Not All Gap Correction Helps: A Geometric View of External Information in LLM Inference

Dingzirui Wang[1]*, Xuanliang Zhang[1], Keyan Xu[1], Qingfu Zhu[1], Wanxiang Che[1]†, Yang Deng[2]
[1]Harbin Institute of Technology [2]Singapore Management University
{dzrwang,xuanliangzhang,kyxu,qfzhu,car}@ir.hit.edu.cn
ydeng@smu.edu.sg

Abstract

Providing external information beyond the query is a common way to improve large language model (LLM) inference, as seen in in-context learning (ICL), retrieval-augmented generation (RAG), and memory-based methods (Mem). Existing work increasingly suggests that useful external information should compensate for the gap between LLMs and user queries, implicitly assuming that stronger gap-correction can lead to better performance. In this paper, we challenge this assumption and show that more gap-corrective external information can instead harm performance. We analyze this phenomenon in Transformers through the lens of reasoning error, defined as the difference between the predicted answer vector and the ground-truth answer vector. Our analysis reveals that external information can be not only under-corrective but also over-corrective, where excessively strong correction increases reasoning error. We further show that the induced correction vector is jointly determined by attention and the complementarity between the external information and the query, and derive conditions under which information is properly corrective in both direction and magnitude. Experiments on four mainstream LLMs and seven reasoning benchmarks across ICL, RAG, and Mem validate our theory and support a theory-guided information selection method.

1 Introduction

Providing query-related information in the input is an effective way to improve the inference performance of large language models (LLMs). This paradigm underlies a wide range of methods, including In-Context Learning (ICL) with demonstrations [6], Retrieval-Augmented Generation (RAG) with external knowledge [19], and Memory-based methods (Mem) with historical interaction [25]. In this paper, we refer to information provided in the input beyond the query as *external information*.

Recent studies suggest that useful external information should not merely resemble the query, but should compensate for what the model lacks in solving it. We refer to such information as *gap-corrective external information*. Existing work has recognized that external information could hurt performance when it is overly long, noisy, or redundant, and has proposed to control such issues via attention-entropy balancing, graph-based evidence organization, or sentence-level context refinement [33, 22, 39][3]. However, these works mainly address the quality and organization of external information while implicitly assuming that stronger gap correction is always preferable. They therefore overlook a crucial distinction between correction that is misaligned with the model's error and correction that is directionally useful but excessive in magnitude. In this work, we argue that not all gap correction is beneficial. The effect of external information is fundamentally geometric: it helps only when it corrects the model in the right direction and by the proper magnitude.

*Work was done during a visit at SMU.
†Corresponding author.
[3]Considering the page limit, we discuss related works in Appendix C.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

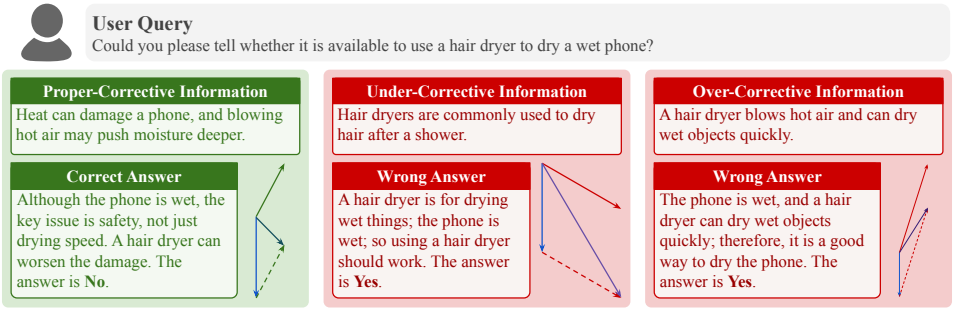


Figure 1: An illustration of different types of information, where green denotes properly corrective information, and red denotes under-corrective or over-corrective information. In the lower-right corner of each part, the blue vector represents the baseline error vector, while the green and red vectors represent the information-induced correction vectors. The cyan and purple vectors denote the corresponding composed vectors, respectively. The composed vector for properly corrective information is shorter than the baseline error vector, showing lower error. In contrast, the composed vectors for under-corrective or over-corrective information are longer than the baseline error vector.

To study this, we measure performance using the reasoning error, which is the vector difference between the predicted answer and the ground truth. We analyze how external information affects this error in Transformers. Our findings are summarized in Table 1. First, we show that external information can be not only *under-corrective*, failing to compensate for what the model lacks, but also *over-corrective*, where the model attends too strongly to the provided information, resulting in an increase in the reasoning error, as shown in Figure 1. Second, we characterize the factors that determine the external *information-induced correction vector*, showing that it is governed by both attention and the complementarity between the external information and the query. Third, based on this analysis, we derive the conditions under which external information is *properly corrective*, where the induced correction must satisfy a joint direction–magnitude requirement so that it is neither under-corrective nor over-corrective, ensuring to decrease the reasoning error.

For ICL, RAG, and Mem, we conduct experiments on four mainstream LLMs and seven reasoning benchmarks, empirically verifying the existence of over-corrective external information. Further experiments show that the effects of different factors on corrective strength are consistent with our theoretical predictions and that the conditions for information to be properly corrective also align with our derivations. In addition, based on our theoretical analysis, we propose an information selection method that maximizes the norm of the correction vector while satisfying the proper-correction condition. Compared with the baselines, it achieves an average relative error decrease of 5.9%, demonstrating that our theoretical findings can effectively guide practice.

Our contributions can be summarized as follows:

- We present a geometric view of how external information affects LLM inference, showing that stronger gap correction is not always better but could harm performance.
- We identify the key factors behind the information-induced correction vector, showing that its strength depends on both relevance, measured by attention to the external information, and complementarity, measured by its difference from the query in the value space.
- We establish the conditions for external information to be properly corrective, where the induced correction must have both the right direction and the proper magnitude to reduce reasoning errors.

2 Preliminary

In this section, we introduce some necessary mathematical notation. **Notably, our simplifications of the model input–output mapping and Transformer architecture follow common practice in prior theoretical studies [42, 24] and are further supported by the consistency between our theoretical discovery and the experimental results in §4.** We regard the external information, user query, and output as single semantic vectors after encoding them separately. In Appendix E, we discuss that our conclusions continue to hold when the inputs and outputs consist of multiple tokens. Specifically, let

Table 1: The main findings and corresponding evidence of this paper.

Finding	Evidence	Experiment
Over-corrective external information could harm performance.	Theorem 1	§4.2
Corrective strength of external information is governed by both attention and complementarity between query and information.	Corollary 2, Corollary 4	§4.3
The proper-corrective external information should satisfy a joint direction-magnitude requirement $\rho(t,x) > \frac{\|g(t,x)\|}{2\|e(x)\|}$.	Corollary 3, Corollary 5	§4.4

$d \in \mathbb{N}$ denote the embedding dimension. We consider the input matrix $E = [\, t,\, x\,] \in \mathbb{R}^{d\times 2}$, where $t \in \mathbb{R}^d$ represents the external information and $x \in \mathbb{R}^d$ represents the user query. The prediction function $f : \mathbb{R}^{L\times d} \to \mathbb{R}^{L\times d}$ maps an input of arbitrary length L to an output of the same length. Given a target $y \in \mathbb{R}^d$, we view the output at the position corresponding to x, denoted f_x, as the model readout and define the **information-conditioned error vector** $e(t,x) := f_x(t,x) - y \in \mathbb{R}^d$. As a comparison, we define the **baseline error vector** $e(x) := f_x(x) - y$, where $e(t,x) - e(x)$ reflects how external information affect the error.

In this paper, we primarily focus on the Transformer architecture. In light of prior studies showing that many components of the Transformer do not materially affect the overall properties [42, 10], we adopt a suitably simplified Transformer architecture in this paper, allowing us to focus more sharply on the fundamental properties of the model itself. We further show in Appendix I that the components omitted in our analysis do not affect the main conclusions of this work. Specifically, the model we consider consists of one or more Transformer blocks, where each block contains an attention layer and an MLP layer with a residual connection.

Attention layer Let $W_Q, W_K, W_V, W_O \in \mathbb{R}^{d\times d}$ be the query matrix, key matrix, value matrix, and output projection matrix, respectively. For any vector $z \in \mathbb{R}^d$, define $q_z := W_Q z,\ k_z := W_K z,\ v_z := W_V z$. Let the score matrix $S \in \mathbb{R}^{2\times 2}$ have entries

$$S_{uw} := \frac{\langle q_u,\, k_w\rangle}{\sqrt{d}}, \qquad u, w \in \{t, x\}.$$

Applying the $\mathrm{softmax}$ operation row-wise, we define the attention weights

$$\alpha_{u\leftarrow w} := \frac{\exp(S_{uw})}{\sum_{w'\in\{t,x\}} \exp(S_{uw'})}, \qquad u, w \in \{t, x\}.$$

Then the attention output at position $u \in \{t, x\}$ is defined as

$$\mathrm{Attn}_u(E) := W_O \left(\sum_{w\in\{t,x\}} \alpha_{u\leftarrow w}\, v_w \right) \in \mathbb{R}^d.$$

MLP layer Given weights $W_1, W_2 \in \mathbb{R}^{d\times d}$. Let the nonlinear activation function $\sigma : \mathbb{R}^d \to \mathbb{R}^d$ be 1-Lipschitz. We define the MLP layer with the residual connection by

$$\mathrm{MLP}(z) := z + W_2\, \sigma(W_1 z), \qquad z \in \mathbb{R}^d.$$

Finally, the output of a single block is defined as

$$\mathrm{Block}_u(E) := \mathrm{MLP}\big(E_{\bullet(u)} + \mathrm{Attn}_u(E)\big) \in \mathbb{R}^d.$$

3 How External Information Affects Performance

In this part, we use the reasoning error to measure performance and progressively discuss how external information affects the reasoning error. For ease of exposition, we first prove in the single-information, single-layer setting that there exists over-corrective external information that can increase the error, and we characterize the factors that affect the corrective effect of external information, as well as the conditions under which it is properly corrective. We then generalize the above results to the multi-information and multi-layer setting, making the conclusions of this paper more realistic. All proofs of this section are present in Appendix D.

3.1 Single Information, Single Layer

In this part, we analyze how a single piece of external information affects the reasoning error in a single-layer Transformer. Let

$$\delta_{\text{attn}}(t, x) = \text{Attn}_x([t, x]) - \text{Attn}_x([x])$$

be the change in the attention output after introducing external information t. Let the Jacobian matrix of the MLP be

$$J_{\text{MLP}}(z) := I + W_2\, J_\sigma(W_1 z)\, W_1,$$

where $\|J_\sigma(\cdot)\| \leq 1$ holds almost everywhere[4]. Then we can derive that:

Theorem 1 (Correction of External Information)**.**

$$e(t, x) = e(x) + \int_0^1 J_{\text{MLP}}\big(\text{Attn}_x([x]) + s\,\delta_{\text{attn}}(t, x)\big)\delta_{\text{attn}}(t, x)\,\mathrm{d}s$$

Theorem 1 gives the change in the error after introducing external information. We can see that the second term on the right-hand side measures how external information corrects the baseline error, so we define the **information-corrective vector** as:

$$g(t, x) := \int_0^1 J_{\text{MLP}}\big(\text{Attn}_x([x]) + s\,\delta_{\text{attn}}(t, x)\big)\delta_{\text{attn}}(t, x)\,\mathrm{d}s$$

As illustrated in Figure 1, there are two cases that lead to an increase in error: *(i) under-corrective*: the induced correction is not directionally aligned with the baseline error, meaning that the provided information does not correct what the model lacks; *(ii) over-corrective*: the induced correction points in the right direction, but its norm is too large, meaning that the model pays excessive attention to the provided information.

