haoyu Yang, He Liu, He Zhao, Hongwei Liu, Hongxi Yan, Huan Liu, Huilong Chen, Ji Li, Jiajing Zhao, Jiamin Ren, Jian Jiao, Jiani Zhao, Jianyang Yan, Jiaqi Wang, Jiayi Gui, Jiayue Zhao, Jie Liu, Jijie Li, Jing Li, Jing Lu, Jingsen Wang, Jingwei Yuan, Jingxuan Li, Jingzhao Du, Jinhua Du, Jinxin Liu, Junkai Zhi, Junli Gao, Ke Wang, Lekang Yang, Liang Xu, Lin Fan, Lindong Wu, Lintao Ding, Lu Wang, Man Zhang, Minghao Li, Minghuan Xu, Mingming Zhao, Mingshu Zhai, Pengfan Du, Qian Dong, Shangde Lei, Shangqing Tu, Shangtong Yang, Shaoyou Lu, Shijie Li, Shuang Li, Shuang-Li, Shuxun Yang, Sibo Yi, Tianshu Yu, Wei Tian, Weihan Wang, Wenbo Yu, Weng Lam Tam, Wenjie Liang, Wentao Liu, Xiao Wang, Xiaohan Jia, Xiaotao Gu, Xiaoying Ling, Xin Wang, Xing Fan, Xingru Pan, Xinyuan Zhang, Xinze Zhang, Xiuqing Fu, Xunkai Zhang, Yabo Xu, Yandong Wu, Yida Lu, Yidong Wang, Yilin Zhou, Yiming Pan, Ying Zhang, Yingli Wang, Yingru Li, Yinpei Su, Yipeng Geng, Yitong Zhu, Yongkun Yang, Yuhang Li, Yuhao Wu, Yujiang Li, Yunan Liu, Yunqing Wang, Yuntao Li, Yuxuan Zhang, Zezhen Liu, Zhen Yang, Zhengda Zhou, Zhongpei Qiao, Zhuoer Feng, Zhuorui Liu, Zichen Zhang, Zihan Wang, Zijun Yao, Zikang Wang, Ziqiang Liu, Ziwei Chai, Zixuan Li, Zuodong Zhao, Wenguang Chen, Jidong Zhai, Bin Xu, Minlie Huang, Hongning Wang, Juanzi Li, Yuxiao Dong, and Jie Tang. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models, 2025. URL https://arxiv.org/abs/2508.06471.

Harsh Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. AppWorld: A controllable world of apps and people for benchmarking interactive coding agents. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16022–16076, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.850. URL https://aclanthology.org/2024.acl-long.850/.

Ioannis Tsamardinos and Constantin F. Aliferis. Towards principled feature selection: Relevancy, filters and wrappers. In Christopher M. Bishop and Brendan J. Frey (eds.), *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, volume R4 of *Proceedings of Machine Learning Research*, pp. 300–307. PMLR, 03–06 Jan 2003. URL https://proceedings.mlr.press/r4/tsamardinos03a.html. Reissued by PMLR on 01 April 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Petar Veličković, Christos Perivolaropoulos, Federico Barbero, and Razvan Pascanu. Softmax is not enough (for sharp size generalisation). In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=S4JmmpnSPy.

Zilong Wang, Yuedong Cui, Li Zhong, Zimin Zhang, Da Yin, Bill Yuchen Lin, and Jingbo Shang. Officebench: Benchmarking language agents across multiple applications for office automation, 2024. URL https://arxiv.org/abs/2407.19056.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=_VjQlMeSB_J.

John Yang, Carlos E Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik R Narasimhan, and Ofir Press. SWE-agent: Agent-computer interfaces enable automated software engineering. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=mXpq6ut8J3.

Bohan Yu and Yekun Chai. EvolKV: Evolutionary KV cache compression for LLM inference. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 1673–1689, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.88. URL https://aclanthology.org/2025.findings-emnlp.88/.

Honghua Zhang, Liunian Harold Li, Tao Meng, Kai-Wei Chang, and Guy Van Den Broeck. On the paradox of learning to reason from data. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI '23, 2023. ISBN 978-1-956792-03-4. doi: 10.24963/ijcai.2023/375. URL https://doi.org/10.24963/ijcai.2023/375.

Zijian Zhou, Ao Qu, Zhaoxuan Wu, Sunghwan Kim, Alok Prakash, Daniela Rus, Bryan Kian Hsiang Low, and Paul Pu Liang. MEM1: Learning to synergize memory and reasoning for efficient long-horizon agents. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=XY8AaxDSLb.

Adam Zweiger, Xinghong Fu, Han Guo, and Yoon Kim. Fast KV compaction via attention matching. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=t01SKTj8pP.