

SELFELICIT: Your Language Model Secretly Knows Where is the Relevant Evidence

Zhining Liu¹, Rana Ali Amjad², Ravinarayana Adkathimar², Tianxin Wei¹, Hanghang Tong¹

¹University of Illinois Urbana-Champaign, ²Amazon Science

¹{liu326, twei10, htong}@illinois.edu ²{raamjad, adkathi}@amazon.com

Abstract

Providing Language Models (LMs) with relevant evidence in the context (either via retrieval or user-provided) can significantly improve their ability to provide better-grounded responses. However, recent studies have found that LMs often struggle to fully comprehend and utilize key evidence from the context, especially when it contains noise and irrelevant information—an issue common in real-world scenarios. To address this, we propose SELFELICIT, an inference-time approach that helps LMs focus on key contextual evidence through self-guided explicit highlighting. By leveraging the inherent evidence-finding capabilities of LMs using the attention scores of deeper layers, our method automatically identifies and emphasizes key evidence within the input context, facilitating more accurate and grounded responses *without* additional training or iterative prompting. We demonstrate that SELFELICIT brings consistent and significant improvement on multiple evidence-based QA tasks for various LM families while maintaining computational efficiency. Our code and documentation are available at <https://github.com/ZhiningLiu1998/SelfElicit>.

1 Introduction

Language models have demonstrated remarkable capabilities (OpenAI, 2022, 2023; Kaddour et al., 2023), yet hallucinations and factual inaccuracies remain a significant limitation (Ji et al., 2023; Wang et al., 2023a). Furnishing the input context with relevant evidence passage is a popular approach to enhance LM’s ability to generate correct and well-grounded responses (Ram et al., 2023; Jiang et al., 2023b; Wei et al., 2025b). However, recent studies find that LMs can fail to properly leverage supporting facts within context, leading to incorrect answers despite available evidence (Shi et al., 2024; Zhao et al., 2024). Such failure largely stems from the noise and irrelevant information in context (Wu

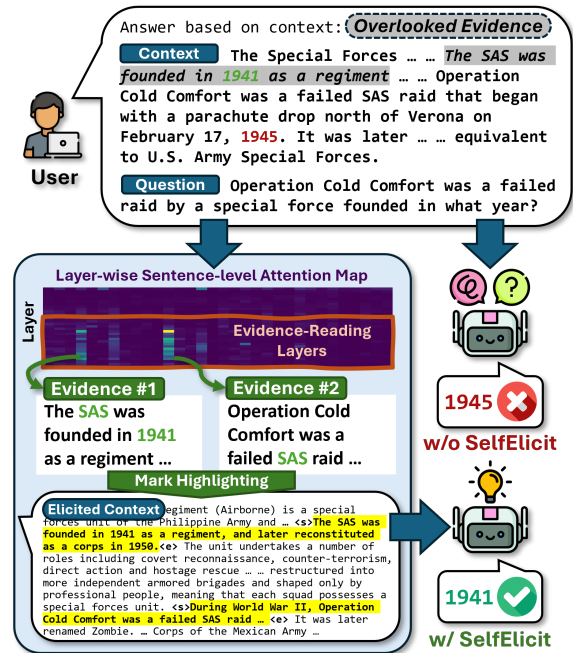


Figure 1: SELFELICIT workflow on a real example with Llama3.1-8B. By locating and explicitly highlighting the initially overlooked 2nd-hop evidence (“SAS was founded in 1941 ...”) within the context, SELFELICIT guides the model to arrive at the correct answer “1941”.

et al., 2024; Cuconasu et al., 2024), which is usually inevitable in practice (Gao et al., 2023).

Recently, improved prompting (Zhou et al., 2023) or decoding (Shi et al., 2024) methods have been proposed to enhance the focus on context. However, they treat the whole context as a single entity, overlooking the fact that not all information provided in the context is important. Meanwhile, another line of work highlights the crucial role of attention mechanisms in identifying key contextual information, leveraging attention patterns for KV cache compression (Li et al., 2024a) and document reordering (Peysakhovich and Lerer, 2023). Motivated by these insights, we delve deeper into the capabilities of attention-based mechanisms for fine-grained, inference-time evidence highlighting,

enhancing LMs’ ability to concentrate on key supporting information within the context. Our contributions in this work are as follows:

- By analyzing the attention scores during response generation, we demonstrate that the LMs have an inherent ability to identify the relevant evidence in the context, regardless of whether they respond correctly or not. This observation holds across various LM families.
- Leveraging this inherent ability, we propose an inference time context augmentation approach, SELFELICIT, that highlights the important evidence in the context and improves LM’s ability to provide correct and well-grounded responses. SELFELICIT is **efficient** for inference, **training-free**, and **robust** to noise and hyper-parameter choice.
- In a comprehensive study, we show the significant performance improvement that SELFELICIT brings across model families and tasks. We also offer extensive studies and analyses on various characteristics of SELFELICIT, including but not limited to its evidence-eliciting accuracy, effect of different design choices, robustness under noisy/informative contexts, runtime efficiency, etc.

Fig. 1 illustrates the SELFELICIT workflow. By identifying and highlighting key evidence within the context, SELFELICIT guides the LM towards important information that it might otherwise overlook, thus leading to the correct response.

2 Preliminaries

Problem description. Given an LM Φ , question \mathbf{q} , context \mathbf{c} and QA prompt template τ_{QA} , we obtain the generated answer \mathbf{g} for the question by combining context and question as input: $\mathbf{g} \leftarrow \Phi(\tau_{\text{QA}}(\mathbf{c}, \mathbf{q}))$. We use the following template as baseline to test different LMs’ base ability in leveraging contextual evidence for QA tasks:

A direct prompt template τ_{QA} for context-based QA

Directly answer the question based on the context passage, no explanation is needed. If the context does not contain any evidence, output "I cannot answer based on the given context."

Context: {context}
Question: {question}

Notation. We only outline the notation and details of Transformer models needed for this work, please

see (Vaswani et al., 2017) for more detailed exposition on transformer architecture. Given an input sequence with n tokens, a decoder only transformer LM generates next token by adaptively attending to all previous n tokens. Denote the attention probability vector for attention head h and layer ℓ by $\mathbf{a}^{(\ell, h)} \in \mathbb{R}^n$. Then we define

$$\mathbf{a}^{(\ell)} = \frac{1}{H} \sum_{h=1}^H \mathbf{a}^{(\ell, h)} \quad (1)$$

where H is the total number of heads in a layer and L is the number of layers in the transformer model. Intuitively, $\mathbf{a}^{(\ell)}$ provides a holistic view of layer ℓ ’s attention distribution over the input sequence, summarizing which tokens the model deems most informative at this stage of processing.

3 SELFELICIT

SELFELICIT is an inference-time context augmentation framework to enhance generative LM’s capability to rely on relevant information present in the context in order to reduce hallucinations and provide factually correct grounded response. SELFELICIT requires **no additional training**, is **efficient at inference time** and provides **significant performance improvement** across models and tasks.

3.1 Self-guided Contextual Evidence Eliciting

The first step in SELFELICIT is to leverage the LM’s internal representations to identify the relevant evidence sentences in the context. Consider the context-based QA setup described in Sec. 2. Denote the number of tokens in input $\tau(\mathbf{c}, \mathbf{q})$ by n and the number of sentences¹ in the context \mathbf{c} by m . Define $\mathcal{S} : \{\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_m\}$ where \mathbf{s}_i denotes i -th sentence. Let $(t_{\mathbf{s}_i}^{\text{start}}, t_{\mathbf{s}_i}^{\text{end}})$ denote the start and end token index of i -th sentence. Define sentence-level attention vector $\bar{\mathbf{a}}^{(\ell)} \in \mathbb{R}^m$ from token level attention $\mathbf{a}^{(\ell)}$ as follows

$$\bar{\mathbf{a}}^{(\ell)} := [\bar{a}_1^{(\ell)}, \bar{a}_2^{(\ell)}, \dots, \bar{a}_m^{(\ell)}] \in \mathbb{R}^m, \quad \text{where } \bar{a}_i^{(\ell)} = \frac{1}{|t_{\mathbf{s}_i}^{\text{end}} - t_{\mathbf{s}_i}^{\text{start}} + 1|} \sum_{j=t_{\mathbf{s}_i}^{\text{start}}}^{t_{\mathbf{s}_i}^{\text{end}}} \mathbf{a}_j^{(\ell)}. \quad (2)$$

Intuitively $\bar{a}_i^{(\ell)}$ tells the relative importance of each context sentence \mathbf{s}_i at layer ℓ . In Fig. 2 we plot

¹The sentences in a document can be extracted with many natural language processing packages, in this paper we use spaCy (Honnibal and Montani, 2017).

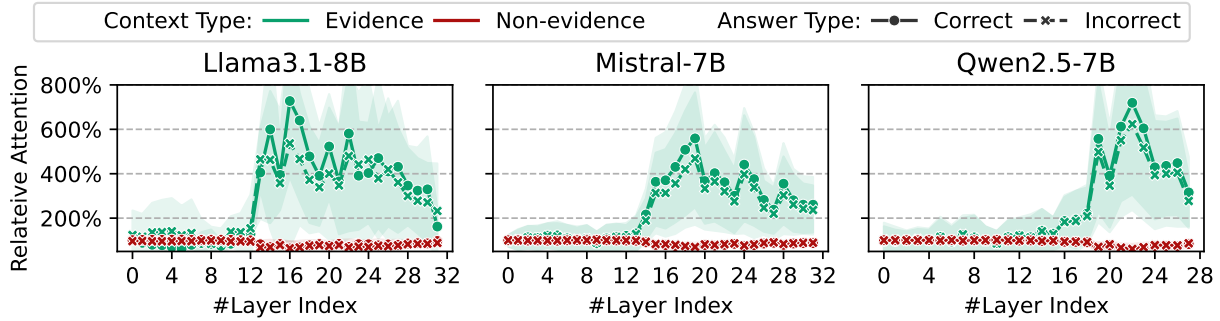


Figure 2: Relative attention³ to the evidence/non-evidence sections (y-axis) across the layers (x-axis) for different LM families on HotpotQA. Deeper layers pay much greater attention to crucial evidence (green lines) in the context, even when LM responds incorrectly (dashed lines). Best viewed in color.

how $\bar{a}_i^{(\ell)}$ varies across layers for evidence vs non-evidence sentences in the context when generating the first response token. We can clearly see that, regardless of whether the model responds correctly or not, the deeper layers of the LM pay significantly higher attention to the relevant evidence² in the context. This observation holds across models families and datasets, see Appendix B.2.

We leverage this insight to define the **sentence evidence scores** e_i for each sentence s_i in \mathcal{S} , which we use at the inference time to predict the evidence containing sentences in the context, as follows: We use a pre-selected subset of layers, denoted by \mathcal{L}_{ER} and referred to as **evidence-reading** layers, and aggregate $\bar{a}_i^{(\ell)}$ to get e_i , i.e.,

$$\mathbf{e} := [e_1, e_2, \dots, e_m] \in \mathbb{R}^m, \quad (3)$$

where
$$e_i = \frac{1}{|\mathcal{L}_{ER}|} \sum_{\ell \in \mathcal{L}_{ER}} \bar{a}_i^{(\ell)}$$

In Section 4.3 (RQ4) we study different choices of \mathcal{L}_{ER} for their ability to identify the relevant evidence. We find that using the last 50% of the layers as \mathcal{L}_{ER} performs consistently well across models and datasets so we use it as default for experiments.

Next, we introduce an evidence thresholding parameter $\alpha \in [0, 1]$ and use it in combination e_i to predict evidence sentences \mathcal{S}_{SE}

$$\mathcal{S}_{SE} = \{s_i | s_i \in \mathcal{S}; e_i \geq \alpha \cdot \max(\mathbf{e})\}, \quad (4)$$

A higher value of α will result in fewer sentences being highlighted, and vice versa. It can be seen as

²In Appendix B.2 we explain how we obtain the ground truth for relevant evidence in the context.