Based on Theorem 1, we can conclude that:

Corollary 2 (Factors that Affect the Corrective Effect of External Information)**.** *For a single-layer Transformer with a single external information t, let* $L_{\text{MLP}} := 1 + \|W_1\|\,\|W_2\|$*. Then*

$$\|g(t, x)\| \leq L_{\text{MLP}}\alpha_{x\leftarrow t}\,\|W_O\|\,\|v_t - v_x\|.$$

Corollary 2 gives the factors that affect the corrective effect of external information, including: *(i) relevance* $\alpha_{x\leftarrow t}$, namely how much attention the query pays to external information; *(ii) complementarity* $\|v_t - v_x\|$, namely how different the external information is from the query in the value projection space. A larger relevance or complementarity increases the strength of the induced correction, which can be helpful for decreasing error but can also push the model into the over-corrective regime when the correction becomes too large.

Based on Theorem 1 and Corollary 2, we provide the conditions under which external information is properly corrective:

Corollary 3 (When External Information is Properly Corrective)**.** *For a single-layer Transformer with a single external information t, if* $g(t, x) \neq 0$ *and* $e(x) \neq 0$*, define*

$$\rho(t, x) := \frac{\langle -e(x),\, g(t, x)\rangle}{\|e(x)\|\,\|g(t, x)\|} \in [-1, 1].$$

Then

$$\|e(t, x)\| < \|e(x)\| \quad \Longleftrightarrow \quad \frac{\|g(t, x)\|}{2\,\|e(x)\|} < \rho(t, x).$$

In Corollary 3, $\rho(t, x)$ denotes the cosine of the angle between the correction vector $g(t, x)$ and the negative baseline error $-e(x)$. We refer to it as the **correction-error angle**. Corollary 3 shows that external information can reduce the error only when the information satisfies a joint *direction-magnitude* condition. Under such condition, under-corrective and over-corrective correspond respectively to $\rho(t, x) < 0$ and $0 < \rho(t, x) < \frac{\|g(t,x)\|}{2\,\|e(x)\|}$.

[4]In this paper, $\|\cdot\|$ denotes the Euclidean norm for vectors and the operator norm induced by the Euclidean norm for matrices.

3.2 Multiple Information, Multiple Layer

We finally extend the analysis to the realistic setting where a Transformer processes multiple pieces of external information across multiple layers. Consider a Transformer with L layers. Let the input consist of the query position x and an existing external information set $T = \{t_1, \ldots, t_n\}$.

Let $h_x^{(\ell)}(T, x)$ denote the hidden state and let $\mathrm{Attn}_x^{(\ell)}(T, x)$ denote the attention output at position x in layer ℓ. Define the reasoning error with the existing external information by

$$e^{(L)}(T, x) := h_x^{(L)}(T, x) - y.$$

Now suppose a new piece of external information t is added on top of the existing set T. We suppose each residual block is $L_{\mathrm{blk}}^{(j)}$-Lipschitz. At layer ℓ, its induced attention change at position x is

$$\delta_{\mathrm{attn}}^{(\ell)}(t; T, x) := \mathrm{Attn}_x^{(\ell)}(T \cup \{t\}, x) - \mathrm{Attn}_x^{(\ell)}(T, x).$$

In analogy to the single-layer case, define the local correction vector at layer ℓ as

$$g^{(\ell)}(t; T, x) := \int_0^1 J_{\mathrm{MLP}^{(\ell)}}\big(\mathrm{Attn}_x^{(\ell)}(T, x) + s\, \delta_{\mathrm{attn}}^{(\ell)}(t; T, x)\big)\, \delta_{\mathrm{attn}}^{(\ell)}(t; T, x)\, \mathrm{d}s.$$

To propagate this local correction to the final layer, let

$$\Phi^{(\ell+1:L)}(z) := f^{(L)} \circ \cdots \circ f^{(\ell+1)}(z),$$

and define

$$\tilde{g}^{(\ell)}(t; T, x) := \int_0^1 J_{\Phi^{(\ell+1:L)}}\big(h_x^{(\ell)}(T, x) + s\, g^{(\ell)}(t; T, x)\big)\, g^{(\ell)}(t; T, x)\, \mathrm{d}s.$$

By the residual structure, the overall correction induced by the newly introduced external information is

$$g^{(L)}(t; T, x) := \sum_{\ell=1}^{L} \tilde{g}^{(\ell)}(t; T, x),$$

and therefore $e^{(L)}(t; T, x) = e^{(L)}(T, x) + g^{(L)}(t; T, x)$.

The next two corollaries show that the same logic from the previous conclusion still holds, where the effect of new external information is governed by both its magnitude and its direction relative to the current error:

Corollary 4 (Factors that Affect the Corrective Effect of External Information in Multi-Layer Transformers)**.** *Let* $L_{\mathrm{MLP}}^{(\ell)} := 1 + \|W_1^{(\ell)}\|\, \|W_2^{(\ell)}\|$. *Then the overall correction vector satisfies*

$$\|g^{(L)}(t; T, x)\| \le \sum_{\ell=1}^{L} \Big(\prod_{j=\ell+1}^{L} L_{\mathrm{blk}}^{(j)} \Big) L_{\mathrm{MLP}}^{(\ell)} \alpha_{x \leftarrow t}^{(\ell)} \|W_O^{(\ell)}\| \|\Delta v^{(\ell)}(t; T, x)\|.$$

Corollary 4 shows that, at each layer, the strength of the correction induced by the new external information is still determined by two factors: *(i) relevance* $\alpha_{x \leftarrow t}^{(\ell)}$, and *(ii) complementarity* $\|\Delta v^{(\ell)}(t; T, x)\|$. Compared with the single-layer case, complementarity is now measured against the effective state already formed by the query and the existing external information, and the upper layers further scale this correction through $\prod_{j=\ell+1}^{L} L_{\mathrm{blk}}^{(j)}$.

Corollary 5 (When External Information is Properly Corrective in Multi-Layer Transformers)**.** *If* $\|g^{(L)}(t; T, x)\| > 0$ *and* $e^{(L)}(T, x) \neq 0$*, define*

$$\rho^{(L)}(t; T, x) := \frac{\langle -e^{(L)}(T, x),\, g^{(L)}(t; T, x) \rangle}{\|e^{(L)}(T, x)\|\, \|g^{(L)}(t; T, x)\|} \in [-1, 1].$$

Then

$$\|e^{(L)}(t; T, x)\| < \|e^{(L)}(T, x)\| \quad \Longleftrightarrow \quad \rho^{(L)}(t; T, x) > \frac{\|g^{(L)}(t; T, x)\|}{2\, \|e^{(L)}(T, x)\|}.$$

Corollary 5 shows that even in the multi-information, multi-layer setting, external information reduces the reasoning error only when the induced correction is properly corrective relative to the current error. Hence, deeper architectures do not change the basic criterion. New external information helps only when it points in the right direction and has the proper scale. Otherwise, it could still lead to under-correction or over-correction.

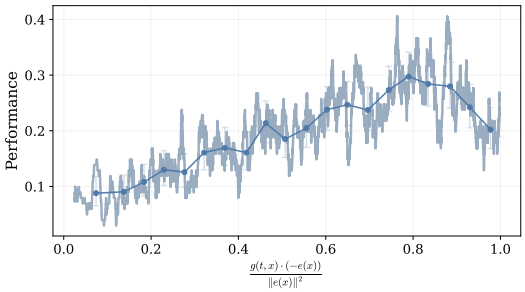


Figure 2: The performance change under different relative corrective effect when using contextual information. Since performance of each data is a boolean value, we plot it against the first variable by grouping samples into quantile-based bins, computing the empirical accuracy and Wilson confidence interval within each bin, and overlaying a rolling-average curve to visualize the trend.

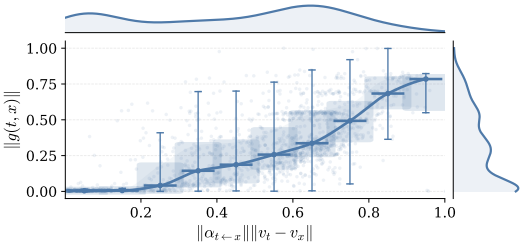


Figure 3: The variation of information-corrective vector norm $\|g(t,x)\|$ with respect to the relevance and complementarity $\alpha_{x\leftarrow t}\|v_t - v_x\|$, where each point represents one data instance. The curves at the top and to the right show the distributions of the data points along the x-axis and y-axis, respectively. The Pearson correlation coefficient for the fitted points is 0.770.

4 Experiment

4.1 Experiment Setup

Dataset We mainly conduct experiments on three types of mainstream tasks to fully validate our theoretical conclusions, which include: *(i)* Math: MATH [9], CHAMP [26]; *(ii)* Reason: TheoremQA [2], MMLU-Redux [7], GPQA [29]; *(iii)* QA: NaturalQuestions [15], FinQA [3]. We introduce these datasets in detail in Appendix F.

Model We conduct experiments on four mainstream LLMs, including Llama-3.1-8B-Instruct (Llama-3.1-8B) [8], DeepSeek-R1-Distill-Llama-8B (Llama-R1-8B) [5], Qwen3-8B [38], and Ministral-8B-Instruct-2410 (Ministral-8B) [28]. Given that these LLMs cover different families, we believe they can adequately validate our theoretical conclusions.

Setup We first filter the bad cases under the no-information setting, ensuring that the experimental data consist of queries that the models cannot solve on their own. To reflect mainstream ways of leveraging external information, we construct external information as follows: *(i)* ICL: we use demonstrations from the training set as information; *(ii)* RAG: we use the knowledge provided in the dataset as information; *(iii)* Mem: at test time, we use historical test instances as information. The detailed methodological design and experimental configurations are provided in Appendix G, and the practical computation of error and correction vectors is detailed in Appendix H. We merge the data from all datasets, models, and methods as our experimental data. To uniformly reflect the relationship, we normalize the results of all settings. All our experiments are run on two H100-80G and are based on Transformers [36] and vLLM [16], which cost about 20 minutes under each setting.

4.2 RQ1: Could More Gap-Corrective External Information Harm Performance?

To verify that more gap-corrective external information could harm performance, we analyze performance under different levels of relative corrective effect. As discussed in §3.1, we use $\frac{g(t,x)\cdot(-e(x))}{\|e(x)\|^2}$ to measure the relative corrective effect, i.e., the projection of the information-corrective vector onto the negative direction of the baseline error, normalized by the magnitude of the baseline error. We then further normalize it so that the results can be examined consistently across different datasets and models. The experimental results are shown in Figure 2. We observe that: *(i)* Overall, performance first increases and then decreases as the corrective effect grows, indicating that an excessively strong corrective effect can indeed hurt performance. *(ii)* The turning point is around 0.8. Moreover, compared with over-correction, performance degrades more noticeably under under-correction, suggesting that for the datasets and models in our experiments, under-correction remains the primary performance bottleneck. We will further investigate this issue in §5.1.

4.3 RQ2: How Different Factors Affect Correction Strength?

In this part, we adapt experiments to validate the conclusion of Corollary 2 on the factors that govern the *corrective strength* of external information. Recall that the norm of the information-corrective vector is upper-bounded by terms jointly determined by *relevance* ($\alpha_{x\leftarrow t}$) and *complementarity* ($\|v_t - v_x\|$). Accordingly, we examine how the corrective strength $\|g(t,x)\|$ varies with these two factors. The experimental results of the change curve are shown in Figure 3.

From the figure, we observe that: *(i)* The norm of the information-corrective vector is overall positively correlated with $\alpha_{x\leftarrow t}\|v_t - v_x\|$, which supports our theoretical claim that stronger relevance and complementarity generally lead to stronger correction. *(ii)* This relationship is not strictly linear. This is consistent with Corollary 2, which characterizes $\alpha_{x\leftarrow t}\|v_t - v_x\|$ as governing an *upper bound* on the corrective strength rather than its exact value. As a result, many examples exhibit substantially weaker correction than the bound would permit. *(iii)* Most samples are concentrated in the upper half of the relevance–complementarity plane, while their corrective strength remains mostly below 0.1. This suggests that even when the provided external information is relatively relevant and complementary, the induced correction is often still weak. In other words, many candidates are potentially *properly corrective*, but not sufficiently strong to fully offset the baseline error, which further explains why external information often yields only limited gains in practice.

