# ATOMIC: ATTENTION REALLOCATION TO MITIGATE IRRELEVANT CONTEXT IN SMALL LANGUAGE MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Small language models (SLMs) have recently gained increasing attention due to their strong cost efficiency, computational efficiency, and competitive performance across a wide range of reasoning tasks. However, when reasoning inputs contain irrelevant context, SLMs are substantially more vulnerable to distraction than large language models, leading to severe performance degradation. Through a fine grained analysis of inference time attention, we identify a severe underlying issue: SLMs exhibit a markedly weaker ability to distinguish relevant context (RC) from irrelevant context (IC), as evidenced by consistently lower AUROC scores when treating attention over RC and IC as a binary discrimination problem. As a result, SLMs often assign high attention to distracting tokens in irrelevant context and erroneously incorporate them as evidence during reasoning, which disrupts effective evidence aggregation. To address this limitation, we propose ATOMIC. Our approach first performs task conditioned context level filtering to suppress irrelevant information, then introduces an Evidence Likelihood Score to amplify attention on critical evidence tokens within relevant context, and finally reallocates attention during answer inference to further reinforce key evidence. Extensive experiments across multiple mathematical reasoning benchmarks demonstrate that our method achieves state of the art performance. Our code is available at: https://anonymous.4open.science/r/ATOMIC.

## 1 INTRODUCTION

Small Language Models (SLMs) (Belcak et al., 2025; Wang et al., 2025a) offer compelling advantages in cost efficiency, computational efficiency, and deployability compared to Large Language Models (LLMs) (Zhao et al., 2026; Chang et al., 2024; Naveed et al., 2025), and have shown competitive performance across reasoning, scientific, and code generation tasks (Shao et al., 2024; Reddy & Shojaee, 2025; Li et al., 2023a). In certain tasks, SLMs have even been shown to outperform their larger counterparts (Zhang et al., 2024; Hsieh et al., 2023), providing compelling evidence of their practical advantages. Standard formulations of reasoning tasks assume that the input is composed of relevant context (RC) sufficient to support correct inference. However, a critical challenge faced by SLMs is their vulnerability to distraction from irrelevant context (IC) (Li et al., 2023b; Shi et al., 2023; Yang et al., 2025b) in mathematical reasoning tasks (Cobbe et al., 2021; Hendrycks et al., 2021). In this work, we provide a finer-grained analysis showing that this IC-induced failure mode is surprisingly severe in SLMs, stemming from their limited ability to discriminate RC from IC, which causes highly distracting tokens to be erroneously treated as evidence and incorporated into the reasoning process. This misallocation of attention directly interferes with evidence aggregation and leads to substantially degraded reasoning performance. This phenomenon is often attributed to the limited number of parameters in SLMs, which is assumed to constrain their semantic capacity (Yan et al., 2025). Nevertheless, a systematic investigation into the underlying reasons for this significant vulnerability in SLMs remains largely unexplored. This paper seeks to fill this gap by analyzing the root causes of SLMs' sensitivity to irrelevant context and by proposing effective strategies to mitigate the impact of irrelevant contextual information on SLM performance.

The phenomenon of large language models being distracted by IC is still at an early stage of investigation. For example, one line of work (Zhu et al., 2025) attributes the performance drop to the



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

model's failure to attend to the correct reasoning direction, while another study (Wang et al., 2025b) argues that language models struggle to effectively filter out irrelevant information. Other work (Wu et al., 2024) suggests that LLMs rely heavily on semantic similarity rather than factual grounding, making highly related but misleading context particularly disruptive. However, these works largely remain at a behavioral level and do not provide a fine-grained, internal explanation of how distraction arises during inference. More critically, they offer limited insight into why this issue becomes surprisingly severe in SLMs. In addition, as we have experienced, properly evaluating the severity of IC-induced failures itself remains a critical yet unresolved challenge: existing evaluations often rely on coarse end-task accuracy or qualitative examples, which fail to disentangle whether errors stem from flawed evidence selection, spurious attention allocation, or downstream reasoning mistakes. As a result, both the true extent of this vulnerability and its downstream consequences on reasoning reliability remain poorly quantified and insufficiently understood.

To address the limited ability of small language models (SLMs) to distinguish relevant context (RC) from irrelevant context (IC), we propose ***A**ttention Reallocation to **M**itigate **I**rrelevant **C**ontext in Small Language Models (***ATOMIC***)*, a principled inference time framework that explicitly mitigates distraction from IC and enhances evidence utilization from RC. Specifically, we first perform task conditioned context level filtering to proactively suppress irrelevant information and reduce attention allocation to IC. Then, we introduce an Evidence Likelihood Score (ELS) to identify and amplify attention on evidence tokens within RC that are critical for correct reasoning. Importantly, we formulate a principled evaluation protocol that casts RC and IC discrimination as an internal decision problem at inference time, enabling quantitative assessment of a model's ability to separate evidence from distraction. While AUROC is a standard metric, our contribution lies in formulating the objective over token level attention signals at inference time, enabling a direct and quantitative measure of evidence discrimination under irrelevant context, rather than task accuracy. During answer inference, we reallocate attention scores to suppress irrelevant context while reinforcing critical evidence tokens within relevant context, thereby improving reasoning robustness. Our main contributions are summarized as follows:

- **Identification of Irrelevant Context Vulnerability.** We conduct a systematic analysis of how small language models respond to irrelevant context, uncovering fundamental limitations in their ability to distinguish relevant from irrelevant information during reasoning.
- **A Novel Reasoning Framework for Irrelevant Context.** We propose ATOMIC, which suppresses irrelevant context while explicitly extracting evidence tokens within relevant context, thereby improving the reasoning robustness of SLMs in the presence of irrelevant context.
- **Comprehensive Evaluation.** Extensive experiments across multiple benchmarks and model families demonstrate consistent improvements over strong baselines, indicating the effectiveness of the proposed ATOMIC method.

## 2 PRELIMINARIES

### 2.1 NOTATIONS

We adopt standard notation conventions tailored to this work. Lowercase italic letters (e.g., $\alpha$, $\beta$) denote scalar values, while lowercase bold letters (e.g., $\boldsymbol{v}$) represent vectors, such as attention score vectors or hidden state vectors. Uppercase calligraphic letters (e.g., $\mathcal{Q}, \mathcal{R}$) denote sets or spaces. We consider an input space $\mathcal{Q}$ that is explicitly composed of a relevant context and an irrelevant context, such that $\mathcal{Q} = \mathcal{Q}^{\text{rel}} \cup \mathcal{Q}^{\text{irr}}$. Each input sequence $Q \in \mathcal{Q}$ consists of a collection of tokens, $Q = \{q_1^{\text{rel}}, \ldots, q_k^{\text{rel}}, \ q_1^{\text{irr}}, \ldots, q_l^{\text{irr}}\}$, where $q_i^{\text{rel}}$ denotes tokens from the relevant context and $q_j^{\text{irr}}$ denotes tokens from the irrelevant context. A language model is denoted by a function $f$, which maps an input token sequence to an output sequence, $f : \mathcal{Q} \to \mathcal{R}$. The output space $\mathcal{R}$ contains token sequences $R = (r_1, r_2, \ldots, r_m)$, where each $r_i$ represents an output token generated by the model. During inference, the language model produces intermediate representations such as token-level attention distributions. We use uppercase bold italic letters to denote matrices (e.g. $\boldsymbol{A}$), such as the attention matrix at a specific layer and head, whose entries represent attention weights between token pairs. We use lowercase bold italic letters to denote vectors (e.g. $\boldsymbol{v}$), such as the attention score vector of a particular response token with respect to the input tokens, with each component corresponding to one input position.

### 2.2 DEFINITIONS

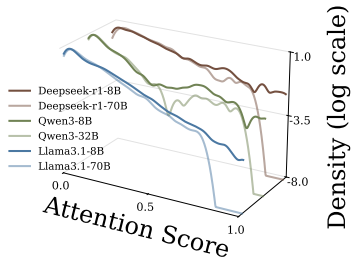


Figure 1: IC attention-score distributions across LLMs and SLMs.

During language model inference, the input context often contains a mixture of information that is necessary for solving the task and information that is incidental, redundant, or unrelated to the target reasoning objective. We refer to such unrelated or non essential information as *irrelevant context* (IC). While an ideal language model should selectively attend to task relevant context and ignore IC, in practice large and especially SLMs may allocate non negligible attention or internal capacity to IC. This can interfere with evidence aggregation, distort intermediate representations, and ultimately lead to incorrect outputs.