³Relative attention to the evidence/non-evidence section is defined as the ratio of section-average attention per token (APT) to the context-average APT, e.g., a 600% evidence relative attention means that LM is paying 6x attention to each token in evidence sentences compared to context average.

a trade-off parameter between precision and recall in evidence elicitation, which can be tuned based on task requirements. SELFELICIT is generally robust to α , see Section 4.3 (RQ5), with $\alpha = 0.5$ yielding significant performance improvement over baseline across models and tasks. Therefore, We use $\alpha = 0.5$ as our default setting for experiments.

We note that similar analyses can be presented at a token level to identify evidence tokens instead of sentences. However, as discussed in Appendix B.4, augmenting the original context by highlighting complete sentences is semantically more meaningful and beneficial than individual tokens.

3.2 Contextual Evidence Highlighting

The second step of SELFELICIT is to modify the original context to highlight the evidence sentences identified in Sec. 3.1 and to modify the original instructions in Sec. 2 to guide the model to use the highlighted evidence. Prioritizing simplicity and efficiency, we introduce a prompt augmentation strategy to achieve this. Specifically for highlighting, in the raw context passage \mathbf{c} we place text markers "`<start_important>`" and "`<end_important>`" before and after each sentence in \mathcal{S}_{SE} , resulting in the new context \mathbf{c}^* . Furthermore, we update the task instructions to guide the LM's attention towards the highlighted sentences. The response is generated using the mark-highlighted \mathbf{c}^* and the updated prompt template \mathcal{T}_{SEQA} .

SELFELICIT Prompt Template \mathcal{T}_{SEQA}

```
{Original QA Instructions} Within the context, <start_important>
and <end_important> are used to mark the important evidence sen-
tences, read carefully. Do not include the markers in the output.
Context: {elicited_context}
Question: {question}
```

The overall procedure is summarized in Algorithm 1. We note that SELFELICIT is designed to prioritize efficiency and simplicity, generating only

one additional token during inference to identify \mathcal{S}_{SE} (line 1 in Alg. 1). The highlighting method also outperforms more complex strategies, detailed further in Appendices B.5 and B.7.

Algorithm 1 SELFELICIT

Input Language model Φ , question \mathbf{q} , context \mathbf{c} , prompt templates τ_{QA} and τ_{SEQA} , evidence-reading layers \mathcal{L}_{ER} , eliciting threshold α .

- 1: Generate one token with $\Phi(\tau_{QA}(\mathbf{c}, \mathbf{q}))$;
- 2: Compute $\bar{\mathbf{a}}^{(\ell)}$ for all $\ell \in \mathcal{L}_{ER}$; \triangleright Eq. (2)
- 3: Get evidence score \mathbf{e} with \mathcal{L}_{ER} ; \triangleright Eq. (3)
- 4: Select evidence sentences \mathcal{S}_{SE} ; \triangleright Eq. (4)
- 5: Derive new context \mathbf{c}^* by highlighting \mathcal{S}_{SE} ;
- 6: Generate answer $\mathbf{g}_{SE} \leftarrow \Phi(\tau_{SEQA}(\mathbf{c}^*, \mathbf{q}))$

Output: The final answer \mathbf{g}_{SE}

4 Experiments

We conduct comprehensive experiments across six LMs from different families with varying sizes and on four single- and multi-hop reasoning open-book QA tasks from various domains to investigate:

- **RQ1:** How does SELFELICIT perform in terms of improving answer quality and accuracy?
- **RQ2:** How do the evidence scores and the context sentences highlighted by SELFELICIT correlate with relevant evidence?
- **RQ3:** Robustness to noise in the context?
- **RQ4&5:** How do different choices of evidence-reading layers & threshold α affect SELFELICIT?

4.1 Experimental Setup

Datasets and Metrics. We test 4 datasets: HotpotQA (Yang et al., 2018) and the MRQA version (Fisch et al., 2019) of NewsQA (Trischler et al., 2017), TriviaQA (TQA) (Joshi et al., 2017), and Natural Questions (NQ) (Kwiatkowski et al., 2019). These datasets feature context passages from diverse sources (e.g., web/Wikipedia/news reports), requiring the model to reason over a single or multiple pieces of evidence within the context. This provides a comprehensive test of SELFELICIT in real-world applications. For all datasets, we use the official validation split on HuggingFace for testing. We apply greedy decoding to get deterministic answers. Exact Match (EM) and Token-level F1 scores are used for QA performance evaluation.

Models and Baselines. We test six open-source instruction fine-tuned models: Llama-3.1 (8B, 70B) (Dubey et al., 2024), Mistral (7B, 12B) (Jiang

et al., 2023a), and Qwen2.5 (7B, 32B) (Yang et al., 2024). We compare SELFELICIT to the following prompting/evidence-eliciting approaches:

- **COT (Wei et al., 2022a):** Chain-of-thought prompting encourages the model to reason through intermediate steps before reaching a final answer. While CoT promotes step-by-step reasoning based on the evidence in the context, it does not explicitly highlight important information within the context. Therefore, we use it as a natural baseline for validating the advantage of explicit evidence elicitation in SELFELICIT. We implement it by adding the COT prompt "Think step by step to provide the answer." at the end of the instruction.
- **FULLELICIT:** A naive approach that highlights the entire context as important. Comparing with it shows the necessity of fine-grained, sentence-level evidence elicitation.
- **PROMPTELICIT:** This method leverages the LM itself for generative evidence extraction. It involves two steps: first, the LM is prompted to select the most relevant evidence sentences from the context passage that can help answer the question. Then, we highlight the extracted evidence and get new context for QA following Section 3.2. Note that this method involves generative evidence extraction in iterative prompting, which requires the model to generate a large amount of additional tokens. This serves as a strong baseline for evaluating whether the quality of evidence selected by SELFELICIT can match that of evidence extracted through generative approach. We use the following prompt for generative evidence extraction step:

PROMPTELICIT Prompt Template τ_{PE}

```
Please find the supporting evidence sentences from the context for the question, then copy-paste the original text to output. Template for output: '- [sentence1] - [sentence2] ...'
Context: {context}
Question: {question}
```

Since our focus is on analyzing and improving the inherent ability of the generative LM to provide factually grounded responses based on relevant information provided in the context, we do not compare to any dataset specific and task specific fine-tuning methods. For a fair comparison, we do not tune the hyper-parameters for SELFELICIT. They are fixed as \mathcal{L}_{ER} being the last 50% layers and $\alpha = 0.5$.

Table 1: Results of applying different evidence-eliciting methods to 6 LMs across 4 context-based QA tasks. We report EM and Token F1 scores (in $\times 10^{-2}$) with the gains over direct QA. The "average" columns present the average QA performance and the inference time (per sample) across all datasets. The best results are **bolded**.

Model	Method	Dataset								Average			
		HotpotQA		NewsQA		TQA		NQ		Ranking		Inference	
		EM	Token F1	EM	Token F1	EM	Token F1	EM	Token F1	EM	Token F1	Time (ms)	
Llama-3.1	8B	BASE	58.9	57.7	64.3	56.2	72.8	66.1	59.7	61.6	4.38	4.75	224.1
		COT	60.4	58.6	64.9	55.8	74.4	67.4	59.6	62.1	3.75	4.00	224.8
		FULLELICIT	60.7	59.0	65.9	56.6	72.8	66.3	61.1	62.5	3.12	3.25	226.3
		PROMPTELICIT	66.3	66.3	62.8	56.7	76.0	69.6	61.8	65.6	2.75	2.00	1672.0
		SELFELICIT	68.5	69.5	66.9	60.8	79.4	72.7	64.0	67.8	1.00	1.00	264.1
	70B	BASE	71.8	74.2	66.7	57.4	78.0	71.2	59.3	63.2	3.25	3.00	1389.8
		COT	72.4	74.0	67.2	56.4	77.0	69.9	60.3	63.1	3.12	4.50	1394.2
		FULLELICIT	71.3	73.8	66.2	56.8	77.5	70.7	58.2	61.8	4.25	4.50	1408.1
		PROMPTELICIT	73.4	76.2	63.1	58.2	77.0	72.3	64.0	68.5	3.38	2.00	8124.0
		SELFELICIT	75.9	79.0	69.2	62.1	80.0	74.4	65.4	68.7	1.00	1.00	1566.9
Mistral	7B	BASE	70.4	45.6	61.4	32.6	81.8	47.9	65.7	29.9	3.75	4.25	538.4
		COT	71.4	44.0	60.4	31.4	82.2	48.2	67.5	29.3	2.25	4.75	560.8
		FULLELICIT	70.6	47.6	62.0	34.0	81.3	51.5	66.8	30.5	3.00	3.00	541.4
		PROMPTELICIT	71.0	64.1	57.7	42.2	81.9	62.3	65.6	42.8	4.00	1.25	1877.8
		SELFELICIT	74.3	61.6	60.6	41.7	83.6	61.3	66.4	43.4	2.00	1.75	431.3
	12B	BASE	59.2	65.9	51.9	51.6	72.9	68.7	54.8	61.6	5.00	4.75	281.9
		COT	62.0	69.0	53.0	52.6	75.6	71.1	55.3	62.0	3.25	3.25	284.6
		FULLELICIT	59.8	65.8	53.1	51.9	73.7	68.8	55.5	62.6	2.88	4.00	283.9
		PROMPTELICIT	62.8	73.1	52.2	56.6	79.9	77.6	55.5	65.7	2.38	1.75	1455.0
		SELFELICIT	63.6	72.9	54.9	58.6	82.6	79.9	55.3	66.0	1.50	1.25	339.1
Qwen2.5	7B	BASE	65.2	65.8	58.3	45.4	77.4	66.1	62.2	59.9	3.75	3.75	245.2
		COT	70.7	37.9	59.2	31.9	78.6	41.6	63.8	32.3	2.00	5.00	421.5
		FULLELICIT	65.5	65.7	57.5	48.2	77.1	67.3	64.4	60.6	3.50	2.75	249.6
		PROMPTELICIT	64.7	67.7	54.9	46.3	75.4	66.8	64.5	65.0	4.25	2.25	1165.1
		SELFELICIT	69.1	71.4	59.6	50.8	78.1	67.8	65.0	64.7	1.50	1.25	289.4
	32B	BASE	71.8	68.4	60.0	44.7	79.0	69.3	62.7	59.3	3.12	3.25	928.2
		COT	71.3	67.1	60.0	43.5	79.5	66.8	59.5	55.3	3.62	5.00	998.6
		FULLELICIT	71.3	68.2	61.6	45.7	78.5	68.8	63.2	58.1	3.25	3.75	936.3
		PROMPTELICIT	71.3	74.5	59.0	51.5	78.0	69.9	64.8	68.1	4.00	2.00	5109.8
		SELFELICIT	73.3	75.0	65.6	57.3	82.1	74.8	66.8	69.8	1.00	1.00	980.7

4.2 Main Results

RQ1: SELFELICIT consistently improves grounded factuality. Table 1 compares SELFELICIT to the other methods. We observe that

- As expected FULLELICIT does not provide meaningful and consistent improvement since it doesn't elicit fine-grained evidence.
- COT also does not provide a meaningful and consistent gain which highlights the fact that asking the model to reason carefully does not improve its ability to leverage the relevant evidence while generating the response. COT results also show variation across LM families where we don't observe meaningful gains for Llama and Mistral models. When combined with Qwen models, COT generates significantly longer answers with intermediate steps. While this results in a slight increase in EM scores in a few cases, it also leads to a large drop (up to 27.9) in F1 scores due to the inclusion of redundant information, as well

as longer inference times (e.g., 171.9% more inference time for Qwen2.5-7B).

- SELFELICIT significantly and consistently improves the performance across all datasets and models of different sizes (5.0%-11.7% gain over baseline). Even when compared to computationally expensive (average inference time increase of 878%/939% for Llama3.1-8B/70B) iterative prompting approach PROMPTELICIT, SELFELICIT outperforms for 40 out of 48 model-task metric pairs while incurring a fraction of the computational cost increment (only $\sim 3\text{-}5\%$ increase when compared to PROMPTELICIT).

Appendix B.1 provides a more detailed discussion of computational efficiency of the approaches.