4.4 RQ3: When is External Information Properly Corrective?

In this part, we empirically validate the conclusion of Corollary 3 regarding whether external information is properly corrective. We analyze how the corrective direction and strength jointly affect error reduction. The results are shown in Figure 4, from which we make two observations: *(i)* Almost all examples with decreased error fall into the region $\frac{\|g(t,x)\|}{2\|e(x)\|} < \rho(t,x)$. This shows that properly corrective information is sufficiently aligned with the negative baseline error, and its magnitude is not excessively large relative to that alignment. This observation provides direct empirical support for the error-reduction condition in Corollary 3. *(ii)* There are also many examples of increased error around $\rho(t,x) = 1$ and $\frac{\|g(t,x)\|}{2\|e(x)\|} = 1$. This indicates that even when the corrective direction is nearly ideal, an excessively large corrective magnitude can still harm performance. In other words, these cases are *over-corrective*, where the model overemphasizes the contextual information and correspondingly underweights the user query, which leads to a substantial increase in error.

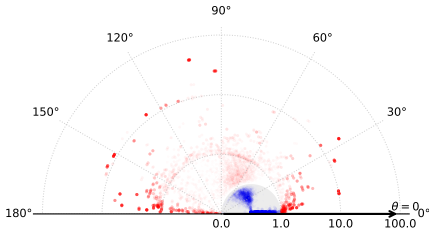


Figure 4: Error change with respect to the information-corrective vector $g(t,x)$, where each point in the figure represents one instance. The length from the axis origin to each point represents $\frac{\|g(t,x)\|}{2\|e(x)\|}$, and the angle with the positive x-axis represents $\arccos \rho(t,x)$. Blue and red indicate that the error change is smaller or larger than 0, and darker colors indicate larger absolute values. The gray region corresponds to $\frac{\|g(t,x)\|}{2\|e(x)\|} < \rho(t,x)$.

5 Theory-Guided Information Selection

In this section, we investigate how the theoretical insights developed above can be translated into practical strategies for external information selection during LLM inference. To this end, we study lightweight, theory-guided improvements on Llama-3.1-8B and Qwen3-8B, all of which introduce only minimal additional computational overhead.

5.1 Distribution of Corrective Patterns

To understand why external information sometimes hurts performance, we analyze examples with $\|e(t,x)\| > \|e(x)\|$. According to Theorem 1, such harmful cases mainly fall into two categories: *(i) over-corrective*, where $\frac{\|g(t,x)\|}{2\|e(x)\|} > \rho(t,x)$, meaning that the corrective vector is not poorly directed but is too large relative to its alignment; *(ii) under-corrective*, where $\rho(t,x) < 0$, meaning that the corrective vector points away from the negative direction of the baseline error.

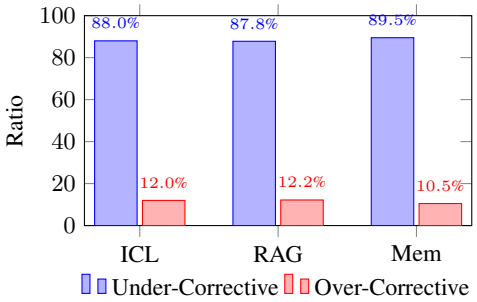


Figure 5: Distribution of harmful correction where the information-conditioned error vector is larger than the baseline error vector. Under-Corrective indicates that the corrective direction is misaligned with the negative baseline error. Over-Corrective indicates that the corrective magnitude is excessively large.

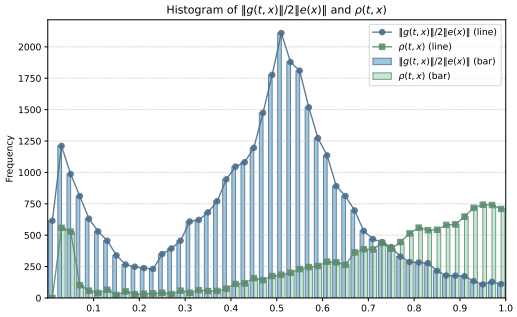


Figure 6: Distribution of the corrective strength ratio $\frac{\|g(t,x)\|}{2\|e(x)\|}$ and the alignment score $\rho(t,x)$ on examples where the error decreases after using external information cross all experimental settings. The blue part denotes the corrective strength and the green part denotes the angle.

Table 2: Performance of ICL with top-1 selection of five runs. Zero denotes inference without external information. The best result under each setting is marked in **bold**.

Model	Method	MATH	GPQA	FinQA
Llama-3.1-8B	Zero	46.0	31.3	48.6
	BM25	46.8	42.4	49.2
	DICL [12]	47.4	43.0	51.3
	GenICL [43]	47.6	43.0	51.6
	TopicK [14]	47.0	42.9	49.8
	Ours	$\mathbf{49.6 \pm 0.7}$	$\mathbf{46.5 \pm 2.8}$	$\mathbf{53.1 \pm 1.2}$
Qwen3-8B	Zero	87.0	64.6	60.1
	BM25	88.0	63.6	64.2
	DICL [12]	88.0	64.6	67.1
	GenICL [43]	88.4	65.2	67.3
	TopicK [14]	86.2	64.1	64.2
	Ours	$\mathbf{90.2 \pm 0.5}$	$\mathbf{65.7 \pm 2.2}$	$\mathbf{69.5 \pm 1.7}$

Table 3: Performance of RAG with top-1 selection of five runs. TQA denotes TheoremQA and NQ denotes NaturalQuestions. InstructRAG refers to the InstructRAG-ICL [35].

Model	Method	CHAMP	TQA	NQ
Llama-3.1-8B	Zero	14.1	26.8	38.5
	BM25	22.2	37.6	74.7
	DPR [13]	23.0	38.4	75.6
	InstructRAG [35]	24.1	39.2	76.3
	ReFeedL [41]	24.8	40.0	77.1
	Ours	$\mathbf{27.4 \pm 1.0}$	$\mathbf{42.5 \pm 0.5}$	$\mathbf{79.4 \pm 0.5}$
Qwen3-8B	Zero	39.6	54.6	44.2
	BM25	42.6	62.8	73.8
	DPR [13]	43.7	63.6	74.5
	InstructRAG [35]	44.8	64.4	75.1
	ReFeedL [41]	45.2	65.1	75.6
	Ours	$\mathbf{47.8 \pm 1.4}$	$\mathbf{67.2 \pm 0.4}$	$\mathbf{78.9 \pm 0.3}$

The results are shown in Figure 5. We observe that the dominant failure mode is *under-correction*. This indicates that the main challenge in using external information is not simply to increase the correction magnitude, but to ensure that the induced correction moves the model in the proper corrective direction. In other words, external information may fail not because it is too weak, but because it does not sufficiently align with the direction needed to reduce the baseline error.

We further examine examples where the error decreases after using external information. Figure 6 shows that $\rho(t,x)$ is strongly skewed toward 1, suggesting that helpful external information is usually well aligned with the negative baseline error. However, most values of $\frac{\|g(t,x)\|}{2\|e(x)\|}$ lie between 0.1 and 0.5. This indicates that many successful cases are only *weakly corrective*: they point in the right direction and thus reduce the error, but their corrective strength is still not large enough to fully cancel the original baseline error. Therefore, successful external information is characterized by two aspects: proper directional alignment and sufficient corrective strength.

These observations suggest that reducing reasoning error requires a two-stage use of external information. *(i)* First, one should prioritize *properly corrective* information, namely information whose corrective vector is aligned with the negative baseline error and avoids both over-correction and under-correction. *(ii)* Then, among such candidates, one should prefer those with larger corrective strength, so that the residual error can be reduced more substantially.

5.2 Theory-Guided External Information Selection

Motivated by the above analysis of external information, we design a simple two-stage strategy for external information selection: first identify candidates that are likely to be *properly corrective*, and then choose the one with the largest potential corrective strength.

Table 4: Performance of Mem with top-1 selection of five runs. Zero denotes inference without information. TQA denotes TheoremQA, MMLU denotes MMLU-Redux, and NQ denotes NaturalQuestions. The best performance of each setting is marked in **bold**.

Model	Method	MATH	CHAMP	TQA	MMLU	GPQA	NQ	FinQA
Llama-3.1-8B	Zero	46.0	14.1	26.8	63.8	31.3	38.5	48.6
	BM25	46.4	14.4	27.0	64.1	31.8	38.9	49.0
	DC [31]	47.0	14.8	27.6	64.6	32.3	39.3	49.5
	ExpRAG [34]	47.0	14.8	27.6	64.7	31.8	39.4	49.7
	ReMem [34]	47.2	15.2	27.7	64.9	32.3	39.8	49.6
	Ours	$\mathbf{49.6 \pm 0.7}$	$\mathbf{18.1 \pm 2.3}$	$\mathbf{30.4 \pm 0.5}$	$\mathbf{67.1 \pm 0.4}$	$\mathbf{34.8 \pm 3.1}$	$\mathbf{41.7 \pm 0.9}$	$\mathbf{52.1 \pm 1.5}$
Qwen3-8B	Zero	87.0	39.6	54.6	65.8	64.6	44.2	60.1
	BM25	87.2	40.0	54.8	66.0	65.2	44.6	60.4
	DC [31]	88.0	40.4	55.4	66.5	65.2	45.0	60.9
	ExpRAG [34]	88.0	40.7	55.6	66.6	65.7	45.1	61.1
	ReMem [34]	88.2	40.7	55.6	66.8	65.2	45.3	61.5
	Ours	$\mathbf{89.0 \pm 0.4}$	$\mathbf{43.0 \pm 2.4}$	$\mathbf{58.7 \pm 0.5}$	$\mathbf{69.9 \pm 0.4}$	$\mathbf{68.7 \pm 2.9}$	$\mathbf{47.5 \pm 0.9}$	$\mathbf{63.4 \pm 1.5}$

Corrective Direction Since computing the alignment correction-error angle $\rho(t, x)$ requires the ground-truth answer, it is unavailable at inference time. Following prior work [37, 23], we train a smaller model to predict this quantity. Specifically, we use ELECTRA-Large-Discriminator [4] as the predictor, where the relative error is shown in Appendix J. We merge the training splits of each dataset as supervision. The input consists of the user query, the candidate external information, and the prediction of the LLM without external information, while the output is $\rho(t, x)$.

Corrective Strength Corollary 2 shows that the norm of the corrective vector is upper-bounded by $\alpha_{x\leftarrow t}\|v_t - v_x\|$ up to model-dependent constants. This suggests that the corrective strength of external information is jointly determined by its *relevance* $\alpha_{x\leftarrow t}$ and *complementarity* $\|v_t - v_x\|$. Since $\alpha_{x\leftarrow t}$, v_t, and v_x can be directly extracted from the model, we use $\alpha_{x\leftarrow t}\|v_t - v_x\|$ as a lightweight proxy for the potential corrective strength of each candidate.

To reduce computational cost, for each query, we first use the retrieval procedure in Appendix G to coarsely collect 32 candidate pieces of external information. We then estimate $\rho(t, x)$ for each candidate and compute its corrective-strength proxy $\alpha_{x\leftarrow t}\|v_t - v_x\|$. We first filter out candidates that are unlikely to be properly corrective according to the predicted angle and then select the candidate with the largest corrective-strength proxy among the remaining ones to guide inference.

We evaluate this strategy under multiple settings in Table 2, Table 3, and Table 4, where all methods use top-1 external information during inference. We also discuss how different factors affect the performance in Appendix K. Compared with the corresponding baselines, our method achieves comparable or better performance across settings, yielding an average relative error reduction of 5.9%. These results suggest that our theory is not only descriptive but also practically useful: it can guide external information selection by distinguishing non-corrective, over-corrective, and properly corrective candidates in a principled manner. Besides, the goal of this section is not to propose a new state-of-the-art system but to verify the predictive power of our theory and show that it can support principled design choices in realistic settings.