**Definition 2.1** (Irrelevant Context)**.** Let a language model $f$ map an input token sequence $Q = (q_1, q_2, \ldots, q_n) \in \mathcal{Q}$ to an output $R = f(Q) \in \mathcal{R}$. The *irrelevant context* is defined as a subset of tokens $Q^{\text{irr}} \subseteq Q$ such that removing $Q^{\text{irr}}$ from the input does not affect the correct reasoning outcome of the task.

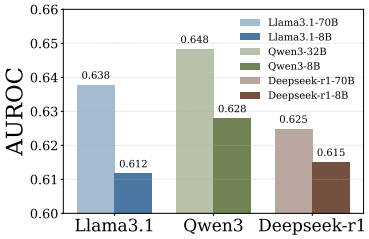


Figure 2: Discrimination between RC and IC, measured by AUROC.

In mathematical reasoning tasks (Shi et al., 2023; Wu et al., 2024; Amini et al., 2019), only a subset of input tokens directly contributes to deriving the correct solution. These tokens encode information such as quantities, relations, or constraints that are essential to the underlying task formulation and must be correctly utilized to produce a valid answer. For example, in a mathematical reasoning task, such tokens may correspond to numerical values, equations, or relational statements that are required for correct problem solving. We refer to such tokens as *evidence tokens*. For example, in the problem *"Steve is 5 feet 6 inches. He grows 6 inches. The height of Emma is 8 feet. How tall is Steve in inches?"*, the tokens corresponding to *"5 feet 6 inches.* and *"6 inches"* constitute evidence tokens, as they are directly involved in the required computation. In contrast, the statement *"The height of Emma is 8 feet"* introduces numerically salient but task irrelevant information and therefore does not contain evidence tokens for this task. SLMs may nevertheless attend to such irrelevant numerical tokens, which can mislead the reasoning process and result in incorrect answer.

**Definition 2.2** (Evidence Token)**.** Let a language model $f$ map an input token sequence $Q = (q_1, q_2, \ldots, q_n) \in \mathcal{Q}$ to an output $R = f(Q) \in \mathcal{R}$. A token $q_i \in Q$ is an *evidence token* if it encodes information that is required to derive the task correct output $R^\star$. Equivalently, removing $q_i$ from input renders the task underdetermined or changes the correct solution.

### 2.3 TOKEN CONTRIBUTION DURING INFERENCE

To analyze how input tokens influence the final prediction during inference, we quantify token contribution using the attention from the final answer token to all input tokens. For each input token, we aggregate its attention weights from the final answer token across all transformer layers and attention heads into a single scalar score, which serves as a unified measure of token contribution. Figure 1 shows that these attention scores are highly sparse: most tokens receive near zero attention, while only a small subset obtain substantially larger scores. This indicates that the model relies on only a limited number of tokens when producing the final answer. Based on these token level attention scores, we further assess whether the model can distinguish relevant context, RC, from irrelevant context, IC. Since each token is labeled as either RC or IC, we treat its aggregated attention score as a continuous decision statistic and compute the AUROC, with RC tokens as positives and IC tokens as negatives. We refer to this metric as *AUROC over Attention Scores*. A higher value indicates stronger separation between RC and IC tokens during inference. As shown in Figure 2, large language models consistently achieve higher *AUROC over Attention Scores* than small language models. This suggests that small language models have weaker evidence discrimination and are therefore more easily distracted by IC during reasoning.

# 3 METHODOLOGY

We introduce a three-stage inference-time framework to improve attention discrimination between relevant context (RC) and irrelevant context (IC) in small language models. Stage I suppresses IC in a task-conditioned manner to reduce spurious attention. Stage II identifies and amplifies evidence tokens that are critical for correct reasoning. Stage III softly reallocates attention scores during answer inference to further suppress residual influence from IC. Together, these stages progressively align attention allocation with evidence relevance, leading to more robust reasoning.

## 3.1 OPTIMIZING ATTENTION DISCRIMINATION

Since the proposed *AUROC over Attention Scores* (Section 2.3) is a non differentiable evaluation measure and cannot be directly optimized, we instead introduce a surrogate objective that explicitly captures the goal of improving attention based discrimination between RC and IC. Our formulation is motivated by the inherent sparsity of attention scores, where only a small fraction of tokens receive high attention during inference. Leveraging this property, we propose a likelihood ratio based attention threshold to partition attention scores into high and low regimes, enabling a tractable optimization objective.

**Likelihood Ratio Attention Threshold.** Let $s \in \mathbb{R}$ denote the attention score assigned to an input token at inference time. We estimate the conditional density functions $p(s \mid RC)$ and $p(s \mid IC)$ from attention scores associated with relevant and irrelevant context tokens, respectively. Based on these densities, we define the log likelihood ratio (LLR) as $\mathrm{LLR}(s) = \log \frac{p(s|RC)}{p(s|IC)}$. We then define a likelihood ratio attention threshold by identifying the first point on the right side of the IC density peak where the likelihoods of RC and IC are equal: $t^* = \min \{s > \arg\max_s p(s \mid IC) \mid \mathrm{LLR}(s) = 0\}$. This construction ensures that tokens classified as having high attention scores lie in a region where the IC density is low relative to RC, thereby reducing the risk of assigning excessive attention to irrelevant context. We note several corner cases that further elucidate the behavior of this threshold. If, on the right side of the IC density peak, the IC density rises above or remains consistently higher than the RC density, the majority of high attention mass is assigned to IC tokens, indicating a failure mode where the model is dominated by irrelevant information. In contrast, when IC density remains consistently lower than RC density beyond the peak, the model exhibits near ideal discrimination between RC and IC, corresponding to strong reasoning robustness.

**Surrogate Objective.** With the threshold $t^*$ defined, we can discard the low score region dominated by sparse attention and focus on the high score regime that directly influences reasoning outcomes. Specifically, by categorizing tokens along two axes, context type (RC versus IC) and attention level (high versus low based on $t^*$), we construct a standard confusion matrix over attention assignments. Based on this matrix, we adopt two complementary metrics as surrogate objectives. The *True Negative Ratio (TNR)* measures the proportion of IC tokens that are correctly assigned low attention scores, with higher values indicating more effective suppression of irrelevant information. The *Precision* measures the proportion of RC tokens among all high attention tokens, reflecting how well the model concentrates attention on relevant evidence. By jointly maximizing TNR and Precision, we obtain a principled surrogate objective that encourages suppression of IC while amplifying attention to RC, providing a tractable and interpretable optimization target for attention reallocation.

## 3.2 STAGE I: IRRELEVANT CONTEXT SUPPRESSION

In this stage, we aim to suppress irrelevant context (IC) at the context level while ensuring that all retained tokens remain faithful to the original input. Rather than filtering tokens based on heuristic thresholds or hard decisions, we rewrite the input context into a compact representation that is fully consistent with the original token sequence and conditioned on the question. The resulting distilled context contains only information that is relevant to the reasoning objective, and serves as the input to subsequent token-level stages.

We consider an input sequence $Q \in \mathcal{Q} = \mathcal{Q}^{\mathrm{rel}} \cup \mathcal{Q}^{\mathrm{irr}}$, which consists of both relevant and irrelevant context tokens. Our goal in Stage I is to construct a new context $\tilde{Q} \subseteq Q$ that preserves all information necessary for answering the question while removing tokens originating from irrelevant context. Importantly, we require that every token in $\tilde{Q}$ corresponds to a token in the original

input sequence, ensuring token-level consistency and avoiding the introduction of any external or hallucinated content.

We then formalize this process as a question-conditioned context distillation problem. Let $f : \mathcal{Q} \to \mathcal{R}$ denote the language model used for inference. The distilled context $\tilde{Q}$ is defined as the minimal subset of tokens from $Q$ that is sufficient for producing an answer consistent with the original input:

$$\tilde{Q} \;=\; \arg\min_{Z \subseteq Q} |Z| \quad \text{s.t.} \quad f(Z) \approx f(Q). \tag{1}$$

The constraint $f(Z) \approx f(Q)$ indicates that the output generated from the distilled context should remain unchanged or equivalent to that produced from the full input, thereby enforcing sufficiency. By optimizing the objective in Eq. 1, Stage I yields a distilled context $\tilde{Q}$ that contains only question-consistent information. Tokens belonging to irrelevant context are excluded because they do not contribute to maintaining output consistency under the minimality constraint. We therefore treat $\tilde{Q}$ as the refined relevant context, which significantly reduces the number of irrelevant tokens entering subsequent token-level evidence scoring and attention reallocation stages. This design improves TNR by construction, while preserving task-relevant evidence required for correct reasoning.