RQ2: SELFELICIT highlights relevant evidence. We demonstrate both qualitatively and quantitatively how the context highlighted by SELFELICIT is the relevant evidence for the task which leads to the performance gains reported in Table 1.

Table 2: Examples of how SELFELICIT helps LM, with **blue** text highlighting the evidence selected by SELFELICIT. Due to space limitation we only display a small portion of the full context, please see Appendix B.8 for full context.

	(Partial) Context Passage	Question & Answers
True or False	(Displaying 123 of 794 Words) Tiger Please is an Indie / Alternative five-piece band from Cardiff, Wales. The band formed in August 2008. The band’s influences are U2, Sigur Rós, Kings of Leon, John Mayer and Counting Crows. They signed with Walnut Tree Records in 2009 and released their debut EP "They Don’t Change Under Moonlight". "Kerrang!" magazine, "Rock Sound" magazine, and "Classic Rock" magazine praised the EP and featured the band on the "Rock Sound" and "Classic Rock" cover-mount albums. Black Rebel Motorcycle Club (often abbreviated as BRMC) is an American rock band from San Francisco, California. The group consists of Peter Hayes (vocal, guitar, harmonica), Robert Levon Been (vocal, bass, guitar), and Leah Shapiro (drums). Former drummer Nick Jago left the band in 2008	Question: Are the bands Tiger Please and Black Rebel Motorcycle Club from the same country? True Answer: No. Base: Yes. ✗ +SelfElicit: No. ✓
Comparison	(Displaying 129 of 227 Words) Home Monthly was a monthly women’s magazine published in Pittsburgh, Pennsylvania in the late 19th century. "The Strategy of the Were-Wolf Dog" is a short story by Willa Cather. It was first published in "Home Monthly" in December 1896. The Count of Crow’s Nest is a short story by Willa Cather. It was first published in "Home Monthly" in October 1896. Mirabella was a women’s magazine published from June 1989 to April 2000. It was created by and named for Grace Mirabella, a former "Vogue" editor in chief, in partnership with Rupert Murdoch. "Nanette: An Aside" is a short story by Willa Cather. It was first published in "Courier" on 31 July 1897 and one month later in "Home Monthly". "The Prodigies" is a short story by Willa Cather	Question: Which women’s magazine was published first, Mirabella or Home Monthly? True Answer: Home Monthly. Base: Mirabella. ✗ +SelfElicit: Home Monthly. ✓
Fact Retrieval	(Displaying 153 of 1014 Words) Lars Lunde (born 21 March 1964) is a Danish former professional football player, who played in the striker position. Lunde got his breakthrough with Brøndby IF in 1983, and he made his debut for the Denmark national football team in October 1983. He was sold to Young Boys Bern in Switzerland, before moving to German club Bayern Munich in 1986. He was a part of the Bayern team which won the German Bundesliga championship in 1987, and he came on as a late substitute when Bayern lost the 1987 European Cup Final to FC Porto. He played the last of his three matches for the Danish national team in April 1987, before leaving Bayern during the 1987–88 season. He went on to play for a number of smaller clubs, ending his career with FC Baden in Switzerland.	Question: Which team did Lars Lunde play for when defeated for the 1987 European Cup Final? True Answer: Bayern Munich BASE: FC Porto. ✗ +SelfElicit: Bayern Munich. ✓
Multi-hop Reasoning	(Displaying 146 of 532 Words) This sent Olympic down to play in the Premier League in 2007. Adelaide City won the title with games to spare after being runaway leaders, finishing the season unbeaten. Norwood is a suburb of Adelaide, about 4 km east of the Adelaide city centre. The suburb is in the City of Norwood Payneham & St Peters, the oldest South Australian local government municipality, with a city population over 34,000. Whyalla railway station was the terminus station of the Whyalla line serving the South Australian city of Whyalla. Walter Frank Giffen (20 September 1861 in Norwood – 28 June 1949 in Adelaide) was an Australian cricketer who played in 3 Tests between 1887 and 1892. He was the brother of the great all-rounder George Giffen. The City of Burnside is a local government area with an estimated population of 44,300 people in the South Australian city of Adelaide.	Question: Walter Giffen is from a suburb of which South Australian city? True Answer: Adelaide Base: Norwood. ✗ +SelfElicit: Adelaide. ✓

For qualitative illustration, in Table 2 we show various types of examples (e.g., true/false, comparison, fact retrieval, multi-hop reasoning) from the HotpotQA dataset. SELFELICIT identifies the most relevant supporting facts across different QA tasks. For instance, in the 2nd "comparison" example that asks which magazine was published first, SELFELICIT highlights the information stating the publication dates of the magazines and related articles. In the 4th "multi-hop reasoning" example, it emphasizes the overlooked 2nd-hop evidence "Norwood is a suburb of Adelaide", which helps the model arrive at the correct answer "Adelaide".

Table 3: SELFELICIT accurately identifies contextual evidence sentences across different datasets and models.

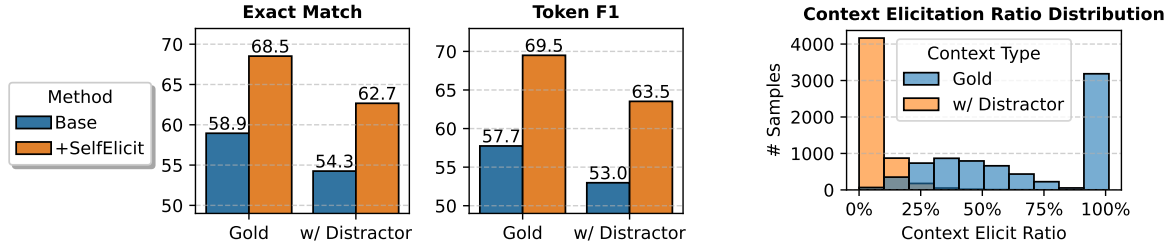
Dataset	Model		
	Llama3.1	Mistral	Qwen2.5
Metric: Sentence-level AUROC			
HotpotQA	91.24	85.35	88.21
NewsQA	92.68	88.68	91.54
TQA	73.27	68.89	70.59
NQ	90.87	85.51	87.43
Metric: Sentence-level NDCG			
HotpotQA	91.36	82.45	87.05
NewsQA	82.79	70.65	82.37
TQA	66.41	63.32	67.19
NQ	91.45	86.45	87.65

For quantitative evaluation, we assess the accuracy of evidence elicitation by checking whether

SELFELICIT assigns higher evidence scores to *ground-truth* evidence sentences. Specifically, for HotpotQA, we use the "supporting_facts" annotations to derive ground-truth evidence labels, while for other datasets, a sentence is treated as ground truth evidence if it contains at least one of the correct answers. Since SELFELICIT computes *continuous* evidence scores for each sentence (see Eq. (3)), we utilize AUROC and NDCG@all to assess the score for accurately classifying/ranking contextual evidence. Table 3 shows that SE can accurately locate evidence sentences across models and datasets, with 80-95 AUROC or NDCG scores in most cases. We note that results on the TQA dataset are relatively lower because TQA often has multiple candidate correct answers and thus a larger set of evidence. During inference, the model typically focuses on evidence for only one of the answers, which reduces the evidence score of other (candidate) evidence sentences. However, this does not hinder SELFELICIT’s ability to help LM respond better, as demonstrated by the results in Table 1.

4.3 Additional Analysis

RQ3: SELFELICIT is effective in presence of context noise. To explore the impact of real world noise on SELFELICIT we study its performance for the "distractor" variant of HotpotQA dataset containing additional distracting information, retrieved



(a) Performance of base and SELFELICIT-augmented LM with/without distracting context information, evaluated by Exact Match and Token F1 score.

(b) Context elicitation ratio of SELFELICIT with/without distracting context information.

Figure 3: SELFELICIT demonstrates robust advantage even in the presence of noisy context (Fig. 3(a)). When the context passages contain more distracting information, SELFELICIT tends to select a significantly smaller portion of text as evidence (Fig. 3(b)) to prevent the LM from being distracted by irrelevant contexts.

from Wikipedia (Yang et al., 2018), in the context. The distractor setting increases the average length of the context by 1443% by introducing irrelevant information. Figure 3(a) shows that SELFELICIT maintains the advantage over the baseline even in the presence of substantial context noise.

However, similar to the baseline, the performance of SELFELICIT for "distractor" variant worsens meaningfully when compared to SELFELICIT performance for the gold context setup, also shown in Figure 3(a). To investigate this we do a deep dive into the evidence elicitation capability of SELFELICIT in the presence of substantial noise. Figure 3(b) contrasts the elicit ratio (i.e., the proportion of elicited evidence within the original context) under gold and distractor settings. With gold context, SELFELICIT tends to select a larger proportion as evidence since most contextual information provided is relevant. By contrast, in the distractor setting where most of the context is irrelevant, the proportion of evidence selected decreases substantially (typically $<10\%$) for the same $\alpha = 0.5$. This shows that even with a static α , SELFELICIT naturally exhibits adaptiveness by selecting a lower evidence ratio in noisy contexts, focusing only on the relevant evidence without distraction from noise.

RQ4: Deeper Layers are better for \mathcal{L}_{ER} . We already saw in Fig. 2 that the deeper layers of all the model families exhibit higher attention scores and a clear ability to distinguish relevant evidence within the context. In Table 4 we verify this further empirically by comparing the evidence elicit accuracy and QA performance of seven different choices of \mathcal{L}_{ER} for Llama3.1-8B evaluated on HotpotQA with $\alpha = 0.5$. We see that, consistent with the observation in Fig. 2, choosing last 50% of the layers as \mathcal{L}_{ER} leads to (close to) best metrics

in terms of both evidence elicitation and task performance. Based on the qualitative observation in Fig. 2, this choice also looks robust across model families which is then further verified by the results in Table 1 where we see a consistent significant improvement across all model families and tasks for this choice of \mathcal{L}_{ER} . While there may be other choices of \mathcal{L}_{ER} which lead to better performance gains for a specific model-dataset pair, finding them requires model and task specific hyper-parameter tuning unlike universally applicable default setting.

Table 4: Effect of different evidence-reading (ER) layer choices on elicit accuracy and QA performance.

ER Layer Span	Elicit Accuracy		QA Performance	
	AUROC	NDCG	EM	Token F1
0%-100%	89.02	75.50	62.57	63.33
0%-50%	70.38	44.99	62.14	62.65
50%-100%	<u>91.55</u>	80.11	64.86	65.23
0%-25%	59.01	37.55	61.86	61.83
25%-50%	74.82	48.80	62.57	62.73
50%-75%	91.66	<u>79.96</u>	<u>63.57</u>	<u>64.14</u>
75%-100%	91.02	78.72	63.43	64.10

RQ5: Balancing elicitation precision and comprehensiveness in choice of α .

The choice of α in SELFELICIT acts as a proxy for trade-off between evidence elicitation precision vs coverage. For example, for $\alpha = 1$ we only select the sentence with highest e_i whereas for $\alpha = 0$ we have the whole context selected, analogue to FULLELICIT. In Figure 4 we show the impact of α on the Token F1 score and evidence elicitation ratio for Llama3.1-8B on four QA datasets. We observe

- For all datasets the performance quickly rises and the evidence elicitation ratio quickly drops with increasing α for smaller values of

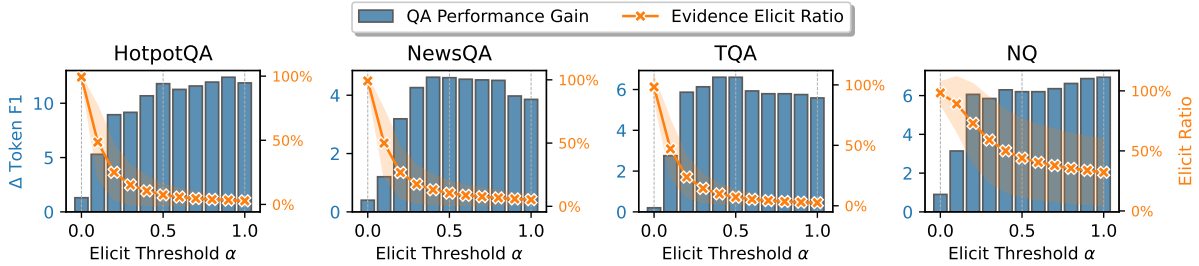


Figure 4: Impact of elicit threshold α (x-axis) on the QA performance gain (blue bars, left y-axis) and evidence elicit ratio (orange lines, right y-axis) of SELFELICIT on four QA tasks. Best viewed in color.