6 Conclusion

In this paper, we study how external information affects LLM reasoning in Transformers. We show that stronger gap correction does not necessarily improve performance. Besides under-corrective information, there also exists an overlooked over-corrective regime where the model over-attends to external information, leading to an increase in reasoning errors. To explain this phenomenon, we characterize the correction vector induced by external information and show that its strength depends on both attention-based relevance and value-space complementarity with the query. Based on this analysis, we establish the conditions for information to be properly corrective, requiring both an appropriate direction and a proper magnitude. Experiments across ICL, RAG, and Mem settings on multiple LLMs and benchmarks empirically verify our theoretical findings. We further show that these insights can guide practice by improving information selection. Overall, our results suggest that effective context engineering should not pursue stronger correction blindly, but instead seek external information that is properly corrective to the missing capability of LLMs for the query.

Acknowledgment

We gratefully acknowledge the support of the National Natural Science Foundation of China (NSFC) via grant 62236004 and 62476073. This research/project is also supported by the National Research Foundation Singapore under the AI Singapore Programme (AISG Award No: AISG3-RPGV-2025-016). Yang Deng is support by the Lee Kong Chian Fellowship awarded by Singapore Management University.

References

[1] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2024.

[2] Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. TheoremQA: A theorem-driven question answering dataset. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7889–7901, Singapore, December 2023. Association for Computational Linguistics.

[3] Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. FinQA: A dataset of numerical reasoning over financial data. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[4] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*, 2020.

[5] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025.

[6] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[7] Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile Van Krieken, and Pasquale Minervini. Are we done with MMLU? In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *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)*, pages 5069–5096, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[8] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe,

Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024.

[9] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[10] Yu Huang, Yuan Cheng, and Yingbin Liang. In-context convergence of transformers. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 19660–19722. PMLR, 21–27 Jul 2024.

[11] Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore, December 2023. Association for Computational Linguistics.

[12] Janak Kapuriya, Manit Kaushik, Debasis Ganguly, and Sumit Bhatia. Exploring the role of diversity in example selection for in-context learning. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 2962–2966, New York, NY, USA, 2025. Association for Computing Machinery.

[13] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online, November 2020. Association for Computational Linguistics.

[14] Wonbin Kweon, SeongKu Kang, Runchu Tian, Pengcheng Jiang, Jiawei Han, and Hwanjo Yu. Topic coverage-based demonstration retrieval for in-context learning. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19911–19923, Suzhou, China, November 2025. Association for Computational Linguistics.

[15] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019.

[16] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

[17] Tiziano Labruna, Jon Ander Campos, and Gorka Azkune. When to retrieve: Teaching LLMs to utilize information retrieval effectively. In Galia Angelova, Maria Kunilovskaya, Marie Escribe, and Ruslan Mitkov, editors, *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 623–632, Varna, Bulgaria, September 2025. INCOMA Ltd., Shoumen, Bulgaria.

[18] 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 H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc., 2020.

[19] Yangning Li, Weizhi Zhang, Yuyao Yang, Wei-Chieh Huang, Yaozu Wu, Junyu Luo, Yuanchen Bei, Henry Peng Zou, Xiao Luo, Yusheng Zhao, Chunkit Chan, Yankai Chen, Zhongfen Deng, Yinghui Li, Hai-Tao Zheng, Dongyuan Li, Renhe Jiang, Ming Zhang, Yangqiu Song, and Philip S. Yu. A survey of RAG-reasoning systems in large language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 12120–12145, Suzhou, China, November 2025. Association for Computational Linguistics.

[20] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024.

[21] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for GPT-3? In Eneko Agirre, Marianna Apidianaki, and Ivan Vulić, editors, *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online, May 2022. Association for Computational Linguistics.

[22] Zhenghao Liu, Mingyan Wu, Xinze Li, Yukun Yan, Shuo Wang, Cheng Yang, Minghe Yu, Zheni Zeng, and Maosong Sun. Graph-anchored knowledge indexing for retrieval-augmented generation, 2026.

[23] Ziyue Liu, Ruijie Zhang, Zhengyang Wang, Mingsong Yan, Zi Yang, Paul D. Hovland, Bogdan Nicolae, Franck Cappello, Sui Tang, and Zheng Zhang. CoLA: Compute-efficient pre-training of LLMs via low-rank activation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4627–4645, Suzhou, China, November 2025. Association for Computational Linguistics.

[24] Yue M. Lu, Mary Letey, Jacob A. Zavatone-Veth, Anindita Maiti, and Cengiz Pehlevan. Asymptotic theory of in-context learning by linear attention. *Proceedings of the National Academy of Sciences*, 122(28), July 2025.

[25] Jinghao Luo, Yuchen Tian, Chuxue Cao, Ziyang Luo, Hongzhan Lin, Kaixin Li, Chuyi Kong, Ruichao Yang, and Jing Ma. From storage to experience: A survey on the evolution of llm agent memory mechanisms. *Preprints*, January 2026.

[26] Yujun Mao, Yoon Kim, and Yilun Zhou. CHAMP: A competition-level dataset for fine-grained analyses of LLMs' mathematical reasoning capabilities. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13256–13274, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[27] Lingrui Mei, Jiayu Yao, Yuyao Ge, Yiwei Wang, Baolong Bi, Yujun Cai, Jiazhi Liu, Mingyu Li, Zhong-Zhi Li, Duzhen Zhang, Chenlin Zhou, Jiayi Mao, Tianze Xia, Jiafeng Guo, and Shenghua Liu. A survey of context engineering for large language models, 2025.

[28] Mistral AI team. Un ministral, des ministraux. `https://mistral.ai/news/ministraux`, oct 2024. Accessed: 2025-11-18.

[29] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024.

[30] Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. REPLUG: Retrieval-augmented black-box language models. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8371–8384, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[31] Mirac Suzgun, Mert Yuksekgonul, Federico Bianchi, Dan Jurafsky, and James Zou. Dynamic cheatsheet: Test-time learning with adaptive memory, 2025.

[32] Yile Wang, Peng Li, Maosong Sun, and Yang Liu. Self-knowledge guided retrieval augmentation for large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10303–10315, Singapore, December 2023. Association for Computational Linguistics.

[33] Yuhao Wang, Ruiyang Ren, Yucheng Wang, Jing Liu, Wayne Xin Zhao, Hua Wu, and Haifeng Wang. Bee-rag: Balanced entropy engineering for retrieval-augmented generation, 2025.

[34] Tianxin Wei, Noveen Sachdeva, Benjamin Coleman, Zhankui He, Yuanchen Bei, Xuying Ning, Mengting Ai, Yunzhe Li, Jingrui He, Ed H. Chi, Chi Wang, Shuo Chen, Fernando Pereira, Wang-Cheng Kang, and Derek Zhiyuan Cheng. Evo-memory: Benchmarking llm agent test-time learning with self-evolving memory, 2025.

[35] Zhepei Wei, Wei-Lin Chen, and Yu Meng. InstructRAG: Instructing retrieval-augmented generation via self-synthesized rationales. In *The Thirteenth International Conference on Learning Representations*, 2025.

[36] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[37] Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V. Le, Tengyu Ma, and Adams Wei Yu. Doremi: optimizing data mixtures speeds up language model pretraining. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

[38] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025.

[39] Ruobing Yao, Yifei Zhang, Shuang Song, Yuhan Liu, Neng Gao, and Chenyang Tu. ParetoRAG: Leveraging sentence-context attention for robust and efficient retrieval-augmented generation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 4137–4151, Suzhou, China, November 2025. Association for Computational Linguistics.

[40] Zijun Yao, Weijian Qi, Liangming Pan, Shulin Cao, Linmei Hu, Liu Weichuan, Lei Hou, and Juanzi Li. SeaKR: Self-aware knowledge retrieval for adaptive retrieval augmented generation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27022–27043, Vienna, Austria, July 2025. Association for Computational Linguistics.

[41] Taewon Yun, Jihwan Oh, Hyangsuk Min, Yuho Lee, Jihwan Bang, Jason Cai, and Hwanjun Song. Refeed: Multi-dimensional summarization refinement with reflective reasoning on feedback. In *Second Conference on Language Modeling*, 2025.

[42] Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. Trained transformers learn linear models in-context. *J. Mach. Learn. Res.*, 25(1), January 2024.

[43] Zheng Zhang, Shaocheng Lan, Lei Song, Jiang Bian, Yexin Li, and Kan Ren. Learning to select in-context demonstration preferred by large language model. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11345–11360, Vienna, Austria, July 2025. Association for Computational Linguistics.

[44] Zihan Zhang, Meng Fang, and Ling Chen. RetrievalQA: Assessing adaptive retrieval-augmented generation for short-form open-domain question answering. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6963–6975, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[45] Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. Expel: Llm agents are experiential learners. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19632–19642, Mar. 2024.

A Limitation

This work focuses on understanding how external information affects LLM inference from the perspective of reasoning error. To make the analysis tractable, we adopt a simplified Transformer formulation and represent the query, external information, and output as semantic vectors. While we further extend the analysis to multi-information and multi-layer settings and empirically validate the conclusions across multiple models and benchmarks, the formulation is still an abstraction of practical LLM inference. Future work may study more fine-grained settings, such as token-level dynamics, longer contexts, and multi-modal external information.

B Social Impact

External information is widely used in LLM applications such as retrieval-augmented generation, in-context learning, and memory-based assistants. By showing that external information should be properly corrective rather than merely relevant or strong, this work may help improve the reliability of such systems and reduce cases where noisy, misleading, or over-emphasized context degrades model reasoning. This can be beneficial for applications that require careful use of retrieved or historical information.

At the same time, our findings also highlight that external information must be selected and used responsibly. If the candidate information itself contains biased, private, outdated, or misleading content, a better selection mechanism may still amplify undesirable effects. Therefore, the proposed perspective should be combined with existing safeguards for data quality, privacy protection, and bias mitigation when deployed in real-world systems.

C Related Work

External information is a standard way to extend LLM inference beyond the knowledge stored in model parameters [19, 6, 27]. Early work mainly follows a query-centric view, where useful external information is assumed to be the most similar or relevant to the input query. Representative methods therefore focus on improving query-information matching, for example by retrieving semantically similar demonstrations or passages, or by further promoting diversity and topic coverage among the selected candidates [21, 13, 18, 30, 14]. More recent work questions this assumption by noting that LLMs already encode internal knowledge, so external retrieval is not always necessary. This has led to adaptive or model-aware methods that aim to provide external information only when the model lacks the knowledge required by the current query. For example, FLARE [11] triggers retrieval when generation becomes uncertain, Self-RAG [1] learns to decide when retrieval is needed and how to critique retrieved evidence, and SKR [32] explicitly models what the LLM knows and does not know. RetrievalQA [44] further benchmarks this setting, while recent analyzes show that retrieval is especially helpful for long-tail or weakly memorized knowledge. Along the same line, ADAPT-LLM [17] and SeaKR [40] learn retrieval triggers based on retrieval necessity or self-aware uncertainty. Similarly, GenICL [43] moves beyond raw similarity by directly optimizing demonstration usefulness.

However, in this paper, we argue that when selecting external information, information that better fills the gap between the model and the query is not always better. We reveal an *over-correction* phenomenon, where the model over-attends to the provided external information, leading to an increase in reasoning errors as a result. Based on this, we analyze the factors that govern corrective strength and derive the conditions that external information must satisfy in order to be properly corrective.

D Proof

D.1 Supporting Lemmas

We first present several lemmas used in the proofs.