### 3.3 Stage II: Evidence Likelihood Scoring

After context-level compression in Section 3.2, we operate on the distilled input $\tilde{Q} \subseteq Q$, which primarily contains tokens from the refined relevant context. At this stage, we assign each token $q \in \tilde{Q}$ an *evidence likelihood score* (ELS), which estimates how likely the token constitutes true reasoning evidence under the given query. Formally, let $\mathcal{E}^\star \subseteq \tilde{Q}$ denote the latent set of tokens that are necessary for correct reasoning. We define the evidence likelihood score of a token $q$ as the posterior log-odds of $q$ being a true evidence token:

$$\mathrm{ELS}(q) \;\triangleq\; \log \frac{p\Big(q \in \mathcal{E}^\star \mid \tilde{Q}\Big)}{p\Big(q \notin \mathcal{E}^\star \mid \tilde{Q}\Big)}. \tag{2}$$

The true posterior is intractable at inference time; therefore, our objective is not to estimate Eq. 2 exactly, but to construct a scoring function that preserves the relative ordering of tokens by their evidence likelihood. The posterior log-odds in Eq. 2 can be rewritten, up to an additive constant, as a log-likelihood ratio between evidence and non-evidence tokens. We approximate this quantity using a linear surrogate: $\mathrm{ELS}(q) \;\approx\; \sum_k w_k \, \phi_k(q, \tilde{Q})$, where $\{\phi_k\}$ are token-level features computed with respect to the distilled context $\tilde{Q}$, and $\{w_k\}$ are fixed weights. The feature set is designed to capture complementary signals of whether a token serves as reasoning evidence. Specifically, it combines semantic relevance to the problem, contextual salience within the refined context, positional prominence, and token-type information. All features are computed deterministically from $q$ and $\tilde{Q}$, without any external supervision. Through this linear composition, tokens that are more likely to be true evidence receive higher $\mathrm{ELS}(q)$ scores, so that under a fixed evidence selection budget, ranking by $\mathrm{ELS}(q)$ prioritizes critical evidence tokens for subsequent inference.

### 3.4 Stage III: Soft Attention Score Reallocation

This stage integrates the outputs of Stage I and Stage II. After irrelevant context has been suppressed through question-conditioned distillation, and ELS have been computed for tokens in the refined context $\tilde{Q}$, we perform a soft reallocation of token-level attention scores during inference. The goal is to bias the model toward attending to high-evidence tokens without enforcing hard selection decisions that may restrict reasoning flexibility in small language models. Let $s(q)$ denote the original attention score assigned to token $q \in \tilde{Q}$ during inference. We define a reallocated score $\tilde{s}(q) = s(q) \cdot w(q)$, where the reweighting function $w(q)$ is defined as a monotonic function of the evidence likelihood score:

$$w(q) = g(\mathrm{ELS}(q)), \tag{3}$$

with $g(\cdot)$ being a non-decreasing function. Since Stage I produces a refined context $\tilde{Q}$ that primarily contains question-consistent tokens, the reallocation operates entirely within $\tilde{Q}$. Tokens with higher

evidence likelihood receive amplified attention scores, while tokens with low evidence likelihood are softly suppressed. This soft reallocation preserves the relative ranking of attention within the refined context, while increasing the separation between evidence and non-evidence tokens. At test time, inference proceeds using the reallocated scores $\tilde{s}(q)$ in place of the original attention scores. Through this evidence token extraction mechanism, high-evidence tokens are more likely to dominate the high-attention regime, thereby improving precision. By jointly suppressing IC and extracting evidence scores, ATOMIC simultaneously improves both *TNR* and *Precision*, leading to enhanced reasoning performance in SLMs.

# 4 Theoretical Analysis

In this section, we provide a theoretical analysis of the core component of our method, *Evidence Likelihood Score* (ELS). Our goal is to justify the design of ELS by showing that, under a reasoning capacity assumption, optimal evidence selection reduces to a ranking problem over tokens. The formal assumption is stated in Appendix A.1. Let $\mathcal{Q}$ denote the set of input tokens, and let $\delta : \mathcal{Q} \to \{0, 1\}$ be a binary selection function, where $\delta(q) = 1$ indicates that token $q$ is selected as reasoning evidence. Under the reasoning capacity assumption, evidence selection is subject to a fixed budget: $\sum_{q\in\mathcal{Q}} \delta(q) = k$. Let $y(q) \in \{0, 1\}$ indicate whether token $q$ is a true evidence token. Since Precision is the proportion of true evidence tokens among the selected $k$ tokens, maximizing Precision is equivalent to maximizing the expected number of true evidence tokens selected. This leads to the following proposition.

**Proposition 4.1** (Precision Maximization)**.** *Under the fixed evidence selection budget* $\sum_{q\in\mathcal{Q}} \delta(q) = k$*, maximizing Precision is equivalent to solving* $\max_\delta \mathbb{E}\Big[\sum_{q\in\mathcal{Q}} \delta(q)\, y(q)\Big]$.

The expectation is taken with respect to the randomness of $y(q)$. Let $p(q) \triangleq \mathbb{P}\big(y(q) = 1 \mid \tilde{Q}\big)$ denote the posterior probability that token $q$ is a true reasoning evidence token, conditioned on the refined context $\tilde{Q}$. By linearity of expectation, the objective in Proposition 4.1 can be rewritten as $\max_\delta \sum_{q\in\mathcal{Q}} \delta(q)\, p(q) \quad \text{s.t.} \quad \sum_{q\in\mathcal{Q}} \delta(q) = k$. This formulation reveals that evidence selection under a fixed budget is fundamentally a ranking problem over tokens according to their posterior evidence probabilities. We formalize the optimal solution in the following theorem.

**Theorem 4.2** (Optimal Evidence Selection)**.** *Let* $p(q) \triangleq \mathbb{P}\big(y(q) = 1 \mid \tilde{Q}\big)$ *denote the posterior probability that token* $q$ *is true evidence. Given budget* $k$*,* $\delta^\star$ *is optimal in expectation if it selects the* $k$ *tokens with the largest posterior probabilities* $p(q)$*. Equivalently,* $\delta^\star$ *preserves the ordering induced by* $p(q)$*.*

Theorem 4.2 shows that, under a fixed budget, optimal evidence selection depends only on the ranking induced by the posterior evidence probability $p(q)$. Thus, any scoring function that preserves this ranking achieves the optimal expected Precision. Our $\mathrm{ELS}(q)$ is designed as a surrogate for the posterior log odds, and Appendix A.3 shows that it preserves the posterior ranking up to a bounded error. Therefore, $\mathrm{ELS}(q)$ yields near optimal expected Precision, providing theoretical justification for our design.

# 5 Experiment

In this section, we conduct a comprehensive evaluation of our proposed ATOMIC. Specifically, we aim to answer the following research questions (RQs): **RQ1**: What is the overall performance under irrelevant context? **RQ2**: How well are attention discrimination improvements aligned with downstream performance? **RQ3**: How effective is each component of the proposed method? We provide detailed experimental analyses addressing these questions in the following subsections.

## 5.1 Experimental Setup

**Models.** We conduct extensive evaluations across multiple model families, each comprising a corresponding LLM and SLM, namely Qwen3-(8B, 32B) (Yang et al., 2025a), Llama3.1-(8B, 70B) (Grattafiori et al., 2024), and Deepseek-r1-(8B, 70B) (Guo et al., 2025). This setup enables

Table 1: Performance comparison between ATOMIC and baseline methods on the GSM-IC (Shi et al., 2023) and GSM-DC (Wu et al., 2024) benchmarks. Experiments are conducted using SLMs with at most 8B parameters from three model families: Qwen3-8B (Yang et al., 2025a), Llama3.1-8B (Grattafiori et al., 2024), and Deepseek-r1-8B (Guo et al., 2025). For brevity, we refer to each SLM by its corresponding model family name in the following experiments. Results are Pass@1 accuracy in percentage (%). The best performance is highlighted in **bold**.

| | GSM-IC (two-step) | | | GSM-IC (multi-step) | | | GSM-DC | | |
|---|---|---|---|---|---|---|---|---|---|
| Method | Qwen | Llama | Deepseek | Qwen | Llama | Deepseek | Qwen | Llama | Deepseek |
| Vanilla | 31.50 | 14.93 | 60.15 | 9.025 | 6.300 | 62.28 | 20.88 | 5.425 | 19.25 |
| CoT | 44.75 | 22.75 | 61.25 | 15.88 | 4.125 | 70.75 | 19.90 | 13.12 | 21.13 |
| LTM | 55.40 | 26.18 | 67.23 | 20.30 | 14.68 | 76.10 | 21.73 | 10.78 | 19.88 |
| Program | 94.00 | 66.38 | 73.88 | 84.13 | 58.00 | 60.88 | 14.95 | 12.15 | 14.75 |
| **ATOMIC** | **96.75** | **77.63** | **81.88** | **87.25** | **65.25** | **81.25** | **52.50** | **35.63** | **41.25** |

a comprehensive assessment of the effectiveness of our method, which is specifically designed for SLMs, across diverse architectures, as well as a direct comparison with their LLM counterparts.