α indicating clearly that evidence elicitation upto a threshold helps all datasets.

- Beyond $\alpha = 0.5$ there is relatively minor variation in the performance with change in α , illustrating robustness of SELFELICIT to choice of $\alpha \in [0.5, 1]$ across datasets. The minor variation is dependent on the nature of the dataset. For datasets that require multi-hop reasoning, relying on multiple pieces of relevant evidence, we achieve optimal performance for α closer to 0.5 whereas for NQ requiring simpler reasoning we achieve the best performance for $\alpha = 1$.

To strike a balanced trade-off between evidence elicitation precision and coverage, we choose $\alpha = 0.5$ as default for experiments which leads to consistent gains across models and datasets (Table 1).

5 Related Works

Context-based question answering. Furnishing LMs with relevant context is an effective way of providing up-to-date external and/or private knowledge (Ji et al., 2023; Asai et al., 2023) to help mitigate hallucination and improve response accuracy (Huang et al., 2023). Retrieval-augmented Generation (RAG) is a widely adopted paradigm for this purpose (Gao et al., 2023; Fan et al., 2024). Despite its popularity, recent studies have pointed out that context retrieved from external sources often contains noise and irrelevant information, leading to confusion for LM (Cuconasu et al., 2024; Wu et al., 2024; Yu et al., 2024). Motivated by this, we explore how to leverage contextual information more effectively at a finer granularity by highlighting critical information within the context. To the best of our knowledge, we are the first to investigate automated contextual evidence highlighting based on LM internal representations.

Factuality and internal representation. Recent

studies have explored ways to understand QA factuality by analyzing the internal representations to identify important attention layers (Yuksekgonul et al., 2024; Chen et al., 2024b) or heads (Halawi et al., 2024) that are crucial for generation correctness or hallucination. However, they primarily focus on how LMs utilize their parametric knowledge in controlled generation (Yuksekgonul et al., 2024; Halawi et al., 2024) or closed-book QA (Chen et al., 2024b), with little discussion on context-based QA that rely on using external non-parametric knowledge. For context-based QA, Peysakhovich and Lerer (2023) propose combating positional bias by iteratively ranking retrieved context documents based on average attention scores. Wang et al. (2023b) employs a fine-tuned Llama-2 evidence extractor to filter context information. Nevertheless, these works either employ document-level average attention over all layers and heads or necessitate training an auxiliary model for generative filtering. In contrast, our work delves into a detailed analysis of the behavior of each attention layer at the sentence level, enabling fine-grained and efficient evidence highlighting at inference time.

6 Conclusion

This paper proposes a novel method SELFELICIT that boosts LM’s ability to utilize context with automatic contextual evidence highlighting. It harnesses the inherent ability of specific attention layers within the LM to differentiate evidence within the context, enabling test-time auto-eliciting that is general, efficient and requires no additional training. Comprehensive experiments on context based QA validate the effectiveness of SELFELICIT, showing a performance improvement of 5.0% to 11.7% across multiple datasets and LMs. Compared to more costly generative evidence extraction, our method achieves better performance with far less additional computational overhead.

Limitations

We evaluated the effectiveness of SELFELICIT across several open-source LMs. However, we can not assess SELFELICIT on proprietary LMs due to the need to access attention scores. Prioritizing efficiency and simplicity, we used a static eliciting threshold in the paper. We discussed the effect of different threshold α choices in Figure 4. Our method is generally robust to different threshold α , with a broad range of values yielding consistently strong improvements, but one can still fine-tune it on specific tasks for better performance. Moreover, an interesting future direction would be to automatically select the optimal elicit thresholds based on the model and input characteristics. We provide a more detailed case analysis and discussion on the potential benefit of dynamic threshold adjustment in Appendix C.1.

Acknowledgments

We thank Danai Koutra for her insights and discussions throughout the project. This work is partially supported by NSF (2134079) and AFOSR (FA9550-24-1-0002), The content of the information in this document does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

- Mengting Ai, Tianxin Wei, Yifan Chen, Zeming Guo, and Jingrui He. 2023. Mlp fusion: Towards efficient fine-tuning of dense and mixture-of-experts language models. *arXiv e-prints*, pages arXiv-2307.
- Mengting Ai, Tianxin Wei, Yifan Chen, Zhichen Zeng, Ritchie Zhao, Girish Varatkar, Bitu Darvish Rouhani, Xianfeng Tang, Hanghang Tong, and Jingrui He. 2025. Resmoe: Space-efficient compression of mixture of experts llms via residual restoration. *arXiv preprint arXiv:2503.06881*.
- Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.
- Yikun Ban, Yuchen Yan, Arindam Banerjee, and Jingrui He. 2021. Ee-net: Exploitation-exploration neural networks in contextual bandits. *arXiv preprint arXiv:2110.03177*.
- Wenxuan Bao, Tianxin Wei, Haohan Wang, and Jingrui He. 2023. Adaptive test-time personalization for federated learning. *Advances in Neural Information Processing Systems*, 36:77882–77914.
- Eunice Chan, Zhining Liu, Ruizhong Qiu, Yuheng Zhang, Ross Maciejewski, and Hanghang Tong. 2024. Group fairness via group consensus. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1788–1808.
- Lingjie Chen, Ruizhong Qiu, Siyu Yuan, Zhining Liu, Tianxin Wei, Hyunsik Yoo, Zhichen Zeng, Deqing Yang, and Hanghang Tong. 2024a. Wapiti: A watermark for finetuned open-source llms. *arXiv preprint arXiv:2410.06467*.
- Shiqi Chen, Miao Xiong, Junteng Liu, Zhengxuan Wu, Teng Xiao, Siyang Gao, and Junxian He. 2024b. In-context sharpness as alerts: An inner representation perspective for hallucination mitigation. In *Forty-first International Conference on Machine Learning*.
- Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Learning Representations (ICLR)*.
- Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonello, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for rag systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 719–729.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6491–6501.
- Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. Mrqa 2019 shared task: Evaluating generalization in reading comprehension. In *EMNLP 2019 MRQA Workshop*, page 1.
- Dongqi Fu, Liri Fang, Zihao Li, Hanghang Tong, Vette I. Torvik, and Jingrui He. 2024. Parametric graph representations in the era of foundation models: A survey and position. *CoRR*, abs/2410.12126.
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

- Kai Guo, Xiaofeng Cao, Zhining Liu, and Yi Chang. 2023. Taming over-smoothing representation on heterophilic graphs. *Information Sciences*, 647:119463.
- Danny Halawi, Jean-Stanislas Denain, and Jacob Steinhardt. 2024. Overthinking the truth: Understanding how language models process false demonstrations. In *The Twelfth International Conference on Learning Representations*.
- Xinrui He, Tianxin Wei, and Jingrui He. 2023. Robust basket recommendation via noise-tolerated graph contrastive learning. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 709–719.
- Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1):411–420.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. [A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions](#). *arXiv preprint arXiv:2311.05232*.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. [Survey of hallucination in natural language generation](#). In *ACM Computing Surveys*.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023a. Mistral 7b. *arXiv preprint arXiv:2310.06825*.
- Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023b. [Active retrieval augmented generation](#). In *Empirical Methods in Natural Language Processing (EMNLP)*.
- Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. [Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension](#). In *Association for Computational Linguistics (ACL)*.
- Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. [Challenges and applications of large language models](#). In *arXiv preprint arXiv:2307.10169*.
- 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. 2019. [Natural questions: A benchmark for question answering research](#). In *Transactions of the Association of Computational Linguistics (TACL)*.
- Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. 2024a. [Snapkv: Llm knows what you are looking for before generation](#). *arXiv preprint arXiv:2404.14469*.
- Yunzhe Li, Junting Wang, Hari Sundaram, and Zhining Liu. 2025. [A zero-shot generalization framework for llm-driven cross-domain sequential recommendation](#). *arXiv preprint arXiv:2501.19232*.
- Zihao Li, Lecheng Zheng, Bowen Jin, Dongqi Fu, Baoyu Jing, Yikun Ban, Jingrui He, and Jiawei Han. 2024b. [Can graph neural networks learn language with extremely weak text supervision?](#) *CoRR*, abs/2412.08174.
- Xiao Lin, Jian Kang, Weilin Cong, and Hanghang Tong. 2024a. [Bemap: Balanced message passing for fair graph neural network](#). In *Learning on Graphs Conference*, pages 37–1. PMLR.
- Xiao Lin, Zhining Liu, Dongqi Fu, Ruizhong Qiu, and Hanghang Tong. 2024b. [Backtime: Backdoor attacks on multivariate time series forecasting](#). *arXiv preprint arXiv:2410.02195*.
- Lihui Liu, Boxin Du, Yi Ren Fung, Heng Ji, Jiejun Xu, and Hanghang Tong. 2021a. [Kompare: A knowledge graph comparative reasoning system](#). In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '21*, page 3308–3318, New York, NY, USA. Association for Computing Machinery.
- Lihui Liu, Boxin Du, Jiejun Xu, Yinglong Xia, and Hanghang Tong. 2022. [Joint knowledge graph completion and question answering](#). In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '22*, page 1098–1108, New York, NY, USA. Association for Computing Machinery.
- Lihui Liu, Blaine Hill, Boxin Du, Fei Wang, and Hanghang Tong. 2024a. [Conversational question answering with language models generated reformulations over knowledge graph](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 839–850, Bangkok, Thailand. Association for Computational Linguistics.
- Zhining Liu, Wei Cao, Zhifeng Gao, Jiang Bian, Hechang Chen, Yi Chang, and Tie-Yan Liu. 2020a. [Self-paced ensemble for highly imbalanced massive data classification](#). In *2020 IEEE 36th international conference on data engineering (ICDE)*, pages 841–852. IEEE.
- Zhining Liu, Jian Kang, Hanghang Tong, and Yi Chang. 2021b. [Imbens: Ensemble class-imbalanced learning in python](#). *arXiv preprint arXiv:2111.12776*.
- Zhining Liu, Ruizhong Qiu, Zhichen Zeng, Hyunsik Yoo, David Zhou, Zhe Xu, Yada Zhu, Kommy Woldemariam, Jingrui He, and Hanghang Tong. 2024b.