Lemma 6. *For a single external information t,*

$$\delta_{\mathrm{attn}}(t, x) = \alpha_{x \leftarrow t} \, W_O \, (v_t - v_x).$$

Proof of Lemma 6. By definition,

$$\mathrm{Attn}_x([t,x]) = W_O\big(\alpha_{x\leftarrow t} v_t + \alpha_{x\leftarrow x} v_x\big).$$

When the input only contains the query token x, the attention weight on x is 1, and hence

$$\mathrm{Attn}_x([x]) = W_O v_x.$$

Therefore,

$$\begin{aligned}\delta_{\mathrm{attn}}(t,x) &= \mathrm{Attn}_x([t,x]) - \mathrm{Attn}_x([x]) \\ &= W_O\big(\alpha_{x\leftarrow t} v_t + \alpha_{x\leftarrow x} v_x - v_x\big) \\ &= W_O\big(\alpha_{x\leftarrow t} v_t - (1-\alpha_{x\leftarrow x}) v_x\big).\end{aligned}$$

Since $\alpha_{x\leftarrow t} + \alpha_{x\leftarrow x} = 1$, we obtain

$$\delta_{\mathrm{attn}}(t,x) = \alpha_{x\leftarrow t}\, W_O\,(v_t - v_x).$$

□

Lemma 7. *For any $z, z' \in \mathbb{R}^d$,*

$$\|\mathrm{MLP}(z) - \mathrm{MLP}(z')\| \le L_{\mathrm{MLP}} \|z - z'\|, \qquad L_{\mathrm{MLP}} := 1 + \|W_1\|\,\|W_2\|.$$

Proof of Lemma 7. Since

$$\mathrm{MLP}(z) = z + W_2\,\sigma(W_1 z),$$

we have

$$\mathrm{MLP}(z) - \mathrm{MLP}(z') = (z - z') + W_2\big(\sigma(W_1 z) - \sigma(W_1 z')\big).$$

Using the triangle inequality and the fact that σ is 1-Lipschitz,

$$\begin{aligned}\|\mathrm{MLP}(z) - \mathrm{MLP}(z')\| &\le \|z - z'\| + \|W_2\|\,\|\sigma(W_1 z) - \sigma(W_1 z')\| \\ &\le \|z - z'\| + \|W_2\|\,\|W_1 z - W_1 z'\| \\ &\le \big(1 + \|W_1\|\,\|W_2\|\big)\|z - z'\|.\end{aligned}$$

Hence

$$\|\mathrm{MLP}(z) - \mathrm{MLP}(z')\| \le L_{\mathrm{MLP}} \|z - z'\|.$$

□

Lemma 8. *Let $T = \{t_1, \ldots, t_n\}$ be the existing external information set. Denote by $\alpha^{(T)}_{x\leftarrow u}$ the attention weights under input $[T, x]$, and define*

$$\bar{v}_{x,T} := \sum_{u \in T \cup \{x\}} \alpha^{(T)}_{x\leftarrow u}\, v_u = \sum_{i=1}^{n} \alpha^{(T)}_{x\leftarrow t_i}\, v_{t_i} + \alpha^{(T)}_{x\leftarrow x}\, v_x.$$

After adding a new external information t, denote by $\alpha^{(T\cup\{t\})}_{x\leftarrow u}$ the attention weights under input $[T \cup \{t\}, x]$. Then

$$\delta_{\mathrm{attn}}(t; T, x) := \mathrm{Attn}_x([T \cup \{t\}, x]) - \mathrm{Attn}_x([T, x]) = \alpha^{(T\cup\{t\})}_{x\leftarrow t} W_O\big(v_t - \bar{v}_{x,T}\big).$$

Proof of Lemma 8. Let s_u be the attention logit from position x to token $u \in T \cup \{x\}$, and let

$$Z := \sum_{u \in T \cup \{x\}} \exp(s_u).$$

Under input $[T, x]$, we have

$$\alpha^{(T)}_{x\leftarrow u} = \frac{\exp(s_u)}{Z}, \qquad \bar{v}_{x,T} = \sum_{u \in T \cup \{x\}} \alpha^{(T)}_{x\leftarrow u} v_u.$$

After adding t, let its logit be s_t. Then under input $[T \cup \{t\}, x]$,

$$\alpha^{(T\cup\{t\})}_{x\leftarrow t} = \frac{\exp(s_t)}{Z + \exp(s_t)}, \qquad \alpha^{(T\cup\{t\})}_{x\leftarrow u} = \frac{\exp(s_u)}{Z + \exp(s_t)} = \big(1 - \alpha^{(T\cup\{t\})}_{x\leftarrow t}\big)\alpha^{(T)}_{x\leftarrow u} \quad (u \in T \cup \{x\}).$$

Therefore,

$$\begin{aligned}\mathrm{Attn}_x([T\cup\{t\},x]) &= W_O\Big(\alpha_{x\leftarrow t}^{(T\cup\{t\})}v_t + \sum_{u\in T\cup\{x\}}\alpha_{x\leftarrow u}^{(T\cup\{t\})}v_u\Big)\\ &= W_O\Big(\alpha_{x\leftarrow t}^{(T\cup\{t\})}v_t + \big(1-\alpha_{x\leftarrow t}^{(T\cup\{t\})}\big)\bar{v}_{x,T}\Big).\end{aligned}$$

Subtracting

$$\mathrm{Attn}_x([T,x]) = W_O\bar{v}_{x,T},$$

we obtain

$$\begin{aligned}\delta_{\mathrm{attn}}(t;T,x) &= W_O\Big(\alpha_{x\leftarrow t}^{(T\cup\{t\})}v_t + \big(1-\alpha_{x\leftarrow t}^{(T\cup\{t\})}\big)\bar{v}_{x,T} - \bar{v}_{x,T}\Big)\\ &= \alpha_{x\leftarrow t}^{(T\cup\{t\})}W_O\big(v_t-\bar{v}_{x,T}\big).\end{aligned}$$

□

D.2 Proof of Theorem 1

Proof of Theorem 1. Let

$$z_0 := \mathrm{Attn}_x([x]), \qquad z_1 := \mathrm{Attn}_x([t,x]).$$

Then by definition,

$$z_1 - z_0 = \delta_{\mathrm{attn}}(t,x).$$

Since

$$e(t,x) - e(x) = \mathrm{MLP}(z_1) - \mathrm{MLP}(z_0),$$

define the line segment

$$z(s) := z_0 + s(z_1 - z_0) = \mathrm{Attn}_x([x]) + s\,\delta_{\mathrm{attn}}(t,x), \qquad s\in[0,1].$$

By the fundamental theorem of calculus,

$$e(t,x) - e(x) = \int_0^1 \frac{\mathrm{d}}{\mathrm{d}s}\mathrm{MLP}\big(z(s)\big)\,\mathrm{d}s.$$

Applying the chain rule gives

$$\frac{\mathrm{d}}{\mathrm{d}s}\mathrm{MLP}\big(z(s)\big) = J_{\mathrm{MLP}}\big(z(s)\big)\,\delta_{\mathrm{attn}}(t,x).$$

Therefore,

$$e(t,x) - e(x) = \int_0^1 J_{\mathrm{MLP}}\big(\mathrm{Attn}_x([x]) + s\,\delta_{\mathrm{attn}}(t,x)\big)\delta_{\mathrm{attn}}(t,x)\,\mathrm{d}s.$$

Hence

$$e(t,x) = e(x) + \int_0^1 J_{\mathrm{MLP}}\big(\mathrm{Attn}_x([x]) + s\,\delta_{\mathrm{attn}}(t,x)\big)\delta_{\mathrm{attn}}(t,x)\,\mathrm{d}s,$$

as claimed. □

D.3 Proof of Corollary 2

Proof of Corollary 2. Recall that

$$g(t,x) := \int_0^1 J_{\mathrm{MLP}}\big(\mathrm{Attn}_x([x]) + s\,\delta_{\mathrm{attn}}(t,x)\big)\delta_{\mathrm{attn}}(t,x)\,\mathrm{d}s.$$

By the triangle inequality and the operator norm inequality,

$$\|g(t,x)\| \le \int_0^1 \Big\|J_{\mathrm{MLP}}\big(\mathrm{Attn}_x([x]) + s\,\delta_{\mathrm{attn}}(t,x)\big)\Big\|\,\|\delta_{\mathrm{attn}}(t,x)\|\,\mathrm{d}s.$$

Since

$$J_{\mathrm{MLP}}(z) = I + W_2\,J_\sigma(W_1 z)\,W_1, \qquad \|J_\sigma(\cdot)\| \le 1 \quad \text{a.e.},$$

we have for almost every z,

$$\|J_{\mathrm{MLP}}(z)\| \le 1 + \|W_1\|\,\|W_2\| = L_{\mathrm{MLP}}.$$

Therefore,

$$\|g(t,x)\| \le L_{\mathrm{MLP}} \|\delta_{\mathrm{attn}}(t,x)\|.$$

By Lemma 6,

$$\delta_{\mathrm{attn}}(t,x) = \alpha_{x\leftarrow t} W_O (v_t - v_x),$$

and thus

$$\|\delta_{\mathrm{attn}}(t,x)\| \le \alpha_{x\leftarrow t}\,\|W_O\|\,\|v_t - v_x\|.$$

Combining the two bounds yields

$$\|g(t,x)\| \le L_{\mathrm{MLP}} \alpha_{x\leftarrow t}\,\|W_O\|\,\|v_t - v_x\|.$$

□

D.4 Proof of Corollary 3

Proof of Corollary 3. By Theorem 1,

$$e(t,x) = e(x) + g(t,x).$$

Therefore,

$$\|e(t,x)\| < \|e(x)\| \quad \Longleftrightarrow \quad \|e(x) + g(t,x)\|^2 < \|e(x)\|^2.$$

Expanding the square on the left-hand side gives

$$\|e(x)\|^2 + 2\langle e(x), g(t,x)\rangle + \|g(t,x)\|^2 < \|e(x)\|^2,$$

which is equivalent to

$$2\langle e(x), g(t,x)\rangle + \|g(t,x)\|^2 < 0.$$

Rearranging yields

$$\langle -e(x), g(t,x)\rangle > \frac{\|g(t,x)\|^2}{2}.$$

Since $g(t,x) \neq 0$ and $e(x) \neq 0$, dividing both sides by $\|e(x)\|\,\|g(t,x)\|$ gives

$$\frac{\langle -e(x), g(t,x)\rangle}{\|e(x)\|\,\|g(t,x)\|} > \frac{\|g(t,x)\|}{2\,\|e(x)\|}.$$

By the definition

$$\rho(t,x) := \frac{\langle -e(x), g(t,x)\rangle}{\|e(x)\|\,\|g(t,x)\|},$$

we conclude that

$$\|e(t,x)\| < \|e(x)\| \quad \Longleftrightarrow \quad \frac{\|g(t,x)\|}{2\,\|e(x)\|} < \rho(t,x).$$

□

D.5 Proof of Corollary 4

Proof of Corollary 4. For each layer ℓ, let

$$v_u^{(\ell)} := W_V^{(\ell)} h_u^{(\ell-1)}, \qquad \bar{v}_{x,T}^{(\ell)} := \sum_{u \in T \cup \{x\}} \alpha_{x\leftarrow u}^{(\ell,T)}\, v_u^{(\ell)},$$

where $\alpha_{x\leftarrow u}^{(\ell,T)}$ denotes the attention weight at layer ℓ under input (T,x). Define

$$\Delta v^{(\ell)}(t;T,x) := v_t^{(\ell)} - \bar{v}_{x,T}^{(\ell)}.$$

Then, by the same argument as Lemma 8, the attention change at layer ℓ can be written as

$$\delta_{\mathrm{attn}}^{(\ell)}(t;T,x) = \alpha_{x\leftarrow t}^{(\ell)} W_O^{(\ell)} \Delta v^{(\ell)}(t;T,x),$$

where $\alpha_{x\leftarrow t}^{(\ell)}$ is the attention weight assigned to the newly added information t at layer ℓ.
By definition,

$$g^{(\ell)}(t;T,x) = \int_0^1 J_{\mathrm{MLP}^{(\ell)}}\big(\mathrm{Attn}_x^{(\ell)}(T,x) + s\,\delta_{\mathrm{attn}}^{(\ell)}(t;T,x)\big)\delta_{\mathrm{attn}}^{(\ell)}(t;T,x)\,\mathrm{d}s.$$

Since

$$J_{\mathrm{MLP}^{(\ell)}}(z) = I + W_2^{(\ell)} J_\sigma\big(W_1^{(\ell)} z\big) W_1^{(\ell)},$$

we have, for almost every z,

$$\big\|J_{\mathrm{MLP}^{(\ell)}}(z)\big\| \le 1 + \|W_1^{(\ell)}\|\,\|W_2^{(\ell)}\| = L_{\mathrm{MLP}}^{(\ell)}.$$