**Benchmarks.** We evaluate model robustness under irrelevant context (IC) using two widely adopted benchmarks that explicitly incorporate distracting information: GSM-IC (Shi et al., 2023) and GSM-DC (Wu et al., 2024). Both benchmarks extend arithmetic reasoning tasks by introducing additional IC. Specifically, GSM-IC categorizes reasoning processes into two-step and multi-step reasoning settings, while GSM-DC constructs dependency graphs and controls the strength of IC by varying the number of distractors.

**Baselines.** We compare our approach against a set of representative and widely adopted baseline methods. Specifically, we include Chain-of-Thought (CoT) (Kojima et al., 2022), which introduces explicit reasoning instructions; Least-to-Most (LTM) (Zhou et al., 2022), which decomposes a problem into a sequence of simpler subproblems that are solved progressively; and Program (Chowdhery et al., 2023), which formulates mathematical problems as executable programs and derives the final answer through program execution. For all baseline methods, we explicitly incorporate instructions that encourage the model to ignore irrelevant context, thereby enhancing their robustness to irrelevant information.

### 5.2 Overall Effectiveness

We conduct a comprehensive evaluation of the proposed ATOMIC method against all baseline approaches across different model families and benchmarks. As shown in Table 1, ATOMIC consistently achieves the best performance across all evaluated benchmarks. Following prior work (Shi et al., 2023), all baseline methods are instruction-tuned to explicitly ignore irrelevant context, which leads to noticeable improvements over the Vanilla setting in most cases. Among the baseline approaches, the Program method achieves performance very close to that of ATOMIC on the GSM-IC benchmark. We attribute this behavior to the fact that Program-based reasoning resembles an explicit information selection process, in which conditions relevant to the reasoning task are actively extracted while irrelevant context is implicitly disregarded, resulting in competitive performance. In contrast, the LTM method sometimes exhibits substantial performance degradation. We hypothesize that this is because LTM relies on subquestion decomposition, where errors introduced by irrelevant context during early decomposition stages can propagate to subsequent reasoning steps. We further compare ATOMIC against both LLMs and SLMs within each model family, as illustrated in Fig. 3. The results demonstrate that ATOMIC yields highly impressive gains: SLMs augmented with ATOMIC achieve performance comparable to, and in some cases surpass, their LLM counterparts. This observation suggests that SLMs already possess sufficient

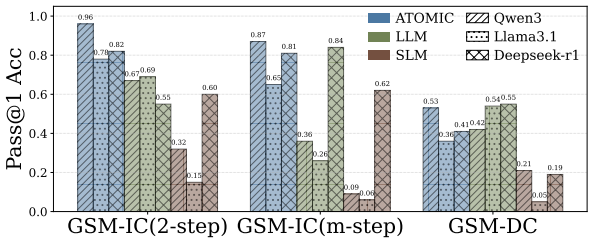


Figure 3: Pass@1 accuracy of ATOMIC and the corresponding SLM and LLM on GSM-IC and GSM-DC.

intrinsic reasoning capability, and that ATOMIC effectively unlocks this potential by mitigating the adverse impact of irrelevant context. Overall, ATOMIC consistently delivers superior performance across model families and benchmarks, providing strong empirical evidence for the effectiveness.

### 5.3 Attention Discrimination Alignment

In this section, we conduct a more detailed analysis of the proposed ATOMIC method, focusing on whether the allocation of attention scores is aligned with our design objective, namely improving attention-based discrimination in the presence of irrelevant context (IC).

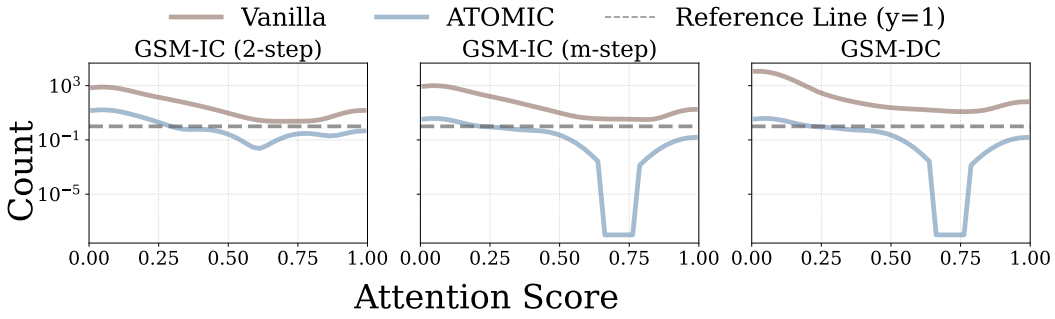


Figure 4: Attention assigned to IC tokens by Vanilla and ATOMIC on GSM-IC and GSM-DC. Histograms are Gaussian-smoothed for clarity; the dashed line marks non-zero attention.

We first analyze the distribution of attention scores, as shown in Fig. 4. We observe that ATOMIC yields substantial improvements over the Vanilla method across all benchmarks. In particular, in the high-attention-score regime, IC tokens are almost entirely absent, indicating that ATOMIC effectively suppresses attention to irrelevant context. This observation provides strong evidence that the proposed IC context suppression mechanism functions as intended. We further analyze the selected *evidence tokens*, with detailed results reported in Appendix B.1. The results show that ATOMIC consistently identifies tokens that are critical during the reasoning process. This selective attention to informative evidence further contributes to improved reasoning performance under IC. Taken together, these findings demonstrate that ATOMIC achieves strong alignment in attention discrimination, effectively distinguishing relevant information from distracting context.

Table 2: Ablation results of ATOMIC on the GSM-DC benchmark across different SLMs within their respective model families. We evaluate the individual contributions of Stage I and Stage II, as well as their combined effect. Results are reported using Pass@1 accuracy.

| Method | Qwen3 | Llama3.1 | Deepseek-r1 |
|---|---|---|---|
| **ATOMIC**$_{\text{Stage I}}$ | 35.63 | 28.63 | 31.38 |
| **ATOMIC**$_{\text{Stage II}}$ | 43.00 | 31.63 | 38.75 |
| **ATOMIC**$_{\text{Stage I + II}}$ | 52.50 | 35.63 | 41.25 |

### 5.4 Ablation Study

In this section, we conduct a detailed analysis to evaluate the effectiveness of the design choices in each stage of the proposed ATOMIC method. As shown in Table 2, we evaluate different stage configurations across multiple model families on the GSM-DC benchmark. We observe that each individual stage yields a significant improvement over the Vanilla method, demonstrating the effectiveness of the design at each stage. Furthermore, when Stage I and Stage II are combined, performance is further improved, indicat-

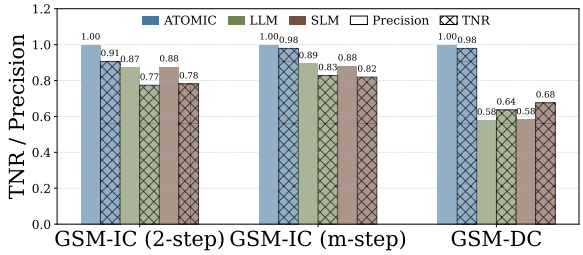


Figure 5: TNR and Precision of ATOMIC, the SLM, and the LLM within the Qwen3 family on GSM-IC and GSM-DC.

ing that the stages are not only individually effective but also complementary in nature. These results confirm the effectiveness of the overall ATOMIC design. We further provide a quantitative analysis following the evaluation protocol described in Section 3.1. As illustrated in Fig. 5, we assess performance from the perspectives of both true negative rate (TNR) and Precision. The results show that ATOMIC consistently outperforms all other methods across both metrics, and in some cases even surpasses the corresponding LLMs. A high TNR indicates effective suppression of irrelevant context, while high Precision reflects an improved ability to distinguish tokens that are critical for the reasoning process. Taken together, both empirical performance gains and quantitative metric analyses consistently validate the effectiveness of the proposed ATOMIC method.