- Class-imbalanced graph learning without class rebalancing. In *Forty-first International Conference on Machine Learning*.
- Zhining Liu, Ruizhong Qiu, Zhichen Zeng, Yada Zhu, Hendrik Hamann, and Hanghang Tong. 2024c. Aim: Attributing, interpreting, mitigating data unfairness. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2014–2025.
- Zhining Liu, Pengfei Wei, Jing Jiang, Wei Cao, Jiang Bian, and Yi Chang. 2020b. Mesa: boost ensemble imbalanced learning with meta-sampler. *Advances in neural information processing systems*, 33:14463–14474.
- OpenAI. 2022. Introducing chatgpt. [URL https://openai.com/blog/chatgpt](https://openai.com/blog/chatgpt).
- OpenAI. 2023. [GPT-4 technical report](https://arxiv.org/abs/2303.08774). *arXiv preprint arXiv:2303.08774*.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.
- Alexander Peysakhovich and Adam Lerer. 2023. Attention sorting combats recency bias in long context language models. *arXiv preprint arXiv:2310.01427*.
- Ruizhong Qiu, Jun-Gi Jang, Xiao Lin, Lihui Liu, and Hanghang Tong. 2024a. TUCKET: A tensor time series data structure for efficient and accurate factor analysis over time ranges. *Proceedings of the VLDB Endowment*, 17(13).
- Ruizhong Qiu, Zhiqing Sun, and Yiming Yang. 2022. DIMES: A differentiable meta solver for combinatorial optimization problems. In *Advances in Neural Information Processing Systems*, volume 35, pages 25531–25546.
- Ruizhong Qiu and Hanghang Tong. 2024. Gradient compressed sensing: A query-efficient gradient estimator for high-dimensional zeroth-order optimization. In *Proceedings of the 41st International Conference on Machine Learning*.
- Ruizhong Qiu, Dingsu Wang, Lei Ying, H Vincent Poor, Yifang Zhang, and Hanghang Tong. 2023. Reconstructing graph diffusion history from a single snapshot. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1978–1988.
- Ruizhong Qiu, Zhe Xu, Wenxuan Bao, and Hanghang Tong. 2024b. Ask, and it shall be given: On the Turing completeness of prompting. *arXiv*, 2411.01992.
- Ruizhong Qiu, Weiliang Will Zeng, Hanghang Tong, James Ezick, and Christopher Lott. 2024c. How efficient is LLM-generated code? A rigorous & high-standard benchmark. *arXiv*, 2406.06647.
- Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. [In-context retrieval-augmented language models](#). In *Association for Computational Linguistics (ACL)*.
- Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791.
- Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordani, Philip Bachman, and Kaheer Suleman. 2017. Newsqa: A machine comprehension dataset. *ACL 2017*, page 191.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, et al. 2023a. [Survey on factuality in large language models: Knowledge, retrieval and domain-specificity](#). In *arXiv preprint arXiv:2310.07521*.
- Zhiruo Wang, Jun Araki, Zhengbao Jiang, Md Rizwan Parvez, and Graham Neubig. 2023b. Learning to filter context for retrieval-augmented generation. *arXiv preprint arXiv:2311.08377*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022a. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Tianxin Wei, Yifan Chen, Xinrui He, and Jingrui He. 2025a. [Connecting domains and contrasting samples: A ladder for domain generalization](#).
- Tianxin Wei, Fuli Feng, Jiawei Chen, Ziwei Wu, Jinfeng Yi, and Xiangnan He. 2021. Model-agnostic counterfactual reasoning for eliminating popularity bias in recommender system. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 1791–1800.
- Tianxin Wei, Zeming Guo, Yifan Chen, and Jingrui He. 2023. Ntk-approximating mlp fusion for efficient language model fine-tuning. In *International Conference on Machine Learning*, pages 36821–36838. PMLR.
- Tianxin Wei and Jingrui He. 2022. Comprehensive fair meta-learned recommender system. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1989–1999.

- Tianxin Wei, Bowen Jin, Ruirui Li, Hansi Zeng, Zhengyang Wang, Jianhui Sun, Qingyu Yin, Hanqing Lu, Suhang Wang, Jingrui He, et al. 2024a. Towards unified multi-modal personalization: Large vision-language models for generative recommendation and beyond. *arXiv preprint arXiv:2403.10667*.
- Tianxin Wei, Ruizhong Qiu, Yifan Chen, Yunzhe Qi, Jiacheng Lin, Wenju Xu, Sreyashi Nag, Ruirui Li, Hanqing Lu, Zhengyang Wang, Chen Luo, Hui Liu, Suhang Wang, Jingrui He, Qi He, and Xianfeng Tang. 2024b. [Robust watermarking for diffusion models: A unified multi-dimensional recipe](#).
- Tianxin Wei, Ziwei Wu, Ruirui Li, Ziniu Hu, Fuli Feng, Xiangnan He, Yizhou Sun, and Wei Wang. 2020. Fast adaptation for cold-start collaborative filtering with meta-learning. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 661–670. IEEE.
- Tianxin Wei, Yuning You, Tianlong Chen, Yang Shen, Jingrui He, and Zhangyang Wang. 2022b. Augmentations in hypergraph contrastive learning: Fabricated and generative. *Advances in neural information processing systems*, 35:1909–1922.
- Zhepei Wei, Wei-Lin Chen, and Yu Meng. 2025b. [InstructRAG: Instructing retrieval-augmented generation via self-synthesized rationales](#). In *The Thirteenth International Conference on Learning Representations*.
- Zhepei Wei, Wenlin Yao, Yao Liu, Weizhi Zhang, Qin Lu, Liang Qiu, Changlong Yu, Puyang Xu, Chao Zhang, Bing Yin, Hyokun Yun, and Lihong Li. 2025c. [Webagent-r1: Training web agents via end-to-end multi-turn reinforcement learning](#). *Preprint, arXiv:2505.16421*.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.
- Siye Wu, Jian Xie, Jiangjie Chen, Tinghui Zhu, Kai Zhang, and Yanghua Xiao. 2024. How easily do irrelevant inputs skew the responses of large language models? *arXiv preprint arXiv:2404.03302*.
- Haobo Xu, Yuchen Yan, Dingsu Wang, Zhe Xu, Zhichen Zeng, Tarek F Abdelzaher, Jiawei Han, and Hanghang Tong. Slog: An inductive spectral graph neural network beyond polynomial filter. In *Forty-first International Conference on Machine Learning*.
- Zhe Xu, Ruizhong Qiu, Yuzhong Chen, Huiyuan Chen, Xiran Fan, Menghai Pan, Zhichen Zeng, Mahashweta Das, and Hanghang Tong. 2024. Discrete-state continuous-time diffusion for graph generation. In *Advances in Neural Information Processing Systems*, volume 37.
- Yuchen Yan, Yuzhong Chen, Huiyuan Chen, Xiaoting Li, Zhe Xu, Zhichen Zeng, Lihui Liu, Zhining Liu, and Hanghang Tong. 2024a. Thecn: Temporal heterophilic graph convolutional network. *arXiv preprint arXiv:2412.16435*.
- Yuchen Yan, Yuzhong Chen, Huiyuan Chen, Minghua Xu, Mahashweta Das, Hao Yang, and Hanghang Tong. 2023a. From trainable negative depth to edge heterophily in graphs. *Advances in Neural Information Processing Systems*, 36:70162–70178.
- Yuchen Yan, Yongyi Hu, Qinghai Zhou, Lihui Liu, Zhichen Zeng, Yuzhong Chen, Menghai Pan, Huiyuan Chen, Mahashweta Das, and Hanghang Tong. 2024b. Pacer: Network embedding from positional to structural. In *Proceedings of the ACM on Web Conference 2024*, pages 2485–2496.
- Yuchen Yan, Yongyi Hu, Qinghai Zhou, Shurang Wu, Dingsu Wang, and Hanghang Tong. 2024c. Topological anonymous walk embedding: A new structural node embedding approach. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 2796–2806.
- Yuchen Yan, Baoyu Jing, Lihui Liu, Ruijie Wang, Jinning Li, Tarek Abdelzaher, and Hanghang Tong. 2023b. Reconciling competing sampling strategies of network embedding. *Advances in Neural Information Processing Systems*, 36:6844–6861.
- Yuchen Yan, Lihui Liu, Yikun Ban, Baoyu Jing, and Hanghang Tong. 2021. Dynamic knowledge graph alignment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 4564–4572.
- Yuchen Yan, Qinghai Zhou, Jinning Li, Tarek Abdelzaher, and Hanghang Tong. 2022. Dissecting cross-layer dependency inference on multi-layered interdependent networks. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 2341–2351.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. [Hotpotqa: A dataset for diverse, explainable multi-hop question answering](#). In *Empirical Methods in Natural Language Processing (EMNLP)*.
- Hangting Ye, Zhining Liu, Wei Cao, Amir M Amiri, Jiang Bian, Yi Chang, Jon D Lurie, Jim Weinstein, and Tie-Yan Liu. 2023. Web-based long-term spine treatment outcome forecasting. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3082–3092.
- Hyunsik Yoo, SeongKu Kang, Ruizhong Qiu, Charlie Xu, Fei Wang, and Hanghang Tong. 2025a. Embracing plasticity: Balancing stability and plasticity in

- continual recommender systems. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*.
- Hyunsik Yoo, Ruizhong Qiu, Charlie Xu, Fei Wang, and Hanghang Tong. 2025b. Generalizable recommender system during temporal popularity distribution shifts. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.
- Hyunsik Yoo, Zhichen Zeng, Jian Kang, Ruizhong Qiu, David Zhou, Zhining Liu, Fei Wang, Charlie Xu, Eunice Chan, and Hanghang Tong. 2024. Ensuring user-side fairness in dynamic recommender systems. In *Proceedings of the ACM on Web Conference 2024*, pages 3667–3678.
- Erxin Yu, Jing Li, and Chunpu Xu. 2024. PopALM: Popularity-aligned language models for social media trendy response prediction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12867–12878.
- Mert Yuksekogonul, Varun Chandrasekaran, Erik Jones, Suriya Gunasekar, Ranjita Naik, Hamid Palangi, Ece Kamar, and Besmira Nushi. 2024. [Attention satisfies: A constraint-satisfaction lens on factual errors of language models](#). In *International Conference on Learning Representations (ICLR)*.
- Zhichen Zeng, Boxin Du, Si Zhang, Yinglong Xia, Zhining Liu, and Hanghang Tong. 2024a. Hierarchical multi-marginal optimal transport for network alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16660–16668.
- Zhichen Zeng, Ruizhong Qiu, Zhe Xu, Zhining Liu, Yuchen Yan, Tianxin Wei, Lei Ying, Jingrui He, and Hanghang Tong. 2024b. Graph mixup on approximate gromov–wasserstein geodesics. In *Forty-first International Conference on Machine Learning*.
- Zhichen Zeng, Si Zhang, Yinglong Xia, and Hanghang Tong. 2023a. Parrot: Position-aware regularized optimal transport for network alignment. In *Proceedings of the ACM Web Conference 2023*, pages 372–382.
- Zhichen Zeng, Ruike Zhu, Yinglong Xia, Hanqing Zeng, and Hanghang Tong. 2023b. Generative graph dictionary learning. In *International Conference on Machine Learning*, pages 40749–40769. PMLR.
- Yuheng Zhang, Dian Yu, Tao Ge, Linfeng Song, Zhichen Zeng, Haitao Mi, Nan Jiang, and Dong Yu. 2025. Improving llm general preference alignment via optimistic online mirror descent. *arXiv preprint arXiv:2502.16852*.
- Zheng Zhao, Emilio Monti, Jens Lehmann, and Haytham Assem. 2024. Enhancing contextual understanding in large language models through contrastive decoding. *arXiv preprint arXiv:2405.02750*.
- Lecheng Zheng, Baoyu Jing, Zihao Li, Hanghang Tong, and Jingrui He. 2024a. [Heterogeneous contrastive learning for foundation models and beyond](#). In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, pages 6666–6676. ACM.
- Lecheng Zheng, Baoyu Jing, Zihao Li, Zhichen Zeng, Tianxin Wei, Mengting Ai, Xinrui He, Lihui Liu, Dongqi Fu, Jiakuan You, Hanghang Tong, and Jingrui He. 2024b. [Pyg-ssl: A graph self-supervised learning toolkit](#). *CoRR*, abs/2412.21151.
- Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14544–14556.

A Reproducibility Details

QA pipeline details. We use PyTorch (Paszke et al., 2019) and HuggingFace Transformers (Wolf et al., 2020) to implement all involved language models (LMs). Specifically, the LMs used in the paper are the instruct fine-tuned version of Llama3.1 8B³ and 70B⁴, Mistral 7B (instruct-v0.3)⁵ and 12B (Nemo)⁶, as well as Qwen2.5 7B⁷ and 32B⁸. We use the official `apply_chat_template` function to get the input tokens from text input ($\tau(c, q)$) and let the model generate the answer based on it. We use greedy decoding with a temperature of 0 to get deterministic answers.

Dataset details. For all the datasets used, we test the LM on the official public validation split, with 7,405/4,212/7,785/12,836 context-question pairs for HotopotQA/NewsQA/TQA/NQ. The context passages in those datasets are retrieved from diverse sources including Wikipedia (Yang et al., 2018; Kwiatkowski et al., 2019), news articles (Trischler et al., 2017), and web search queries (Joshi et al., 2017). The detailed statistics and descriptions are reported in Table 5. We use Llama-3.1 tokenizer to compute the number of tokens in the context passage.