Hence

$$\|g^{(\ell)}(t;T,x)\| \le L_{\mathrm{MLP}}^{(\ell)} \|\delta_{\mathrm{attn}}^{(\ell)}(t;T,x)\|.$$

Combining this with the above expression for $\delta_{\mathrm{attn}}^{(\ell)}(t;T,x)$ gives

$$\|g^{(\ell)}(t;T,x)\| \le L_{\mathrm{MLP}}^{(\ell)} \alpha_{x\leftarrow t}^{(\ell)} \|W_O^{(\ell)}\| \|\Delta v^{(\ell)}(t;T,x)\|.$$

Next, by definition,

$$\tilde{g}^{(\ell)}(t;T,x) := \int_0^1 J_{\Phi^{(\ell+1:L)}}\big(h_x^{(\ell)}(T,x) + s\,g^{(\ell)}(t;T,x)\big) g^{(\ell)}(t;T,x)\,\mathrm{d}s.$$

Since each residual block is $L_{\mathrm{blk}}^{(j)}$-Lipschitz, the composition

$$\Phi^{(\ell+1:L)} = f^{(L)} \circ \cdots \circ f^{(\ell+1)}$$

is

$$\Big(\prod_{j=\ell+1}^{L} L_{\mathrm{blk}}^{(j)}\Big)\text{-Lipschitz.}$$

Therefore, its Jacobian norm is bounded almost everywhere by the same constant:

$$\big\|J_{\Phi^{(\ell+1:L)}}(z)\big\| \le \prod_{j=\ell+1}^{L} L_{\mathrm{blk}}^{(j)} \qquad \text{for a.e. } z.$$

It follows that

$$\begin{aligned}
\|\tilde{g}^{(\ell)}(t;T,x)\| &\le \int_0^1 \Big\|J_{\Phi^{(\ell+1:L)}}\big(h_x^{(\ell)}(T,x) + s\,g^{(\ell)}(t;T,x)\big)\Big\| \|g^{(\ell)}(t;T,x)\|\,\mathrm{d}s \\
&\le \Big(\prod_{j=\ell+1}^{L} L_{\mathrm{blk}}^{(j)}\Big)\|g^{(\ell)}(t;T,x)\|.
\end{aligned}$$

Finally, since

$$g^{(L)}(t;T,x) = \sum_{\ell=1}^{L} \tilde{g}^{(\ell)}(t;T,x),$$

we obtain

$$\begin{aligned}
\|g^{(L)}(t;T,x)\| &\le \sum_{\ell=1}^{L} \|\tilde{g}^{(\ell)}(t;T,x)\| \\
&\le \sum_{\ell=1}^{L}\Big(\prod_{j=\ell+1}^{L} L_{\mathrm{blk}}^{(j)}\Big)\|g^{(\ell)}(t;T,x)\| \\
&\le \sum_{\ell=1}^{L}\Big(\prod_{j=\ell+1}^{L} L_{\mathrm{blk}}^{(j)}\Big) L_{\mathrm{MLP}}^{(\ell)} \alpha_{x\leftarrow t}^{(\ell)} \|W_O^{(\ell)}\| \|\Delta v^{(\ell)}(t;T,x)\|.
\end{aligned}$$

This is exactly the desired bound. □

D.6 Proof of Corollary 5

Proof of Corollary 5. By construction,

$$e^{(L)}(t;T,x) = e^{(L)}(T,x) + g^{(L)}(t;T,x).$$

Therefore,

$$\|e^{(L)}(t;T,x)\| < \|e^{(L)}(T,x)\| \quad \Longleftrightarrow \quad \|e^{(L)}(T,x) + g^{(L)}(t;T,x)\|^2 < \|e^{(L)}(T,x)\|^2.$$

Expanding the square gives

$$\|e^{(L)}(T,x)\|^2 + 2\langle e^{(L)}(T,x), g^{(L)}(t;T,x)\rangle + \|g^{(L)}(t;T,x)\|^2 < \|e^{(L)}(T,x)\|^2,$$

which is equivalent to

$$2\langle e^{(L)}(T,x), g^{(L)}(t;T,x)\rangle + \|g^{(L)}(t;T,x)\|^2 < 0.$$

Rearranging yields

$$\langle -e^{(L)}(T,x), g^{(L)}(t;T,x)\rangle > \frac{\|g^{(L)}(t;T,x)\|^2}{2}.$$

Since $\|g^{(L)}(t;T,x)\| > 0$ and $e^{(L)}(T,x) \neq 0$, dividing both sides by $\|e^{(L)}(T,x)\|\,\|g^{(L)}(t;T,x)\|$ gives

$$\frac{\langle -e^{(L)}(T,x), g^{(L)}(t;T,x)\rangle}{\|e^{(L)}(T,x)\|\,\|g^{(L)}(t;T,x)\|} > \frac{\|g^{(L)}(t;T,x)\|}{2\,\|e^{(L)}(T,x)\|}.$$

By the definition

$$\rho^{(L)}(t;T,x) := \frac{\langle -e^{(L)}(T,x),\, g^{(L)}(t;T,x)\rangle}{\|e^{(L)}(T,x)\|\,\|g^{(L)}(t;T,x)\|},$$

we conclude that

$$\|e^{(L)}(t;T,x)\| < \|e^{(L)}(T,x)\| \quad \Longleftrightarrow \quad \rho^{(L)}(t;T,x) > \frac{\|g^{(L)}(t;T,x)\|}{2\,\|e^{(L)}(T,x)\|}.$$

□

E Multi-Token Query and Multi-Token External Information

In the main text, we analyze the corrective effect of external information by treating both the user query and each piece of external information as a single token. In practice, however, a user query is typically a sequence of tokens, and each piece of external information may also contain multiple tokens. In this appendix, we show that the conclusions in the main text remain qualitatively unchanged in this more realistic setting. The main difference is that a multi-token query or a multi-token external information span should be viewed as an *attention-weighted virtual token* at the target position.

E.1 Single Layer

Let the user query prefix before the target position be

$$Q = \{q_1, \ldots, q_m\},$$

and let the already provided external information tokens be collected into a set

$$U = \{u_1, \ldots, u_n\}.$$

We still focus on the target position x, which denotes the position whose hidden state is used to produce the final answer. For a single-layer Transformer, the attention output at position x on the token set $U \cup Q \cup \{x\}$ is

$$\mathrm{Attn}_x(U,Q,x) = \sum_{z \in U \cup Q \cup \{x\}} \alpha_{x \leftarrow z}(U,Q,x)\, W_O v_z.$$

Now suppose we additionally provide a new piece of external information

$$\tau = \{t_1, \ldots, t_k\},$$

which itself is a multi-token span. Denote by

$$\alpha_{x\leftarrow\tau} := \sum_{j=1}^{k} \alpha_{x\leftarrow t_j}(U, Q, \tau, x)$$

the total attention mass assigned by position x to this information span after it is introduced. When $\alpha_{x\leftarrow\tau} > 0$, define its attention-weighted virtual value by

$$\bar{v}_\tau := \frac{1}{\alpha_{x\leftarrow\tau}} \sum_{j=1}^{k} \alpha_{x\leftarrow t_j}(U, Q, \tau, x)\, v_{t_j}.$$

Similarly, define the effective value state formed by the original query and previously provided information as

$$\bar{v}_{x;U,Q} := \sum_{z\in U\cup Q\cup\{x\}} \alpha_{x\leftarrow z}(U, Q, x)\, v_z.$$

The next proposition shows that, from the viewpoint of the target position x, a multi-token information span behaves exactly like a single virtual token.

Proposition 9 (Virtual-Token Reduction for a Multi-Token Information Span)**.** *Assume $\alpha_{x\leftarrow\tau} > 0$. Then the change in the attention output caused by introducing the multi-token external information span τ is*

$$\delta_{\mathrm{attn}}(\tau; U, Q, x) := \mathrm{Attn}_x(U, Q, \tau, x) - \mathrm{Attn}_x(U, Q, x) = \alpha_{x\leftarrow\tau}\, W_O \left(\bar{v}_\tau - \bar{v}_{x;U,Q}\right).$$

If $\alpha_{x\leftarrow\tau} = 0$, then $\delta_{\mathrm{attn}}(\tau; U, Q, x) = 0$.

Proof. Let

$$Z_{\mathrm{old}} = \sum_{z\in U\cup Q\cup\{x\}} e^{s_z}, \qquad Z_{\mathrm{new}} = Z_{\mathrm{old}} + \sum_{j=1}^{k} e^{s_{t_j}},$$

where s_z is the attention score between position x and token z. For each old token $z \in U \cup Q \cup \{x\}$, its score does not change after adding τ, hence

$$\alpha_{x\leftarrow z}(U, Q, \tau, x) = \frac{e^{s_z}}{Z_{\mathrm{new}}} = \frac{Z_{\mathrm{old}}}{Z_{\mathrm{new}}} \alpha_{x\leftarrow z}(U, Q, x) = (1 - \alpha_{x\leftarrow\tau})\alpha_{x\leftarrow z}(U, Q, x).$$

Therefore,

$$\begin{aligned}
\delta_{\mathrm{attn}}(\tau; U, Q, x) &= \sum_{j=1}^{k} \alpha_{x\leftarrow t_j}(U, Q, \tau, x)\, W_O v_{t_j} \\
&\quad + \sum_{z\in U\cup Q\cup\{x\}} \Big(\alpha_{x\leftarrow z}(U, Q, \tau, x) - \alpha_{x\leftarrow z}(U, Q, x)\Big)\, W_O v_z \\
&= \sum_{j=1}^{k} \alpha_{x\leftarrow t_j}(U, Q, \tau, x)\, W_O v_{t_j} - \alpha_{x\leftarrow\tau} \sum_{z\in U\cup Q\cup\{x\}} \alpha_{x\leftarrow z}(U, Q, x)\, W_O v_z \\
&= \alpha_{x\leftarrow\tau}\, W_O \left(\bar{v}_\tau - \bar{v}_{x;U,Q}\right).
\end{aligned}$$

□

Proposition 9 shows that the only quantities that matter at the span level are: *(i)* the total relevance $\alpha_{x\leftarrow\tau}$, namely the total attention mass paid to this information span; and *(ii)* the span-level complementarity $\|\bar{v}_\tau - \bar{v}_{x;U,Q}\|$, namely how different the span's virtual value is from the current effective state.

We can now directly extend the error decomposition in the main text.

Corollary 10 (Correction of a Multi-Token Information Span)**.** *Let*

$$e(U,Q,x) := f_x(U,Q,x) - y, \qquad e(\tau;U,Q,x) := f_x(U,Q,\tau,x) - y.$$

Then

$$e(\tau;U,Q,x) = e(U,Q,x) + g(\tau;U,Q,x),$$

where

$$g(\tau;U,Q,x) := \int_0^1 J_{\mathrm{MLP}}\Big(\mathrm{Attn}_x(U,Q,x) + s\,\delta_{\mathrm{attn}}(\tau;U,Q,x)\Big)\delta_{\mathrm{attn}}(\tau;U,Q,x)\,\mathrm{d}s.$$

Hence, all conclusions in the single-token case remain valid after replacing the single-token value v_t with the virtual value $\bar{v}_\tau$, and replacing the original query-side state v_x with the effective state $\bar{v}_{x;U,Q}$.

Corollary 11 (Factors that Affect the Corrective Effect of a Multi-Token Information Span)**.** *Let* $L_{\mathrm{MLP}} := 1 + \|W_1\|\,\|W_2\|$*. Then*

$$\|g(\tau;U,Q,x)\| \le L_{\mathrm{MLP}}\,\alpha_{x\leftarrow\tau}\,\|W_O\|\,\|\bar{v}_\tau - \bar{v}_{x;U,Q}\|.$$

Corollary 11 shows that increasing the number of tokens in the information span does not by itself guaranty a stronger corrective effect. Its effect is governed only through: *(i)* how much total attention the span attracts, and *(ii)* how far its attention-weighted virtual value moves away from the current effective state. Therefore, a long but redundant span may have little effect, whereas a short but highly attended and complementary span may induce a large correction and even lead to over-correction.