## 6 Related Works

**Irrelevant Context.** A growing body of prior work studies the role of irrelevant context in language model reasoning. KAFT (Li et al., 2023b) introduces this research direction by analyzing how models interact with contextual information. The authors argue that models should use context to update or correct predictions when the context contains task relevant information. When the context is irrelevant, models should ignore it and rely on internal knowledge. Irrelevant context is later formulated more concretely through benchmark construction. GSM-IC (Shi et al., 2023) and GSM-DC (Yang et al., 2025b) introduce standardized benchmarks for studying irrelevant context in mathematical reasoning tasks. Both benchmarks augment arithmetic problems with additional irrelevant information. GSM-IC injects irrelevant context according to different reasoning step settings. GSM-DC constructs dependency graphs and introduces irrelevant context with varying difficulty levels. Subsequent studies further investigate why irrelevant context degrades model performance. One line of work (Wu et al., 2024) shows that large language models tend to rely on semantic similarity rather than factual or causal relevance when assessing information relevance. As a result, semantically related but incorrect information can mislead model reasoning. ATOMIC analyzes irrelevant context from an attention based perspective and introduces a principled mechanism for improving attention allocation under irrelevant context.

**Small Language Models.** Small language models (SLMs) have recently attracted significant attention as accessible and cost-effective alternatives for efficient language model deployment (Wang et al., 2025a; Lu et al., 2024). Recent work has explored both pretraining and knowledge distillation to improve their capabilities. TinyLlama (Zhang et al., 2024) investigates extensive pretraining for a compact 1.1B-parameter model, while SmolLM2 (Ben Allal et al., 2025) emphasizes data-centric, multi-stage training with carefully designed data mixtures. In knowledge distillation, MiniLLM (Gu et al., 2024) uses a reverse Kullback–Leibler divergence objective to transfer generative capabilities from larger teachers to smaller students. For reasoning-oriented training, chain-of-thought (CoT) knowledge distillation (Magister et al., 2023) fine-tunes smaller models on teacher-generated reasoning traces, while Distilling Step-by-Step (Hsieh et al., 2023) uses teacher-generated rationales as auxiliary supervision within a multi-task learning framework. Beyond model development, recent work argues that SLMs can efficiently handle many specialized tasks in agentic systems (Belcak et al., 2025). SC-Retriever (Jung & Joe, 2025) combines SLMs with retrieval-augmented generation for information retrieval tasks, illustrating their potential in practical applications. In contrast to these works, ATOMIC focuses on improving the robustness of SLMs under irrelevant context. By enhancing their ability to handle irrelevant context, ATOMIC supports the broader applicability of SLMs in real-world reasoning scenarios.

## 7 Conclusion

In this paper, we propose ATOMIC, a method designed to mitigate the performance degradation of SLMs caused by irrelevant context during reasoning. We first identify a systematic misalignment in attention score allocation when irrelevant context is present, which negatively affects SLM reasoning performance. To address this issue, ATOMIC introduces targeted suppression of irrelevant context and explicit extraction of *evidence tokens* that are critical for reasoning. Extensive experimental results across multiple benchmarks and model families demonstrate the effectiveness of our method.

## AI USE STATEMENT

Generative AI tools were used solely for language polishing, including grammar correction and improvements to clarity and presentation. They were not used to formulate research ideas, design or conduct experiments, analyze data, generate results, or draw conclusions. All AI-assisted edits were reviewed and verified by the authors, who take full responsibility for the content of this paper.

## ETHICS STATEMENT

This work evaluates inference-time robustness on established mathematical reasoning benchmarks and does not involve human participants or the collection of personal or sensitive data. ATOMIC uses existing model inputs and attention signals and does not introduce new training data. By reducing sensitivity to irrelevant context, the method may improve the reliability and accessibility of compact language models in resource-constrained settings. Relevance filtering and attention reallocation can nevertheless suppress information that is important but difficult to identify. The method should therefore not be interpreted as a guarantee of correctness, and applications in high-stakes settings require task-specific evaluation and appropriate human oversight.

## REFERENCES

Aida Amini, Saadia Gabriel, Peter Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. *arXiv preprint arXiv:1905.13319*, 2019.

Peter Belcak, Greg Heinrich, Shizhe Diao, Yonggan Fu, Xin Dong, Saurav Muralidharan, Yingyan Celine Lin, and Pavlo Molchanov. Small language models are the future of agentic ai. *arXiv preprint arXiv:2506.02153*, 2025.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, Colin Raffel, Leandro von Werra, and Thomas Wolf. SmolLM2: When smol goes big – data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*, 2025. doi: 10.48550/arXiv.2502.02737. URL https://arxiv.org/abs/2502.02737.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3), March 2024. ISSN 2157-6904.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240): 1–113, 2023.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=5h0qf7IBZZ.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

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. *arXiv preprint arXiv:2103.03874*, 2021.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8003–8017, 2023.

Taemoon Jung and Inwhee Joe. An intelligent docent system with a small large language model (sllm) based on retrieval-augmented generation (rag). *Applied Sciences*, 15(17):9398, 2025.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

Chengshu Li, Jacky Liang, Andy Zeng, Xinyun Chen, Karol Hausman, Dorsa Sadigh, Sergey Levine, Li Fei-Fei, Fei Xia, and Brian Ichter. Chain of code: Reasoning with a language model-augmented code emulator. *arXiv preprint arXiv:2312.04474*, 2023a.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. Large language models with controllable working memory. In *Findings of the association for computational linguistics: ACL 2023*, pp. 1774–1793, 2023b.

Zhenyan Lu, Xiang Li, Dongqi Cai, Rongjie Yi, Fangming Liu, Xiwen Zhang, Nicholas D Lane, and Mengwei Xu. Small language models: Survey, measurements, and insights. *arXiv preprint arXiv:2409.15790*, 2024.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. Teaching small language models to reason. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1773–1781. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-short.151. URL https://aclanthology.org/2023.acl-short.151/.

Iman Mirzadeh, Keivan Alizadeh-Vahid, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=AjXkRZIvjB.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *ACM Trans. Intell. Syst. Technol.*, 16(5), August 2025. ISSN 2157-6904.

Chandan K Reddy and Parshin Shojaee. Towards scientific discovery with generative ai: Progress, opportunities, and challenges. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 28601–28609, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pp. 31210–31227. PMLR, 2023.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475. URL https://aclanthology.org/2022.tacl-1.31/.

Fali Wang, Minhua Lin, Yao Ma, Hui Liu, Qi He, Xianfeng Tang, Jiliang Tang, Jian Pei, and Suhang Wang. A survey on small language models in the era of large language models: Architecture, capabilities, and trustworthiness. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 6173–6183, 2025a.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Yanbo Wang, Zixiang Xu, Yue Huang, Chujie Gao, Siyuan Wu, Jiayi Ye, Xiuying Chen, Pin-Yu Chen, and Xiangliang Zhang. Breaking focus: Contextual distraction curse in large language models. In *Will Synthetic Data Finally Solve the Data Access Problem?*, 2025b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022. URL https://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Siye Wu, Jian Xie, Jiangjie Chen, Tinghui Zhu, Kai Zhang, and Yanghua Xiao. How easily do irrelevant inputs skew the responses of large language models? *arXiv preprint arXiv:2404.03302*, 2024.

Tianqiang Yan, Ziqiao Lin, Lin Zhang, Zhenglong Sun, and Yuan Gao. Entrospect: Information-theoretic self-reflection elicits better response refinement of small language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 24563–24577, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Minglai Yang, Ethan Huang, Liang Zhang, Mihai Surdeanu, William Yang Wang, and Liangming Pan. How is llm reasoning distracted by irrelevant context? an analysis using a controlled benchmark. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 13340–13358, 2025b.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*, 2024.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Zican Dong, Yupeng Hou, Beichen Zhang, Yingqian Min, Junjie Zhang, Peiyu Liu, et al. A survey of large language models. *Frontiers of Computer Science*, 20(12):2012627, 2026.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*, 2022.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V. Le, and Ed H. Chi. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WZH7099tgfM.

Youxiang Zhu, Ruochen Li, Danqing Wang, Daniel Haehn, and Xiaohui Liang. Focus directions make your language models pay more attention to relevant contexts. *arXiv preprint arXiv:2503.23306*, 2025.

# A Supplementary Technical Details

## A.1 Limited Reasoning Capacity Assumption

In this appendix, we formalize the reasoning capacity constraint that underlies the theoretical analysis of *Evidence Likelihood Score* (ELS) in the main text. Our objective is to capture the practical limitation that a language model cannot effectively aggregate an unbounded number of tokens during multi step reasoning. Instead, due to architectural, computational, and attention sparsity constraints, only a limited subset of tokens can meaningfully contribute to the final reasoning outcome.

**Assumption A.1** (Limited Reasoning Capacity)**.** Due to architectural, computational, and attention sparsity constraints, a language model can effectively utilize at most $k$ tokens when performing multi step reasoning for a given question. Formally, the model selects a subset of tokens whose cardinality does not exceed $k$, and only these tokens meaningfully contribute to the final reasoning outcome.