³<https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>

⁴<https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct>

⁵<https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3>

⁶<https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407>

⁷<https://huggingface.co/Qwen/Qwen2.5-7B-Instruct>

⁸<https://huggingface.co/Qwen/Qwen2.5-32B-Instruct>

Table 5: Dataset description and statistics.

Dataset	#ContextTokens		#Samples	Source
	Avg.	Max		
HotpotQA	1251.6	3346	7405	Wikipedia
NewsQA	625.24	940	4212	CNN News Article
TQA	892.89	1974	7785	Bing Search Query
NQ	249.38	2679	12836	Wikipedia

SELFELICIT implementation details. Unless otherwise specified, all SELFELICIT results in this paper use the default settings stated in the paper: the last 50% of layers are designated as evidence-reading layers, with an elicitation threshold $\alpha = 0.5$. We use SpaCy (Honnibal and Montani, 2017) to segment the context passages into sentences, and SELFELICIT performs the sentence-level evidence score computation and elicitation based on this segmentation. Note that this is only for SELFELICIT, for PROMPTELICIT, we directly search for exact text-level matches of the generatively extracted evidence to ensure that all extracted information is utilized.

B Additional Results and Analysis

B.1 Computational Efficiency

Due to space constraints, we report only the average running time per sample across all datasets in Table 1. Table 6 provides a detailed breakdown of efficiency results for each dataset, including the average inference time per sample (i.e., context-question pair) and the total number of tokens generated during inference. Note that Mistral-12B generates shorter answers (7.34 tokens on average) for each sample compared to Mistral-7B (24.06 tokens), thus showing a shorter inference time for each sample as it generates fewer tokens. We observe that: (i) While FULLELICIT is highly efficient as a naive method, it barely helps with the QA performance, as shown in Table 1. (ii) PROMPTELICIT, based on generative evidence extraction, requires the model to generate a large volume of extra tokens during inference, up to 852% more than direct QA, resulting in considerably slower inference times. (iii) As discussed before, the response to COT is not consistent for different LMs. Generally, for Llama and Mistral models, COT have little impact on the QA performance, and inference time as well. For Qwen-series models, COT obtains slightly better EM by generating the intermediate reasoning steps before answering, but this also leads to a great drop in F1 score and significantly higher inference time (e.g., +72.0% for Qwen2.5-7B) due to the genera-

tion of intermediate step tokens. (iv) SELFELICIT, by helping the model focus on contextual evidence, generates shorter but more accurate answers. Additionally, as SELFELICIT does not involve generative evidence extraction, it remains computationally efficient, with only 3-5% of the inference time overhead compared to PROMPTELICIT.

B.2 More Layer-wise Attention Visualization

In Figure 2, we present the differences in attention paid by different layers of the model to evidence versus non-evidence context information on the HotpotQA (Yang et al., 2018) dataset. We chose HotpotQA because it provides human-annotated sentence-level "supporting_facts" annotations. In this section, we extend our analysis to other datasets. Since human annotations are unavailable for these datasets, we use a simple rule to distinguish evidence from non-evidence sentences: any sentence containing at least one correct answer is considered evidence. While this is not a rigorous approach for labeling evidence sentences, it still allows us to visualize patterns and roughly validate whether our observations hold across different datasets. As shown in Figure 5, deeper layers of various LMs consistently demonstrate a strong ability to differentiate evidence across datasets.

B.3 Impact of Evidence Highlighting on Model Attention

To assess the effectiveness of SELFELICIT in directing model attention toward relevant context, we analyzed the average attention score per token for the selected evidence sentences before and after highlighting. We conducted this analysis across four QA datasets using Llama-3.1-8B, averaging attention scores over all layers and heads. As shown in Table 7, we observe a consistent increase in attention scores for the highlighted evidence across all datasets. This demonstrates that the highlighting strategy effectively shifts model focus toward critical contextual information, reinforcing its ability to retrieve and utilize evidence more effectively.

B.4 Sentence-level versus Token-level Eliciting

We now show that sentence-level evidence eliciting holds a significant advantage over token-level eliciting due to its ability to highlight evidence with better semantic coherence. Table 8 presents the experimental results using the llama3.1-8B model across four datasets. We compare the normal version SELFELICIT (sentence) and a new token-level

Table 6: Detailed runtime and efficiency results. we report the average inference time and number of generated tokens for each sample (i.e., context-question pair) for each dataset.

Model	Method	Dataset								Average		
		HotpotQA		NewsQA		TQA		NQ		Inference	#Generated	
		Time (ms)	#Tokens	Time (ms)	#Tokens	Time (ms)	#Tokens	Time (ms)	#Tokens	Time/sample (ms)	Tokens	
Llama-3.1	8B	BASE	161.46	6.68	290.76	11.79	230.18	6.72	213.18	10.42	223.90 (+0.0%)	8.90 (+0.0%)
		COT	163.14	6.52	290.86	11.77	230.21	6.76	213.85	10.36	224.51 (+0.3%)	8.85 (-0.6%)
		FULLELICIT	163.55	6.49	290.79	11.83	236.30	6.97	213.64	10.47	226.07 (+1.0%)	8.94 (+0.4%)
		PROMPTELICIT	1419.08	75.78	1718.02	85.18	2412.91	116.81	1117.64	61.40	1666.91 (+644.5%)	84.79 (+852.5%)
		SELFELICIT	204.65	6.98	318.74	10.54	297.71	7.44	233.78	9.83	263.72 (+17.8%)	8.70 (-2.3%)
	70B	BASE	910.68	5.17	1850.04	11.32	1397.99	5.95	1421.32	11.18	1395.01 (+0.0%)	8.40 (+0.0%)
		COT	917.37	5.15	1856.47	10.61	1398.79	5.88	1423.61	10.88	1399.06 (+0.3%)	8.13 (-3.2%)
		FULLELICIT	915.09	5.41	1851.00	11.13	1448.56	6.32	1434.97	10.31	1412.41 (+1.2%)	8.29 (-1.3%)
		PROMPTELICIT	8545.64	69.39	7712.69	55.77	8283.76	56.14	7139.59	60.06	7920.42 (+467.8%)	60.34 (+618.1%)
		SELFELICIT	1198.53	6.05	1854.49	9.48	1672.19	6.15	1398.81	9.88	1531.00 (+9.7%)	7.89 (-6.1%)
Mistral	7B	BASE	468.26	21.60	564.08	24.60	479.97	18.67	647.02	31.38	539.83 (+0.0%)	24.06 (+0.0%)
		COT	489.55	23.31	568.91	25.27	502.71	19.59	688.56	34.02	562.43 (+4.2%)	25.55 (+6.2%)
		FULLELICIT	477.01	19.74	567.08	23.57	479.99	18.61	647.23	31.78	542.83 (+0.6%)	23.42 (-2.7%)
		PROMPTELICIT	1444.44	70.58	2043.11	97.09	2230.07	102.03	1788.88	89.59	1876.62 (+247.6%)	89.82 (+273.3%)
		SELFELICIT	330.42	12.74	493.56	18.33	440.59	13.32	462.43	21.00	431.75 (-20.0%)	16.34 (-32.1%)
	12B	BASE	205.88	5.62	380.22	10.46	287.35	5.32	251.74	7.97	281.30 (+0.0%)	7.34 (+0.0%)
		COT	208.52	5.57	381.90	10.09	287.36	5.35	258.02	7.41	283.95 (+0.9%)	7.11 (-3.2%)
		FULLELICIT	208.39	5.65	380.59	10.54	289.87	5.41	254.23	7.90	283.27 (+0.7%)	7.37 (+0.4%)
		PROMPTELICIT	1273.57	46.77	1560.04	52.11	1749.14	54.64	1226.14	46.18	1452.22 (+416.3%)	49.93 (+580.0%)
		SELFELICIT	278.07	6.29	428.54	9.02	385.85	5.45	260.08	7.47	338.14 (+20.2%)	7.06 (-3.9%)
Qwen2.5	7B	BASE	143.78	6.36	303.72	13.53	247.26	8.26	286.31	15.49	245.27 (+0.0%)	10.91 (+0.0%)
		COT	449.74	25.10	407.30	20.06	370.10	15.93	460.24	26.51	421.85 (+72.0%)	21.90 (+100.7%)
		FULLELICIT	145.24	6.18	306.38	12.89	255.94	6.85	291.09	16.35	249.66 (+1.8%)	10.57 (-3.1%)
		PROMPTELICIT	1109.49	62.77	1227.45	64.27	1312.63	64.84	1005.11	58.32	1163.67 (+374.4%)	62.55 (+473.3%)
		SELFELICIT	194.21	6.90	357.74	12.92	292.41	7.59	313.18	16.05	289.38 (+18.0%)	10.87 (-0.4%)
	32B	BASE	624.64	6.55	1271.39	15.35	773.30	6.51	1044.09	14.50	928.35 (+0.0%)	10.73 (+0.0%)
		COT	625.01	7.46	1447.84	18.24	871.12	8.13	1048.22	14.59	998.05 (+7.5%)	12.10 (+12.8%)
		FULLELICIT	636.89	6.63	1275.19	15.06	780.95	6.80	1052.68	14.21	936.43 (+0.9%)	10.67 (-0.5%)
		PROMPTELICIT	5556.76	82.86	5468.44	76.23	5290.17	69.83	4096.52	62.62	5102.97 (+449.7%)	72.89 (+579.5%)
		SELFELICIT	738.98	6.32	1220.44	12.29	1011.58	7.02	948.09	11.47	979.77 (+5.5%)	9.28 (-13.5%)

Table 7: Effect of highlighting on model attention, measured as average attention per token (APT) for selected evidence sentences.

APT ($\times 10^{-3}$)	HotpotQA	NewsQA	TQA	NQ
Before Highlighting	0.4362	0.3465	0.2090	0.9631
After Highlighting	0.6351	0.5783	0.3451	1.1917
After/Before Ratio	145.61%	166.91%	165.10%	123.73%

variant SELFELICIT (token) by treating each token as a separate "sentence". It is evident that token-level eliciting performs worse than sentence-level eliciting on all datasets. The underlying reason is that tokenization does not ensure each token represents a complete entity or concept, resulting in highlighted evidence that lacks full semantic meaning. For example, we notice that a year within the context, say "2002", can be tokenized into two tokens "200" and "2" and token-level eliciting could end up highlighting only the first token "200.". Similar "partial highlighting" issues also occur with names, locations, and other uncommon phrases. In contrast, sentence-level eliciting highlights the entire sentence, maintaining semantic continuity in the processed context, which better aids the model in utilizing context information for QA.

Table 8: Sentence-level versus token-level eliciting.