Corollary 12 (When a Multi-Token Information Span is Properly Corrective)**.** *If* $g(\tau;U,Q,x) \neq 0$ *and* $e(U,Q,x) \neq 0$*, define*

$$\rho(\tau;U,Q,x) := \frac{\langle -e(U,Q,x),\, g(\tau;U,Q,x)\rangle}{\|e(U,Q,x)\|\,\|g(\tau;U,Q,x)\|} \in [-1,1].$$

Then

$$\|e(\tau;U,Q,x)\| < \|e(U,Q,x)\| \quad\Longleftrightarrow\quad \frac{\|g(\tau;U,Q,x)\|}{2\,\|e(U,Q,x)\|} < \rho(\tau;U,Q,x).$$

Thus, even when the newly provided external information contains multiple tokens, the same direction-magnitude criterion still characterizes whether it is properly corrective. Under-correction and over-correction still correspond, respectively, to poor directional alignment and excessively large correction magnitude.

E.2 Effect of a Multi-Token Query

We now specialize the above formulation to the case where there is no previously provided external information, namely $U = \varnothing$, and the user query itself contains multiple tokens. In this case, the effective query state becomes

$$\bar{v}_{x;Q} = \sum_{z\in Q\cup\{x\}} \alpha_{x\leftarrow z}(Q,x)\,v_z.$$

Equivalently, if

$$\alpha_{x\leftarrow Q} := \sum_{q\in Q}\alpha_{x\leftarrow q}(Q,x), \qquad \bar{v}_Q := \frac{1}{\alpha_{x\leftarrow Q}}\sum_{q\in Q}\alpha_{x\leftarrow q}(Q,x)\,v_q,$$

then

$$\bar{v}_{x;Q} = \alpha_{x\leftarrow x}v_x + \alpha_{x\leftarrow Q}\bar{v}_Q.$$

Therefore, relative to the single-token analysis in the main text, a multi-token user query only changes the reference state from v_x to the effective query state $\bar{v}_{x;Q}$. The qualitative conclusion remains unchanged: external information is helpful only when its induced correction points toward the missing direction of the current query state and has an appropriate scale. Hence, a longer query mainly affects the corrective effect by changing the current effective state against which external information is compared.

E.3 Multiple Layers

The same reduction also applies layer by layer in a multi-layer Transformer. Let the query prefix be Q, let the already provided external information tokens be U, and let $\tau = \{t_1, \dots, t_k\}$ be a newly added multi-token information span. At layer ℓ, define

$$\alpha^{(\ell)}_{x\leftarrow\tau} := \sum_{j=1}^{k} \alpha^{(\ell)}_{x\leftarrow t_j}(U, Q, \tau, x),$$

$$\bar{v}^{(\ell)}_{\tau} := \frac{1}{\alpha^{(\ell)}_{x\leftarrow\tau}} \sum_{j=1}^{k} \alpha^{(\ell)}_{x\leftarrow t_j}(U, Q, \tau, x)\, v^{(\ell)}_{t_j},$$

and

$$\bar{v}^{(\ell)}_{x;U,Q} := \sum_{z\in U\cup Q\cup\{x\}} \alpha^{(\ell)}_{x\leftarrow z}(U, Q, x)\, v^{(\ell)}_{z}.$$

Then the attention change induced by introducing τ at layer ℓ satisfies

$$\delta^{(\ell)}_{\mathrm{attn}}(\tau; U, Q, x) = \alpha^{(\ell)}_{x\leftarrow\tau}\, W^{(\ell)}_{O}\, \big(\bar{v}^{(\ell)}_{\tau} - \bar{v}^{(\ell)}_{x;U,Q}\big).$$

Accordingly, the local correction vector at layer ℓ is

$$g^{(\ell)}(\tau; U, Q, x) := \int_0^1 J_{\mathrm{MLP}^{(\ell)}}\Big(\mathrm{Attn}^{(\ell)}_{x}(U, Q, x) + s\, \delta^{(\ell)}_{\mathrm{attn}}(\tau; U, Q, x)\Big)\delta^{(\ell)}_{\mathrm{attn}}(\tau; U, Q, x)\, \mathrm{d}s,$$

and the overall correction at the final layer is still

$$g^{(L)}(\tau; U, Q, x) = \sum_{\ell=1}^{L} \tilde{g}^{(\ell)}(\tau; U, Q, x),$$

where each $\tilde{g}^{(\ell)}$ is the propagated version of the local correction from layer ℓ to the final layer, defined exactly as in the main text.

Therefore, Corollary 4 and Corollary 5 in the main text continue to hold after replacing the single-token value difference

$$\Delta v^{(\ell)}(t; T, x) = v^{(\ell)}_{t} - \bar{v}^{(\ell)}_{x,T}$$

with the span-level value difference

$$\Delta v^{(\ell)}(\tau; U, Q, x) := \bar{v}^{(\ell)}_{\tau} - \bar{v}^{(\ell)}_{x;U,Q}.$$

In particular, we obtain

$$\|g^{(L)}(\tau; U, Q, x)\| \le \sum_{\ell=1}^{L}\Big(\prod_{j=\ell+1}^{L} L^{(j)}_{\mathrm{blk}}\Big) L^{(\ell)}_{\mathrm{MLP}} \alpha^{(\ell)}_{x\leftarrow\tau} \|W^{(\ell)}_{O}\| \|\Delta v^{(\ell)}(\tau; U, Q, x)\|.$$

Hence, even when both the user query and the provided external information contain multiple tokens, the main conclusions of this paper are unchanged. The only essential modification is that the corrective effect should be measured at the level of *attention-weighted token groups* rather than individual tokens. Under this view: *(i)* relevance becomes the total attention mass assigned to the information span, and *(ii)* complementarity becomes the distance between the span-level virtual value and the current effective state. The same direction-magnitude criterion still determines whether external information is properly corrective, under-corrective, or over-corrective.

F Experimental Datasets

MATH MATH is a competition-style mathematics benchmark designed to measure advanced symbolic and multi-step problem solving. It contains 12.5k problems drawn from high-school math contests and textbooks, spanning multiple subjects (e.g., algebra, geometry, number theory, counting & probability, etc.) and difficulty levels. Each example provides a problem statement together with a full, human-written step-by-step solution and a final boxed answer, which supports evaluating both final-answer correctness and intermediate reasoning. In LLM evaluation, MATH is typically scored by exact match on the final answer (often after normalization). We employ the MATH500 [20] version for our experiments.

CHAMP CHAMP (Concept and Hint-Annotated Math Problems) is a competition-level math dataset built for fine-grained analysis of how models use contextual information. Beyond the original problem and gold solution/answer, each instance is annotated with (1) concepts—general mathematical facts or principles that may be relevant—and (2) hints—problem-specific tricks or guidance that can materially help solve the problem. This structure enables controlled experiments such as adding helpful hints, injecting misleading concepts, or testing whether models can benefit from the "right" side information. CHAMP is also used to study solution faithfulness and verification, not just final-answer accuracy.

TheoremQA TheoremQA is a theorem-driven question answering benchmark that tests whether models can apply domain knowledge (i.e., specific theorems) to solve challenging problems. It consists of expert-curated questions covering hundreds of theorems across multiple disciplines, including mathematics, physics, EE/CS, and finance. Problems are designed so that solving them requires identifying and correctly using an appropriate theorem rather than relying on surface patterns. The dataset emphasizes reasoning and knowledge application: models must connect the question to relevant formal principles and carry out multi-step derivations to reach the final answer. Evaluation commonly focuses on final-answer accuracy (often exact match), with optional analysis of reasoning quality.

MMLU-Redux MMLU-Redux is a cleaned, manually re-annotated variant of the widely used MMLU benchmark, aiming to correct ambiguities and label errors that can distort model comparisons. It preserves the broad subject coverage of MMLU (spanning dozens of academic and professional domains) while providing revised gold answers and improved quality control. The benchmark is intended for more reliable measurement of factual recall and reasoning in a multiple-choice setting, and it is particularly useful for checking whether model rankings are robust to dataset noise. In practice, evaluation remains standard multiple-choice accuracy, but conclusions are less sensitive to flawed items.

GPQA GPQA is a graduate-level, "Google-proof" multiple-choice QA benchmark written by domain experts to be exceptionally difficult even with unrestricted web access. It focuses on deep scientific reasoning in three core areas: biology, chemistry, and physics. The questions are designed to resist shallow retrieval and require substantial conceptual understanding and careful elimination of distractors. Because humans without specialized expertise struggle on the dataset, GPQA is often used to stress-test frontier models and to study oversight and evaluation reliability on hard scientific questions. Performance is typically reported as multiple-choice accuracy, sometimes with analysis by domain and difficulty.

NaturalQuestions NaturalQuestions (NQ) is a large-scale QA dataset built from real, anonymized Google search queries paired with Wikipedia pages from top search results. For each question-page pair, human annotators label a long answer (usually a paragraph or section) and a short answer (one or more spans/entities), or mark the question as unanswerable from the page. This design supports both reading comprehension and retrieval-style evaluation, since the system must identify evidence within long documents. NQ includes diverse information-seeking questions and naturally occurring ambiguity, making it a strong benchmark for open-domain QA pipelines. Common metrics include exact match and token-level F1 for short answers, plus long-answer accuracy.

FinQA FinQA is a financial question answering dataset targeting numerical reasoning over real corporate reports. Each instance pairs a question with evidence drawn from heterogeneous sources such as tables and accompanying textual context. Solving typically requires multi-step arithmetic, aggregation, and logical operations (e.g., computing growth rates, margins, or differences across periods). A key feature of FinQA is the availability of gold "reasoning programs" (structured operation sequences) that make the computation process explicit, enabling explainable evaluation and program-based training or analysis. Systems are commonly evaluated by final numeric answer accuracy (often exact match after normalization), and sometimes by program correctness or execution validity.

Table 5: The prompt of inference during our experiments. {information} can be replaced with demonstrations, retrieved knowledge, and memory.

Prompt of Inference

```
{task_definition}

--

Here are some information to help you solve the given task:
‘‘‘
{information}
‘‘‘

--

Based on the above information, solve the following question:
{question}
```

G Experimental Setup

In this paper, for all three types of external information, we use BM25 to retrieve information relevant to the user query.

The inference prompt we used is shown in Table 5. Specifically, our method designs for different types of information are as follows:

In-Context Learning We use a sample ICL approach, where the input consists of the task definition, demonstrations, and the user query, and the output is the answer to the query. The demonstration pool is the training set for each dataset.

Retrieval-Augmented Generation We directly use the retrieval corpus provided by each dataset and simply concatenate the retrieved results into the input.

Memory Evolution Following [45], we implement evolution with a simple agent system consisting of two components: a Thinker and a Reasoner. The Thinker learns from historical interactions and summarizes experiences, while the Reasoner answers user queries based on the summarized experience. The Thinker prompt we used is shown in Table 6.

H Calculation of Main Parameters

In this section, we introduce how to compute each analysis parameter in §4. To reduce the experimental cost, we calculate the following parameters with the first layer of the model, as we consider the first layer as a single Transformer block.

- $e(x)$ and $e(t, x)$: We directly compute the error as the difference between the pooling of the encoding vectors of each output token and the corresponding token in the ground-truth answer, under the inputs x and $[t, x]$, respectively.
- $g(t, x)$: Our computation follows the definition of $g(t, x)$, where $\mathrm{Attn}_x([x])$ and $\delta_{\mathrm{attn}}(t, x)$ can be obtained directly from the model parameters. For J_σ in J_{MLP}, we derive the Jacobian matrices of the activation functions for each LLM accordingly.
- $\rho(t, x)$: Based on the definition in Corollary 3, we can compute it directly from $e(x)$ and $g(t, x)$.
- $\alpha_{x \leftarrow t}$: We directly extract the attention weights from x to t when feeding $[t, x]$ into the model.

I Effect of Other Structure of Transformer

In this section, we analyze how different factors not discussed in §3 affect the conclusion of this paper.

Table 6: The prompt of Thinker during memory evolution.