This assumption reflects the empirical observation that attention distributions in language models are highly sparse, and that only a small subset of tokens can be effectively aggregated as reasoning evidence. Under this constraint, evidence selection is naturally subject to a fixed budget. Let $\mathcal{Q}$ denote the set of input tokens, and let $\delta : \mathcal{Q} \to \{0, 1\}$ be a binary evidence selection function, where $\delta(q) = 1$ indicates that token $q$ is selected as reasoning evidence. Then Assumption A.1 induces the budget constraint $\sum_{q \in \mathcal{Q}} \delta(q) = k$.

This budgeted selection view provides the foundation for the theoretical analysis in the main text. In particular, once the number of usable evidence tokens is limited to $k$, maximizing Precision reduces to identifying the subset of $k$ tokens that is most likely to contain true reasoning evidence. The main text then shows that this objective is solved by ranking tokens according to their posterior evidence probabilities, and that ELS is theoretically justified insofar as it preserves this ranking up to a bounded approximation error.

## A.2 Derivation and Approximation of the Evidence Likelihood Score

In this section, we clarify the intermediate steps connecting the posterior-based definition of the evidence likelihood score and the linear surrogate equation in Section 3.3, and discuss the role of the additive constant in this approximation.

Starting from Eq. 2, the evidence likelihood score is defined as the posterior log-odds of a token being a true evidence token:

$$\mathrm{ELS}(q) = \log \frac{p\left(q \in \mathcal{E}^\star \mid \tilde{Q}\right)}{p\left(q \notin \mathcal{E}^\star \mid \tilde{Q}\right)}. \tag{4}$$

Applying Bayes' rule, this expression can be rewritten as

$$\begin{aligned} \mathrm{ELS}(x) = \log \frac{p\left(q \mid y = 1, \tilde{Q}\right)}{p\left(q \mid y = 0, \tilde{Q}\right)} \\ + \log \frac{p(y = 1 \mid \tilde{Q})}{p(y = 0 \mid \tilde{Q})}. \end{aligned} \tag{5}$$

where $y = 1$ denotes evidence tokens and $y = 0$ denotes non-evidence tokens. The second term in Eq. 5 depends only on the prior odds of evidence tokens under the context $\tilde{Q}$ and is independent of the specific token $x$. Therefore, it constitutes an additive constant shared across all tokens in the same input.

Since evidence selection under a fixed budget depends only on the relative ordering of tokens, this constant term does not affect the induced ranking and can be safely ignored. Consequently, the evidence likelihood score can be expressed, up to an additive constant, as a log-likelihood ratio:

$$\begin{aligned} \mathrm{ELS}(q) \propto \log p\Big(q \mid y = 1, \tilde{Q}\Big) \\ - \log p\Big(q \mid y = 0, \tilde{Q}\Big). \end{aligned} \tag{6}$$

In practice, directly estimating the token-level likelihoods in Eq. 6 is intractable at inference time. Rather than approximating both terms explicitly, we construct a ranking-consistent surrogate by modeling the log-likelihood of evidence tokens using a linear composition of token-level features:

$$\log p\Big(q \mid y = 1, \tilde{Q}\Big) \;\approx\; \sum_k w_k \, \phi_k(q, \tilde{Q}), \tag{7}$$

while the contribution of $\log p(q \mid y = 0, \tilde{Q})$ is absorbed into the additive constant and omitted. This yields the linear scoring function in Section 3.3.

Importantly, the objective of this approximation is not to recover calibrated posterior probabilities, but to preserve the relative ordering of tokens by their likelihood of being true evidence. Under a fixed evidence selection budget, any strictly monotonic transformation of the true posterior induces the same optimal selection. Therefore, omitting the constant term and approximating only the evidence-conditioned likelihood is sufficient for the purposes of ranking-based evidence selection.

### A.3 Evidence Likelihood Scoring Approximates Posterior Evidence Ranking

This subsection provides a principled justification for Stage II by relating our evidence likelihood score (ELS) to the latent posterior evidence probability

$$p(q) \;\triangleq\; \mathbb{P}\big(y(q) = 1 \mid \tilde{Q}\big), \tag{8}$$

where $y(q) \in \{0, 1\}$ indicates whether token $q$ is a true evidence token under refined context $\tilde{Q}$. Since $y(q)$ is unobserved in our setting, we do not claim to estimate $p(q)$ exactly. Instead, we show that under a standard score modeling assumption, a linear scoring function of token features can approximate the *posterior log odds* and therefore preserve the posterior induced ranking, which is sufficient for precision optimal budgeted selection (Theorem 4.2).

**Feature representation.** Let $\phi(q, \tilde{Q}) \in \mathbb{R}^d$ denote a deterministic feature vector computed from token $q$ and the distilled context $\tilde{Q}$ produced by Stage I. In our instantiation, $\phi$ concatenates four complementary feature blocks, including query token semantic relevance, contrastive contextual salience, positional prominence, and token type indicators. Stage II computes a linear score

$$\mathrm{ELS}(q) \;\triangleq\; w^\top \phi(q, \tilde{Q}), \tag{9}$$

with fixed weights $w$.

**A standard posterior log odds model.** We adopt a common discriminative score modeling assumption, which posits that the posterior log odds of a binary latent label can be expressed, or well approximated, as an affine function of features.

**Assumption A.2** (Linear Log Odds)**.** There exist $w^\star \in \mathbb{R}^d$ and $b^\star \in \mathbb{R}$ such that the posterior log odds of $q$ being true reasoning evidence are linear:

$$\log \frac{p(q)}{1 - p(q)} = (w^\star)^\top \phi(q, \tilde{Q}) + b^\star, \text{ for all } q \in \tilde{Q}. \tag{10}$$

Assumption A.2 is widely used in NLP and information retrieval whenever one needs a tractable scoring function for ranking or selection. It is the same modeling principle underlying linear classifiers such as logistic regression, linear discriminative rerankers, and feature based scoring functions for retrieval and rationale selection. Importantly, the assumption concerns the existence of a feature space in which log odds is approximately linear, and it does not require calibrated probabilities.

**From log odds approximation to posterior ranking consistency.** Assumption A.2 immediately implies that any score that is an affine transform of $(w^\star)^\top \phi(q, \tilde{Q})$ preserves the ordering induced by $p(q)$, because both the logit map $p \mapsto \log \frac{p}{1-p}$ and affine transforms with positive slope are strictly monotone. Formally:

**Proposition A.3** (Ranking Equivalence)**.** *Under Assumption A.2, for any two tokens $q_i, q_j \in \tilde{Q}$,*

$$\begin{aligned} p(q_i) > p(q_j) \;\iff\; & \\ (w^\star)^\top \phi(q_i, \tilde{Q}) > (w^\star)^\top \phi(q_j, \tilde{Q}). & \end{aligned} \tag{11}$$

*Consequently, selecting the top $k$ tokens by the linear score $(w^\star)^\top \phi(q, \tilde{Q})$ is identical to selecting the top $k$ tokens by posterior evidence probability $p(q)$.*

*Proof.* By Assumption A.2, we have

$$\log \frac{p(q_i)}{1-p(q_i)} - \log \frac{p(q_j)}{1-p(q_j)} = (w^\star)^\top \big(\phi(q_i, \tilde{Q}) - \phi(q_j, \tilde{Q})\big).$$

The logit function is strictly increasing on $(0, 1)$, hence $p(q_i) > p(q_j)$ if and only if $\log \frac{p(q_i)}{1-p(q_i)} > \log \frac{p(q_j)}{1-p(q_j)}$, which holds if and only if $(w^\star)^\top \phi(q_i, \tilde{Q}) > (w^\star)^\top \phi(q_j, \tilde{Q})$. □

**Connection to ELS.** Our ELS in Eq. 9 is a tractable surrogate for the unknown optimal score $(w^\star)^\top \phi(q, \tilde{Q})$. Even when Eq. 10 holds only approximately, the key implication remains: *precise probability calibration is unnecessary*, because Theorem 4.2 requires only the *relative ordering* of tokens by $p(q)$. Therefore, the role of feature design in Stage II is to construct $\phi(q, \tilde{Q})$ so that a simple linear score yields an ordering that is well aligned with the latent posterior ranking.

**Implication for budgeted evidence selection.** Combining Proposition A.3 with Theorem 4.2 yields the following corollary.