Elicit Level	Dataset			
	HotpotQA	NewsQA	TQA	NQ
Metric: EM				
Token	66.5	63.5	74.0	57.5
Sentence	68.5	66.9	79.4	64.0
Metric: Token F1				
Token	68.6	58.4	70.9	61.6
Sentence	69.5	60.8	72.7	67.8

B.5 Alternative Highlighting Strategies

In our primary experiments, we utilized in-context highlighting, where relevant evidence is marked directly within the original passage. This approach aligns with the core insight of SELFELICIT—leveraging the LM’s inherent ability to identify relevant evidence—while maintaining the semantic continuity of the input. However, alternative methods of evidence integration could potentially influence the model’s ability to utilize highlighted information effectively. To explore this, we conducted additional experiments using non-in-context highlighting, where all extracted evidence sentences were aggregated and appended either to the beginning or end of the original context. We evaluated these

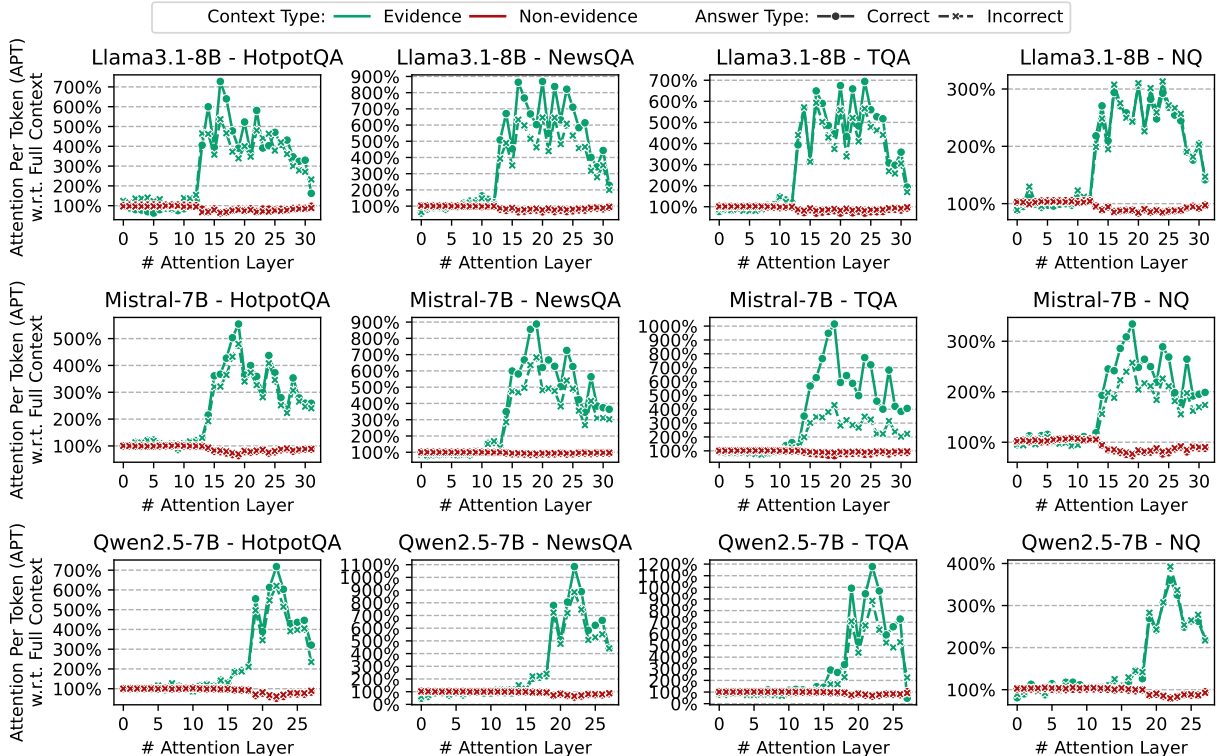


Figure 5: Across different LM families and datasets, the deep attention layers highlight the crucial evidence sentences within the context, even when the LM gives incorrect answers (dashed lines). The X-axis is the depth of attention layers, and the Y-axis is the ratio of the average attention per token (APT) in the evidence/non-evidence sections to the APT across the entire context. For the HotpotQA (Yang et al., 2018) dataset, We leverage the “supporting_facts” annotations to differentiate evidence and non-evidence sentences within the context. For other datasets, we treat a context sentence as evidence if it contains at least one of the correct answers.

variations using Llama-3.1-8B across four datasets and reported F1 scores in Table 9. The results in-

Table 9: Comparison of different highlighting strategies.

Highlighting	HotpotQA	NewsQA	TQA	NQ
-	57.7	56.2	66.1	61.6
In-context	69.5	60.8	72.7	67.8
Attach at Begin	69.8	59.8	69.3	67.8
Attach at End	69.0	58.8	67.0	65.6

indicate that in-context highlighting generally yields the best performance, likely due to its preservation of the original passage’s semantic flow, allowing the model to process highlighted evidence within its natural context. Appending evidence to the beginning or end also improves performance over the base model but is slightly less effective, particularly for multi-hop tasks like HotpotQA, where contextual relationships between facts are crucial.

We emphasize that SELFELICIT adopts in-context highlighting primarily for its simplicity and efficiency. However, our framework is flexible, and alternative highlighting strategies can be seamlessly incorporated. Exploring more effective context augmentation techniques remains an

intriguing direction for future work.

B.6 Potentially Uninformative Context

To further analyze SELFELICIT’s robustness in handling cases where the provided context may lack relevant evidence, we constructed an additional QA task based on HotpotQA. This experiment assesses whether SELFELICIT can correctly reject answers when the context does not support a response. Specifically, we randomly selected 50% of the samples as unanswerable, replacing their original context with unrelated passages from other samples. The correct response in these cases was set to: "I cannot answer based on the given context." We explicitly instructed the LLM to generate this response whenever relevant information was absent. In addition to standard EM and F1 scores, we introduced Rejection Accuracy, which measures the model’s ability to correctly identify and reject uninformative contexts. The results on Llama-3.1-8B are shown in Table 10.

These findings, along with Table 1 in the main paper, demonstrate that SELFELICIT does not degrade—rather, it slightly enhances—the model’s

Table 10: Results on potentially unanswerable QA task.

Method	F1	EM	RejectionAccuracy
Base	78.48	79.58	91.25
SELFELICIT	82.65	81.93	92.86

ability to recognize when the provided context lacks sufficient information for answering. At the same time, SELFELICIT significantly improves the model’s ability to utilize relevant information when it is available by highlighting the most crucial evidence in relation to the user query.

B.7 Highlighting versus Filtering Strategies

In this section, we compare SELFELICIT to the Filco method which filters context using a fine-tuned Llama-2 model (Wang et al., 2023b). Additionally, we explore the impact of highlighting versus filtering evidence. To analyze the impact of different evidence extraction and highlighting strategies, we conduct experiments on four SELFELICIT/Filco variants: (i) **SELFELICIT-highlight**: The standard SELFELICIT approach, where identified evidence sentences within the context are explicitly highlighted. (ii) **SELFELICIT-filter**: Instead of highlighting, this method extracts the identified evidence sentences and concatenates them to form a new, filtered context passage. (iii) **Filco-filter**: The standard Filco approach, which employs a fine-tuned Llama-2 model to generate a filtered version of the context for answering questions. (iv) **Filco-highlight**: This variant highlights the evidence sentences identified by Filco within the original context, rather than filtering them out.

This experimental setup enables a direct comparison between highlighting vs. filtering, as well as SELFELICIT vs. Filco, in their ability to enhance LMs’ utilization of contextual information. We use Llama-3.1-8B-Instruct as the base LM and evaluate performance across four QA datasets. The F1 scores are presented in Table 11.

Table 11: Comparison of SELFELICIT and Filco variants on four QA datasets, evaluated using F1 scores. SELFELICIT-highlight achieves the highest performance across all datasets.

Method	HotpotQA	NewsQA	TQA	NQ
Base	57.7	56.2	66.1	61.6
SELFELICIT-highlight	69.5	60.8	72.7	67.8
SELFELICIT-filter	62.4	55.6	66.5	65.1
Filco-highlight	67.8	57.2	68.9	65.3
Filco-filter	62.6	54.7	66.3	64.5

The key findings are: (i) **SELFELICIT consis-**

tently achieves the highest performance across all datasets, outperforming not only its filtering-based counterpart but also Filco, a more complex, fine-tuned approach. This highlights the effectiveness of inference-time evidence elicitation over training-based filtering. (ii) **Highlighting-based methods outperform their filtering-based counterparts for both SELFELICIT and Filco.** Filtering removes context that may contain auxiliary or indirect supporting information, such as intermediate reasoning steps, leading to information loss. (iii) **Filtering-based methods struggle particularly on NewsQA, which consists of news articles as context.** Unlike structured knowledge sources such as Wikipedia, news articles tend to be semantically cohesive—making it harder to extract self-contained and sufficiently informative evidence sentences. This may explain why Filco was originally evaluated on Wikipedia-based datasets, where evidence extraction is more straightforward.

These findings reinforce that highlighting key evidence within the original context is more effective than filtering, as it preserves both primary evidence and surrounding contextual cues that may contribute to reasoning. Moreover, SELFELICIT, despite being a lightweight inference-time method, surpasses Filco, which relies on additional model fine-tuning—demonstrating the advantage of leveraging pretrained LMs’ inherent ability to locate and utilize relevant evidence.

B.8 Examples with full context

Finally, we present the full context of examples from Table 2 to illustrate SELFELICIT’s ability to identify supporting facts/evidence within noisy long contexts. The results are shown in Tables 12 and 13. These examples are from HotpotQA’s distractor setting, where the context passage contains distracting information retrieved from Wikipedia using the question as a query (Yang et al., 2018). Note that these distractors are not completely irrelevant random noise that can be easily filtered out by retrieval systems. Instead, they appear related to the question but do not actually support answering it, serving as "hard negatives." This scenario is quite common in practice and significantly impacts the performance of retrieval-augmented generation (Wu et al., 2024; Cuconasu et al., 2024). As shown in Tables 12 and 13, the LM struggles to effectively use contextual evidence to provide correct answers under such a situation. Despite this, SELFELICIT accurately highlights the critical

evidence within the noisy long context, helping the model focus on the most relevant facts and thus arrive at the correct answer.

C Additional Discussions

C.1 Failure Case Analysis and Potential Improvements

While SELFELICIT significantly enhances LM’s ability to identify and utilize relevant evidence, there are cases where it fails to aid the model in generating correct answers. Through manual analysis, we found that incomplete evidence highlighting is a primary cause of failure. Consider the following example where Llama-3.1-8B, both with and without SELFELICIT, gives a negative response.

An example case with $\alpha = 0.5$

Context: Alexander S. Wadsworth Commodore Alexander Scammel Wadsworth (1790–April 5, 1851) was an officer of the United States Navy. His more than 40 years of active duty included service in the War of 1812. War of 1812 **The War of 1812 (1812–1815) was a conflict fought between the United States, the United Kingdom, and their respective allies.** Historians in Britain often see it as a minor theater of the Napoleonic Wars; in the United States and Canada, it is seen as a war in its own right ...
Question: Alexander S. Wadsworth served during the war that was fought between which two countries?
Base Model: I cannot answer based on the given context. ✗
SELFELICIT: I cannot answer based on the given context. ✗

We can see that although SELFELICIT successfully highlighted the sentence: "The War of 1812 (1812–1815) was a conflict fought between the United States, the United Kingdom, and their respective allies." as important evidence, this alone was insufficient for the model to infer that Alexander S. Wadsworth served in this war. We observed that lowering α can mitigate such failures by allowing SELFELICIT to highlight more comprehensive evidence. Reducing α to 0.1 enabled the method to additionally highlight the missing link—"Commodore Alexander Scammel Wadsworth (1790–April 5, 1851) was an officer of the United States Navy. His more than 40 years of active duty included service in the War of 1812.", and lead to the correct answer.

An example case with $\alpha = 0.1$

Context: Alexander S. Wadsworth **Commodore Alexander Scammel Wadsworth (1790–April 5, 1851) was an officer of the United States Navy. His more than 40 years of active duty included service in the War of 1812.** War of 1812 **The War of 1812 (1812–1815) was a conflict fought between the United States, the United Kingdom, and their respective allies.** Historians in Britain often see it as a minor theater of the Napoleonic Wars; in the United States and Canada, it is seen as a war in its own right ...
Question: Alexander S. Wadsworth served during the war that was fought between which two countries?
Base Model: I cannot answer based on the given context. ✗
SELFELICIT: United States, the United Kingdom, and their respective allies. ✓

This suggests that a fixed threshold may not be optimal across different questions and contexts, and SELFELICIT could potentially benefit from a more adaptive thresholding mechanism: Instead of using a fixed α , a dynamic threshold that adjusts based on the complexity of the input context or question could provide more flexible evidence selection. Nevertheless, as shown and discussed in Figure 4, SELFELICIT remains robust to the choice of α , with a broad range of values yielding consistently strong improvements. Even with static thresholding, SELFELICIT provides stable and significant gains in LM performance across QA tasks.

C.2 Understanding Layer-Wise Evidence Utilization in Language Models

Our analysis reveals that while intermediate layers of language models effectively emphasize relevant evidence within the context, this emphasis does not necessarily translate into correct final responses. As shown in Figure 2 and 5, attention to key evidence tends to diminish from mid-layers to the final layers, indicating that the final layers do not prioritize contextual evidence as strongly as mid-layers do. This suggests that although models inherently recognize relevant information at intermediate layers, their final token selection during decoding is influenced by additional factors that may override this evidence, leading to incorrect responses.