Prompt of Thinker using Memory Evolution

```
You are an expert meta-reasoning coach for AI agents.

You will receive a SINGLE data sample from an agent's past experience.
This sample may include:  a task description, the agent's
thoughts/actions/observations,
and whether the attempt succeeded or failed.

Your job is to distill this ONE example into ONE HIGH-LEVEL INSIGHT:
a general rule or strategy that would be useful for similar future tasks.

================ INPUT SAMPLE ================
[Task description]
{TASK_DESCRIPTION}

[Episode log / trajectory]
{EPISODE_LOG}
==============================================

Please now produce exactly ONE INSIGHT that:

- Is HIGH-LEVEL and GENERAL.
- It should be a reusable principle that could help on future, similar tasks.
- It should focus on better thinking and acting (e.g., how to search, plan, verify,
explore).

- Is NOT tied to superficial details of this specific sample.
- Do NOT mention specific names, URLs, IDs, or exact numbers from the log
unless they are essential to the principle.
- Do NOT describe the episode step-by-step.
- Do NOT say ‘‘In this trial...’’ or ‘‘In this example...’’.
Instead, phrase it as ‘‘When solving tasks like this, you should...’’.

- Captures what the agent should DO DIFFERENTLY or KEEP DOING.
- If the attempt failed:  focus on what change of strategy would most likely prevent
this failure.
- If the attempt succeeded:  focus on what strategy made it robust and should be
repeated.

Output format (VERY IMPORTANT):

INSIGHT:
- <one single sentence (or two short sentences) stating the general rule>

Do NOT output anything else.  No explanations, no justification, no bullet list of
multiple rules.
Just the single ‘‘INSIGHT:’’ line as specified above.
```

Multi-Head Attention. Multi head attention keeps the same decomposition view because the model still produces an attention output at position x, and the information-corrective vector is simply the composition of the per head contributions through the output projection. **Overall, MHA does not change our main conclusion.**

Attention Variants. Sharing keys and values across heads mainly redistributes where relevance and complementarity reside across heads, while the information-corrective vector is still governed by how much attention is assigned to the context and how informative the context values are in the projected space. **In short, these variants preserve the key conclusion.**

Table 7: The relative error ($\frac{|\texttt{gold_value - pred_value}|}{\texttt{gold_value}}$) of the trained predictor under different settings.

Dataset	Method	Llama-3.1-8B	Qwen3-8B
MATH	ICL	33.8%	39.4%
	Mem	38.3%	25.5%
CHAMP	RAG	27.0%	23.9%
	Mem	34.4%	36.7%
TheoremQA	RAG	26.6%	26.0%
	Mem	28.8%	29.5%
MMLU-Redux	Mem	31.4%	33.4%
GPQA	ICL	34.4%	21.1%
	Mem	37.4%	25.5%
NQ	Mem	39.3%	33.5%
FinQA	ICL	25.2%	38.1%
	Mem	38.1%	27.1%

Feed Forward Network Variants. Gating and alternative nonlinearities may alter the local amplification or attenuation of an update, but error reduction still hinges on whether the induced vector aligns with the negative original error and whether its magnitude is not excessively large. **Thus, the main conclusion remains unchanged.**

Positional Encoding and Attention Bias. These mechanisms mostly reshape attention weights and, therefore, affect relevance. The same relationship between vector direction, vector norm, and the possibility of error reduction continues to apply. **Consequently, our central conclusion still holds.**

Masking and Long Context Visibility Constraints. Visibility constraints decide whether information tokens can contribute at all, but whenever they are visible, their impact on the error can still be interpreted through the same vector based lens. **So they do not overturn the main conclusion either.**

J Performance of Trained Predictor

The relative error of the trained predictor is shown in Table 7.

K Effects of Different Factors on Method Performance

This section studies factors that affect the performance of the theory-guided information-selection method in Section 5. We examine the generalization and accuracy of the learned correction-direction predictor, the contribution of the two theory-derived selection components, and comparisons with stronger selection baselines under the same candidate pool and computational budget.

K.1 Effects of Predictor Generalization and Accuracy

Section 5 uses a learned predictor for the correction-error alignment score $\rho(t, x)$ because the ground-truth answer is unavailable at inference time. We further examine whether this predictor transfers across models and datasets and quantify the gap between predicted and oracle correction-direction filtering.

Cross-model transfer. Table 8 reports downstream accuracy when the predictor trained for one model is applied to another target model. Cross-model transfer reduces downstream accuracy by 0.6–1.3 percentage points across the evaluated settings.

Cross-dataset transfer. Table 9 compares the in-domain setting with the cross-dataset setting. Cross-dataset transfer reduces accuracy by 0.5–1.4 percentage points across the seven benchmarks.

Table 8: Cross-model transfer of the learned correction-direction predictor. Values are average downstream accuracy (%) for each external-information setting.

Predictor Source	Target Model	ICL	RAG	Mem	Avg.
Llama-3.1-8B	Llama-3.1-8B	49.7	49.8	42.0	45.6
Qwen3-8B	Llama-3.1-8B	48.8	48.6	40.9	44.5
Qwen3-8B	Qwen3-8B	75.1	64.6	62.9	66.1
Llama-3.1-8B	Qwen3-8B	74.5	63.3	62.0	65.2

Table 9: Cross-dataset generalization of the learned correction-direction predictor. Values are downstream accuracy (%). "Other Datasets" denotes the cross-dataset training setting.

Target Dataset	In-Domain	Other Datasets
MATH	69.6	68.3
CHAMP	34.1	33.1
TQA	49.7	49.0
MMLU-Redux	68.5	68.0
GPQA	53.9	53.0
NQ	61.9	60.5
FinQA	59.5	58.6

Predicted vs. oracle filtering. Table 10 compares three variants: *No-filter* removes the $\rho(t, x)$-based direction filter, *Predicted* uses the learned predictor, and *Oracle* uses the ground-truth $\rho(t, x)$. Predicted filtering improves upon removing the direction filter, and oracle filtering yields the highest accuracy in each setting.

K.2 Effects of the Selection Components

Our selection strategy contains two theory-derived components: a correction-direction filter based on the predicted $\rho(t, x)$ and a corrective-strength proxy $\alpha_{x \leftarrow t}\|v_t - v_x\|$. We ablate the two components to quantify their individual contributions. In all three external-information settings, using both components consistently outperforms using either component alone. The direction filter is also generally stronger than the strength proxy alone, consistent with the observation in Section 5.1 that incorrect correction direction is a major source of harmful external information.

K.3 Comparison with Stronger Selection Baselines

We further compare the theory-guided selector with stronger and more recent context-selection baselines under the same candidate pool and computational budget. As in Section 5, each query starts from 32 coarsely retrieved candidates and all methods ultimately select top-1 external information for inference. Tables 14-16 report the results for ICL, RAG, and memory-based settings, respectively. Our method achieves the highest or tied-highest accuracy in 25 of the 26 model–benchmark combinations.

Table 10: Effect of correction-direction estimation on downstream accuracy (%).

Model	Setting	No-filter	Predicted	Oracle
Llama-3.1-8B	ICL	48.6	49.7	**54.0**
	RAG	48.1	49.8	**55.2**
	Mem	40.8	42.0	**46.3**
	Average	44.3	45.6	**50.1**
Qwen3-8B	ICL	74.2	75.1	**79.0**
	RAG	63.2	64.6	**69.8**
	Mem	62.2	62.9	**67.4**
	Average	65.2	66.1	**70.7**

Table 11: Ablation of the two selection components in ICL. Values are accuracy (%).

Model	Variant	MATH	GPQA	FinQA
Llama-3.1-8B	Direction Filter Only	49.2	45.6	52.7
	Strength Proxy Only	48.5	44.3	52.2
	Both	**49.6**	**46.5**	**53.1**
Qwen3-8B	Direction Filter Only	89.6	65.5	69.0
	Strength Proxy Only	89.0	65.3	68.2
	Both	**90.2**	**65.7**	**69.5**

Table 12: Ablation of the two selection components in RAG. Values are accuracy (%).

Model	Variant	CHAMP	TQA	NQ
Llama-3.1-8B	Direction Filter Only	26.6	41.9	78.9
	Strength Proxy Only	25.9	40.9	77.8
	Both	**27.4**	**42.5**	**79.4**
Qwen3-8B	Direction Filter Only	47.0	66.7	78.2
	Strength Proxy Only	45.8	65.5	76.2
	Both	**47.8**	**67.2**	**78.9**

Table 13: Ablation of the two selection components in the memory setting. Values are accuracy (%).

Model	Variant	MATH	CHAMP	TQA	MMLU	GPQA	NQ	FinQA
Llama-3.1-8B	Direction Filter Only	48.9	17.6	29.8	66.5	34.3	41.1	51.4
	Strength Proxy Only	47.9	15.8	28.2	65.8	32.8	40.3	50.5
	Both	**49.6**	**18.1**	**30.4**	**67.1**	**34.8**	**41.7**	**52.1**
Qwen3-8B	Direction Filter Only	88.7	42.2	57.8	69.1	67.6	47.1	62.7
	Strength Proxy Only	88.5	41.3	56.5	67.4	66.7	45.8	61.9
	Both	**89.0**	**43.0**	**58.7**	**69.9**	**68.7**	**47.5**	**63.4**

Table 14: Comparison with stronger ICL/context-selection baselines under the same candidate pool and computational budget. Values are accuracy (%).

Model	Method	MATH	GPQA	FinQA
Llama-3.1-8B	Linear-Time DS	48.8	44.4	52.4
	AutoSelect	49.1	45.9	52.8
	DiSP	49.3	46.0	52.7
	Ours	**49.6**	**46.5**	**53.1**
Qwen3-8B	Linear-Time DS	89.3	65.5	67.8
	AutoSelect	89.8	65.6	69.3
	DiSP	**90.2**	65.6	68.8
	Ours	**90.2**	**65.7**	**69.5**

Table 15: Comparison with stronger RAG/context-selection baselines under the same candidate pool and computational budget. Values are accuracy (%).

Model	Method	CHAMP	TQA	NQ
Llama-3.1-8B	Qwen3-Embedding-0.6B	25.9	41.2	78.1
	Qwen3-Reranker-0.6B	26.6	41.6	78.7
	jina-reranker-v3	26.8	42.0	78.9
	Ours	**27.4**	**42.5**	**79.4**
Qwen3-8B	Qwen3-Embedding-0.6B	46.7	66.2	77.5
	Qwen3-Reranker-0.6B	47.2	66.5	78.1
	jina-reranker-v3	47.4	66.5	78.5
	Ours	**47.8**	**67.2**	**78.9**

Table 16: Comparison with stronger memory-selection baselines under the same candidate pool and computational budget. Values are accuracy (%).

Model	Method	MATH	CHAMP	TQA	MMLU	GPQA	NQ	FinQA
Llama-3.1-8B	MemOS	48.6	16.6	28.7	65.6	33.7	40.7	51.0
	SimpleMem	48.9	17.2	29.6	66.5	34.0	41.6	51.8
	MemGuide	49.4	17.7	29.4	66.8	34.0	41.2	51.7
	Ours	**49.6**	**18.1**	**30.4**	**67.1**	**34.8**	**41.7**	**52.1**
Qwen3-8B	MemOS	88.5	41.5	57.2	68.4	67.4	46.5	62.7
	SimpleMem	88.9	42.3	57.6	69.2	67.8	46.8	62.3
	MemGuide	**89.1**	42.7	58.4	69.8	67.8	47.1	62.8
	Ours	89.0	**43.0**	**58.7**	**69.9**	**68.7**	**47.5**	**63.4**

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have discussed the limitation of this paper in Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We have discussed the assumptions in §3 and proof in Appendix D.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have provided the information for main experimental results in §4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Considering that time is tight, we will release source codes as soon as this paper is accepted.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We have discussed the test details in §4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report our experimental results with five runs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We discuss required computer resources in §4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: All our research conform with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the societal impacts of this paper in Appendix B.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our usage of data and models does not have a high risk for misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All our usage of creators or original owners of assets is properly credited and are the license and terms of use explicitly mentioned and properly respected.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: We do not introduce new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: We do not adapt crowdsourcing experiments and research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We have discussed the usage of LLMs in §4.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.