**Corollary A.4** (Precision Optimal Scoring)**.** *Under the limited reasoning capacity in Assumption A.1, the top $k$ tokens ranked by $(w^\star)^\top \phi(q, \tilde{Q})$ form a precision optimal evidence set in expectation. Moreover, any scoring function that is a strictly increasing transform of $(w^\star)^\top \phi(q, \tilde{Q})$ induces the same optimal top $k$ selection.*

**Practical interpretation.** This analysis clarifies what our theory does and does not claim. We do not claim access to the latent posterior $p(q)$, nor do we claim perfect recovery of all evidence tokens. Rather, we justify Stage II as a principled approximation to posterior ranking under constrained reasoning capacity: if the feature induced linear score is aligned with posterior log odds in the standard sense of discriminative score modeling, then ranking tokens by ELS approximates ranking by $p(q)$, which is precisely the criterion required for precision optimal budgeted selection.

# B Supplementary Experiments

## B.1 Evidence Token Extraction

In this section, we introduce the procedure for extracting *evidence tokens* in ATOMIC. Specifically, we compute an evidence likelihood score (ELS) for tokens that are relevant to the reasoning process, as defined in Section 3.3.

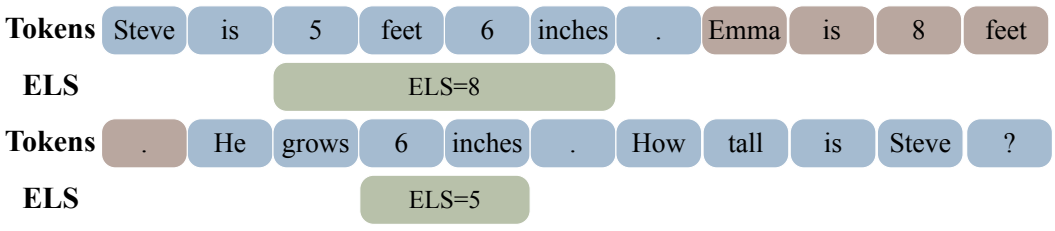


Figure 6: Illustration of the evidence likelihood score (ELS) produced by ATOMIC. Blue blocks denote relevant context (RC) tokens, red blocks denote irrelevant context (IC) tokens, and green blocks represent the ELS corresponding to the tokens above.

As shown in Fig. 6, this figure provides an illustrative example of ELS computed by ATOMIC during a single inference instance. We observe that ATOMIC successfully identifies tokens that are most critical to the final reasoning outcome, and these tokens directly participate in the final computation of the answer. This result demonstrates the effectiveness of Stage II in extracting evidence tokens via the proposed ELS mechanism.

### B.2 Generalization to Additional Distractor Settings

We evaluate Qwen3-8B on two additional benchmarks with distinct types of irrelevant context, using 200 examples per benchmark. For GSM-Symbolic (Mirzadeh et al., 2025), we consider a distractor setting in which seemingly relevant but inconsequential clauses are embedded within the problem statement. For MuSiQue (Trivedi et al., 2022), we augment the context with distractor paragraphs, introducing retrieval-style noise into a non-arithmetic multi-hop question-answering task. We compare ATOMIC with chain-of-thought (CoT) prompting (Wei et al., 2022) and the Program baseline, reporting accuracy for GSM-Symbolic and answer F1 for MuSiQue.

Table 3: Performance on additional distractor settings with Qwen3-8B ($n = 200$ per benchmark). GSM-Symbolic denotes the distractor setting described in the text. Accuracy and F1 are reported on a 0–100 scale; bold indicates the best result in each column.

| Method | GSM-Symbolic Accuracy | MuSiQue + distractors F1 |
|---|---|---|
| CoT | 42.0 | 39.5 |
| Program | 61.5 | 33.0 |
| ATOMIC | **74.0** | **48.5** |

As shown in Table 3, ATOMIC achieves 74.0% accuracy on GSM-Symbolic, exceeding the strongest baseline by 12.5 percentage points, and 48.5 F1 on MuSiQue with distractors, improving over the strongest baseline by 9.0 points. These results extend the observed benefits beyond arithmetic tasks to multi-hop question answering with paragraph-level distractions.

### B.3 Inference Latency and Token Cost

We measure per-instance latency and token usage in local inference runs, averaging over 200 examples each from GSM-IC and GSM-DC. Table 4 reports total latency, the latency of each ATOMIC stage, and input/output token counts aggregated across the calls made by each method. In addition to single-pass CoT, we profile self-consistent CoT (SC-CoT) (Wang et al., 2023) with three generations ($k = 3$), as well as CoT with larger models from the same families.

Table 4: Mean inference latency and token usage over GSM-IC and GSM-DC ($n = 200$ each). Stage times and total latency are in seconds. Token counts are per-instance totals across all calls; "–" indicates that the ATOMIC stage decomposition does not apply.

| Model | Method | Stages I/II/III (s) | Total (s) | Input tokens | Output tokens |
|---|---|---|---|---|---|
| Qwen3-8B | CoT | – | 7.55 | 376.5 | 438.5 |
| Qwen3-8B | SC-CoT ($k = 3$) | – | 21.80 | 1129.5 | 1315.5 |
| Qwen3-8B | **ATOMIC** | 4.13/4.50/4.12 | 12.75 | 1337.0 | 605.0 |
| Qwen3-32B | CoT | – | 6.69 | 376.5 | 662.5 |
| Llama3.1-8B | CoT | – | 27.96 | 373.0 | 4151.5 |
| Llama3.1-8B | SC-CoT ($k = 3$) | – | 80.50 | 1119.0 | 12454.5 |
| Llama3.1-8B | **ATOMIC** | 2.00/3.39/28.12 | 33.51 | 1232.0 | 4874.0 |
| Llama3.1-70B | CoT | – | 54.99 | 373.0 | 4147.5 |

Relative to single-pass CoT on the same model, ATOMIC incurs approximately $1.69\times$ latency on Qwen3-8B and $1.20\times$ latency on Llama3.1-8B. However, it is approximately $1.71\times$ and $2.40\times$ faster than three-sample SC-CoT, respectively, while generating fewer output tokens in both cases. Thus, the additional stages introduce overhead relative to one-pass inference but are less costly than repeated CoT generation in these measurements.

The comparison with larger models reveals a family-dependent latency–storage trade-off. On Llama3.1, ATOMIC with the 8B model takes 33.51 seconds versus 54.99 seconds for 70B CoT, corresponding to a $1.64\times$ speedup, while using 16.1 versus 141.2 GB of BF16 parameter storage, approximately an $8.8\times$ reduction. On Qwen3, ATOMIC with the 8B model takes $1.91\times$ as long as 32B CoT (12.75 versus 6.69 seconds), but requires one quarter of the BF16 parameter storage (16.4

versus 65.6 GB). These storage figures cover model weights only, excluding KV caches and runtime buffers; the latency comparisons are specific to the measured local inference configuration.

### B.4 Comparison with Three-Sample Self-Consistency

To assess whether ATOMIC's improvements can be explained by additional inference calls alone, we compare it with SC-CoT using three generations and majority voting ($k = 3$). This matches the three calls used by ATOMIC, but does not impose identical token or latency budgets. We use the same fixed subset of 200 examples per benchmark for all methods, and additionally report CoT, least-to-most (LTM) prompting (Zhou et al., 2023), and Program as reference baselines.

Table 5: Accuracy (%) on fixed subsets of GSM-IC and GSM-DC ($n = 200$ per benchmark). SC-CoT uses three generations with majority voting, matching ATOMIC's number of inference calls. Other baselines are included for reference. Bold indicates the best result in each row.

| Dataset | Model | CoT | SC-CoT | LTM | Program | ATOMIC |
|---|---|---|---|---|---|---|
| GSM-IC (two-step) | Qwen3-8B | 45.5 | 51.0 | 55.5 | 93.5 | **96.5** |
| | Llama3.1-8B | 23.0 | 28.5 | 26.5 | 66.0 | **77.5** |
| | DeepSeek-R1-7B | 61.0 | 66.5 | 67.0 | 74.0 | **82.0** |
| GSM-DC | Qwen3-8B | 20.0 | 24.5 | 21.5 | 15.0 | **52.0** |
| | Llama3.1-8B | 13.0 | 15.5 | 11.0 | 12.5 | **35.0** |
| | DeepSeek-R1-7B | 21.0 | 24.0 | 20.0 | 15.0 | **41.0** |

Table 5 shows that ATOMIC outperforms SC-CoT in all six settings, with gains ranging from 15.5 to 49.0 percentage points. For the two profiled 8B models, it also achieves lower measured latency and decodes fewer tokens than SC-CoT (Table 4). For example, Qwen3-8B with ATOMIC uses 1337.0 input and 605.0 output tokens, compared with 1129.5 input and 1315.5 output tokens for SC-CoT. These results support the benefit of structured context processing over simply repeating CoT generation three times. Complementarily, the main-paper ablation using only the Stage II intervention, with a single additional pass, achieves 43.00% accuracy on GSM-DC with Qwen, compared with 14.95% for Program; these ablation results are separate from the supplementary subset reported here.

### B.5 Evaluation on 1B and 3B Models

We evaluate ATOMIC on Llama3.2-1B and Llama3.2-3B to examine its behavior below the 7–8B scale. Each model is evaluated on 200 examples from GSM-IC (two-step) and 200 examples from GSM-DC, using CoT, LTM, and Program as baselines.

Table 6: Accuracy (%) with 1B and 3B models ($n = 200$ per benchmark). Bold indicates the best result in each row, including ties.

| Model | Dataset | CoT | LTM | Program | ATOMIC |
|---|---|---|---|---|---|
| Llama3.2-3B | GSM-IC (two-step) | 15.5 | 17.0 | 24.5 | **31.0** |
| | GSM-DC | 8.5 | 7.5 | 11.0 | **13.5** |
| Llama3.2-1B | GSM-IC (two-step) | 1.5 | 0.5 | 2.0 | **2.5** |
| | GSM-DC | 0.5 | 0.0 | **1.5** | **1.5** |

As shown in Table 6, ATOMIC matches or exceeds all baselines in every evaluated setting. With Llama3.2-3B, it improves over the strongest baseline by 6.5 percentage points on GSM-IC and 2.5 points on GSM-DC. However, the 1B model remains near the performance floor across all methods, with ATOMIC reaching only 2.5% and 1.5% accuracy, respectively. We also observe failures to follow the required JSON output format at these smaller scales, highlighting structured instruction following as a practical limitation. Thus, relative gains persist at 3B parameters, whereas absolute performance at 1B remains limited.

### B.6 Explicit Attention-Bias Instantiation

The default ATOMIC pipeline operates through prompting and does not directly modify internal attention matrices. Stage I removes distractors, Stage II identifies salient evidence, and Stage III conditions reasoning on that evidence. To evaluate a direct instantiation of the formal attention-reallocation rule, we inject Stage II evidence-likelihood scores as a pre-softmax bias into the attention logits associated with the corresponding evidence token spans using forward hooks. Stage III then generates the answer with the biased attention. We compare this training-free attention-bias variant with the prompt-based pipeline using Qwen3-8B and 200 examples per benchmark.

Table 7: Accuracy (%) of prompt-based and explicit attention-bias ATOMIC with Qwen3-8B ($n = 200$ per benchmark). Bold indicates the higher observed accuracy.

| Method | GSM-IC (two-step) | GSM-DC |
|---|---|---|
| ATOMIC (prompt) | 96.5 | 52.0 |
| ATOMIC (attention-bias) | **97.0** | **54.5** |

Table 7 shows comparable performance between the two implementations, with the explicit variant improving accuracy by 0.5 percentage points on GSM-IC and 2.5 points on GSM-DC. These results are consistent with the intended evidence-reweighting principle, but do not establish equivalence between prompting and direct attention modification. Accordingly, the theoretical analysis concerns the formal reallocation rule under its stated assumptions, while the default prompting pipeline is evaluated as a functional approximation rather than an exact implementation.

### B.7 Statistical Reliability of Context Separation

We assess the statistical reliability of the SLM–LLM difference in separating relevant context (RC) from irrelevant context (IC), as measured by AUROC. Using the greedy-decoded evaluation outputs, we quantify uncertainty across evaluation instances rather than across stochastic decoding runs. We perform a paired bootstrap over 800 instances with 10,000 resamples, using the same resampled instance indices for the SLM and LLM in each family. For each resample, we recompute the instance-aggregated AUROC and the corresponding LLM-minus-SLM difference. Table 8 reports the mean AUROC, 95% confidence intervals, and the paired-bootstrap $p$-values.

Table 8: SLM–LLM AUROC differences evaluated with a paired bootstrap over 800 instances and 10,000 resamples. Parentheses contain 95% confidence intervals; the gap is LLM minus SLM.

| Model family | SLM AUROC (95% CI) | LLM AUROC (95% CI) | Gap | $p$-value |
|---|---|---|---|---|
| Llama3.1 | 0.6334 (0.5829, 0.6839) | 0.6600 (0.6140, 0.7061) | 0.0267 | 0.0042 |
| Qwen3 | 0.5733 (0.5604, 0.5863) | 0.5901 (0.5758, 0.6045) | 0.0168 | 0.0002 |

The LLM achieves higher AUROC in both families, with reported gaps of 0.0267 for Llama3.1 ($p = 0.0042$) and 0.0168 for Qwen3 ($p = 0.0002$). Both differences are statistically significant at the 0.05 level under this paired instance-resampling analysis. Although the effect sizes are modest and the absolute AUROC values indicate limited separation, the results support a statistically detectable SLM–LLM difference within the evaluated settings.

## C Reproducibility

### C.1 Experimental Settings

In this section, we describe the experimental settings used in our evaluation. The implementations of all baseline methods strictly follow the configurations and protocols established in prior work (Shi et al., 2023). For each benchmark, we sample 800 instances, ensuring that every base question (i.e., the original question before the insertion of irrelevant context) is included in the evaluation. All

backbone models are loaded using the Hugging Face framework, and during inference, we extract attention scores with relative positional information for further analysis.

### C.2 Prompt Templates

In this section, we present all ATOMIC prompt templates used in our experiments, as detailed below.

**ATOMIC Prompt Template: System Input**

You are an inference-time reasoning assistant.

Use Chain-of-Thought (CoT) reasoning.

Instruction:

Feel free to ignore irrelevant information given in the questions.

**ATOMIC Prompt Template: Stage I**

You are given a Question and a Context.

RULES (must follow):
1) Remove all content in the Context that is irrelevant to the Question.
2) In the DISTILLED_CONTEXT, keep only the content that is relevant to the Question.
3) Always retain any condition that is required to fully specify the problem posed by the Question, even if it contains no numeric values.

Question:

{QUESTION}

Context:

{CONTEXT}

Output format (exactly):
DISTILLED_CONTEXT:

**ATOMIC Prompt Template: Stage II**

You are given a Question and a Distilled Context.

RULES (must follow):
1. Extract 1-5 short evidence spans from the distilled context, and explicitly parse any numeric expressions within each span into complete canonical forms.
2. Each span should capture a useful fact or constraint. ($<$ 2 tokens)
3. sim(e): semantic relevance; dctx(e): contextual salience; pos(e): positional prominence; type(e): token-type.
4. ELS(e) = sim(e) + dctx(e) + pos(e) + type(e). (All feature scores are in 0-2)

Output JSON only:

```
{
  "evidence_tokens": [
    {
      "span": "...",
      "sim": 0|1|2,
      "dctx": 0|1|2,
      "pos": 0|1|2,
      "type": 0|1|2,
      "els": int
    }
  ]
}
```

Question:
{QUESTION}

Distilled Context:
{DISTILLED_CONTEXT}

**ATOMIC Prompt Template: Stage III**

You are given a Question, a Distilled Context, and Evidence tokens.
Answer the Question based on the Distilled Context, giving priority to the evidence token spans.

RULES (must follow):
1) Use the Distilled Context to determine the answer.
2) Pay closer attention to the evidence token spans when identifying facts relevant to the question.
3) Perform reasoning strictly over parsed numeric expressions.

Question:

{QUESTION}

Distilled Context:

{DISTILLED_CONTEXT}

Evidence tokens:

{EVIDENCE_TOKENS}

Output format (exactly):

```
{{"final_answer": "<answer>"}}
```

### C.3 HARDWARE

All experiments reported in this paper were executed on a single dedicated workstation. The machine is equipped with one AMD EPYC 7763 CPU featuring 64 physical cores with a 2.45 GHz base frequency, 1000 GB DDR4 system memory, and NVIDIA RTX PRO 6000 (Blackwell) GPU. Unless otherwise noted, every model run, ablation, and evaluation presented in this work was performed on this same hardware configuration to ensure consistency across experimental results.

## D LIMITATIONS

Our evaluation is limited to mathematical reasoning benchmarks with injected irrelevant context (GSM-IC and GSM-DC), so the results may not directly generalize to open-domain QA or broader long-context tasks. ATOMIC also depends on accessible attention signals and an extra inference-time pipeline, which may reduce its applicability to black-box APIs and increase latency. In addition, our evidence scoring relies on ranking-based approximations under a limited reasoning-capacity assumption; while effective in our setting, it does not recover the true posterior evidence distribution and may be less reliable when relevant information is implicit or highly distributed.