This phenomenon aligns with recent findings in closed-book QA and controlled generation settings. Prior work has shown that factually correct tokens often receive higher decoding probabilities in intermediate layers, while incorrect tokens can maintain consistently high probabilities across all layers and still be selected as the final output (Chuang et al., 2024). Similarly, studies on the "overthinking" behavior of LMs indicate that removing the last few layers can sometimes improve factual accuracy, implying that deeper layers may introduce unnecessary complexity or spurious reasoning (Halawi et al., 2024). Our findings provide further empiri-

cal support for these observations, explaining why models can correctly identify relevant evidence yet still produce incorrect answers.

C.3 Potential Risks

Generative AI tools such as language models have an increasing impact on our daily lives in the era of big data and AI (Yan et al., 2021, 2022, 2023a,b, 2024a,b,c; Xu et al., 2024; Ban et al., 2021; Lin et al., 2024a,b; Guo et al., 2023; Wei et al., 2025c, 2024b; Qiu et al., 2024c,b,a, 2023, 2022; Zeng et al., 2023a,b, 2024a,b; Qiu and Tong, 2024; Wei et al., 2025a; Ai et al., 2023; Yoo et al., 2025a,b, 2024; He et al., 2023; Bao et al., 2023; Wei et al., 2023, 2024a; Wei and He, 2022; Wei et al., 2022b, 2020, 2021; Li et al., 2025), such as finance (Chan et al., 2024; Liu et al., 2024c) and healthcare (Ye et al., 2023; Liu et al., 2024b), especially with the recent trends of foundation models (Zheng et al., 2024a; Fu et al., 2024; Zheng et al., 2024b; Li et al., 2024b; Zhang et al., 2025; Ai et al., 2025). This study focuses on enhancing language models' ability to effectively utilize information from contextual documents. However, retrieved documents from the internet may contain unethical or discriminatory content, which the model might read and incorporate into its outputs. While addressing ethical concerns in Retrieval-Augmented Generation (RAG) or fact-based question-answering tasks (Liu et al., 2024a, 2022, 2021a) is beyond the scope of this work, such issues can typically be mitigated by using detectors to filter harmful information from the context documents provided to the language model, or other general techniques such as ensemble multi-model answers (Liu et al., 2021b, 2020a,b) or watermarking (Chen et al., 2024a) AI-generated contents.

C.4 Usage of Artifacts and AI Assistants

All models and datasets used in this study are publicly available on HuggingFace, and we adhered to their respective licenses and terms of use, limiting our work to non-commercial academic research. These models and datasets have been reviewed by their developers/creators to minimize the inclusion of personally identifiable information or offensive content and are widely adopted by the research community. The datasets primarily consist of English-language content and focus on fact-based question-answering tasks. We used AI tools to assist with language refinement during the writing process, but the paper contains no AI-generated

paragraphs. All material has been carefully reviewed to ensure accuracy and adherence to ethical standards.

Table 12: Full context of example #1 & #2 from Table 2 demonstrating how our method assists the LM in context-based QA, with **blue** text highlighting the evidence selected by SELFELICIT. The base LM used is Llama3.1-8B-Instruct.

	Full Context Passage	Question & Answers
True or False	<p>Tiger Please is an Indie / Alternative five-piece band from Cardiff, Wales. The band formed in August 2008. The band's influences are U2, Sigur Rós, Kings of Leon, John Mayer and Counting Crows. They signed with Walnut Tree Records in 2009 and released their debut EP "They Don't Change Under Moonlight". "Kerrang!" magazine, "Rock Sound" magazine, and "Classic Rock" magazine praised the EP and featured the band on the "Rock Sound" and "Classic Rock" cover-mount albums. The band toured with Kids In Glass Houses, InMe, Twin Atlantic and Funeral For A Friend. Taxi Violence is a South African rock band from Cape Town. The group consists of George van der Spy (vocals), Jason Ling (bass), Louis Nel (drums), Rian Zietsman (guitar) and Loedi van Renen (guitar/bass). They have released five studio albums: "Untie Yourself" (2006), "The Turn" (2009), "" (2011), "Soul Shake" (2013), and "Tenfold" (2014). They are influenced by bands such as: Black Rebel Motorcycle Club, Queens of the Stone Age, Led Zeppelin, Jimi Hendrix, Nirvana, and Pearl Jam. This is discography of the American rock band Black Rebel Motorcycle Club Black Rebel Motorcycle Club (often abbreviated as BRMC) is an American rock band from San Francisco, California. The group consists of Peter Hayes (vocal, guitar, harmonica), Robert Levon Been (vocal, bass, guitar), and Leah Shapiro (drums). Former drummer Nick Jago left the band in 2008 to focus on his solo project. Skybombers is a rock band from Melbourne. They were formed as Collusion by Scotch College students Hugh Gurney, Ravi Sharma, Scott McMurtrie and Sam Bethune. They later changed to Skybombers, a name inspired by an icy-pole. Their placing a demo song "It Goes Off" on MySpace brought them their first TV appearances. They had early international attention when "It Goes Off" of their EP "Sirens" made the most-played list on L.A.'s Indie 103.1 and played a showcase gig at The Viper Room. They have toured Australia, Japan and USA. and their debut album "Take Me To Town" was recorded in L.A. with Rick Parker (Black Rebel Motorcycle Club). The band made their way on video game media in 2007 when "It Goes Off" appeared on the soundtrack for "Burnout Dominator", the song later reappeared on "Burnout Paradise" in early 2008. Black Carousel was recorded in LA, again with Rick Parker at the helm. Masaki Liu, sometimes referred to as "Saki", is the engineer and producer operating One Way Studio, a digital recording studio in Benicia, California. Masaki has recorded and produced music for many bands, including Five Iron Frenzy, Black Rebel Motorcycle Club, The Echoing Green, The W's and Yellow Second. Dan Russell is a musician and songwriter in addition to an artist manager and advocate, musician, songwriter, concert promoter, record producer and music supervisor for television and film. A graduate of Walpole High School in Massachusetts and later Barrington College, Russell is known for managing both the American rock band The Call and songwriter Michael Been and has worked in various capacities with such artists as Black Rebel Motorcycle Club, Sam Philips, Mark Heard, U2 and Robin Lane, Ramona Silver, Vigilantes of Love, among others. Black Rebel Motorcycle Club: Live is a DVDs of Black Rebel Motorcycle Club concert footage captured from three sold out shows in Berlin, Dublin and Glasgow, and chronicles the end of the band's 2007 tour in support of "Baby 81". Additionally, it includes intimate, behind-the-scenes footage, glimpses into the making of 2005's Howl and is rounded out with a bonus live album featuring 14 songs. KAV (Kav Sandhu) is a British musician from Leicester UK, based in Los Angeles. KAV played guitar with British band Happy Mondays for 4 years after helping reform the band with frontman Shaun Ryder in 2004. He launched his solo project under moniker "KAV" in 2008 with long-time friend and drummer Jim (James) Portas. His solo material has been compared by the media to everyone from Iggy & The Stooges, Black Rebel Motorcycle Club, Primal Scream, Kasabian, The Rolling Stones & Bob Dylan. He plays live with a full live band, which sometimes features guest musicians from various bands. Vagrant Records is an American record label based in California. It was founded in 1995 by Rich Egan and Jon Cohen. The label focuses on rock but features artists in a variety of other genres including folk, soul, electronic, and pop. It is home to artists such as The 1975, Death Spells, Eels, Bad Suns, Edward Sharpe and the Magnetic Zeroes, CRUISR, Active Child, PJ Harvey, School of Seven Bells, Black Rebel Motorcycle Club, James Vincent McMorrow, Black Joe Lewis, Wake Owl, Blitzen Trapper, and Bombay Bicycle Club. Originally, Vagrant Records was mostly focused on emo bands such as Dashboard Confessional, Saves the Day, The Get Up Kids, and Alkaline Trio.</p>	<p>Question: Are the bands Tiger Please and Black Rebel Motorcycle Club from the same country?</p> <p>True Answer: No.</p> <p>Base: Yes. ✗</p> <p>+SelfElicit: No. ✓</p>
Comparison	<p>Home Monthly was a monthly women's magazine published in Pittsburgh, Pennsylvania in the late 19th century. "The Strategy of the Were-Wolf Dog" is a short story by Willa Cather. It was first published in "Home Monthly" in December 1896. The Count of Crow's Nest is a short story by Willa Cather. It was first published in "Home Monthly" in October 1896. Mirabella was a women's magazine published from June 1989 to April 2000. It was created by and named for Grace Mirabella, a former "Vogue" editor in chief, in partnership with Rupert Murdoch. "Nanette: An Aside" is a short story by Willa Cather. It was first published in "Courier" on 31 July 1897 and one month later in "Home Monthly". "The Prodigies" is a short story by Willa Cather. It was first published in "Home Monthly" in July 1897. "A Resurrection" is a short story by American writer Willa Cather. It was first published in "Home Monthly" in April 1897. Tommy, the Unsentimental is a short story by Willa Cather. It was first published in "Home Monthly" in August 1896. The Princess Baladina is a short story by Willa Cather. It was first published in "Home Monthly" in 1896 under the pseudonym of Charles Douglass. "The Way of the World" is a short story by Willa Cather. It was first published in "Home Monthly" in April 1898.</p>	<p>Question: Which women's magazine was published first, Mirabella or Home Monthly?</p> <p>True Answer: Home Monthly.</p> <p>Base: Mirabella. ✗</p> <p>+SelfElicit: Home Monthly. ✓</p>

Table 13: Full context of example #3 & #4 from Table 2 demonstrating how our method assists the LM in context-based QA, with **blue** text highlighting the evidence selected by SELFELICIT. The base LM used is Llama3.1-8B-Instruct.

	Full Context Passage	Question & Answers
Fact Retrieval	<p>The 1987 European Cup Final was a football match held at the Prater Stadium, Vienna, on 27 May 1987, that saw Porto of Portugal defeat Bayern Munich of West Germany 2–1. Both sides were missing key players: the Portuguese were without their injured striker Fernando Gomes, while the Germans were missing their sweeper, and captain, Klaus Augenthaler, who was suspended, along with striker Roland Wohlfarth and midfielder player Hans Dorfner, who were both injured. The Portuguese side fought back from 1–0 down to win their first European Cup, with the goals coming from a back heel by Rabah Madjer and a volley from Juary, after a Ludwig Kögl header had given Bayern the lead in the first half. The final was the first European Cup final that Bayern, and their captain Lothar Matthäus would lose to successive late goals, repeated 12 years later in the 1999 UEFA Champions League Final against Manchester United. The 1993 European Cup Winners' Cup Final was a football match contested between Parma of Italy and Royal Antwerp of Belgium. The final was held at Wembley Stadium in London, England on 12 May 1993. It was the final match of the 1992–93 European Cup Winners' Cup and the 33rd European Cup Winners' Cup Final. Parma beat Antwerp 3–1 and in doing so became the eighth different Italian team to win a European trophy. The 1970 European Cup Winners' Cup Final was a football match between Manchester City of England and Górnik Zabrze of Poland on 29 April 1970 at Prater Stadium in Vienna, Austria. It was the final match of the 1969–70 European Cup Winners' Cup and the tenth European Cup Winners' Cup final. Both sides made their first appearance in a European final. Manchester City won the match 2–1 thanks to goals by Neil Young and Francis Lee. The victory was City's only European trophy. The 1977 European Cup Final was an association football match between Liverpool of England and Borussia Mönchengladbach of Germany on 25 May 1977 at the Stadio Olimpico in Rome, Italy (the venue was decided in Bern by the UEFA Executive Committee on 17 September 1976). The showpiece event was the final match of the 1976–77 season of Europe's premier cup competition, the European Cup. Both teams were appearing in their first European Cup final, although the two sides had previously met in the 1973 UEFA Cup Final, which Liverpool won 3–2 on aggregate over two legs. The 1988 European Cup Winners' Cup Final was a football match contested between Mechelen of Belgium and the defending champions, Ajax of Netherlands. It was the final match of the 1987–88 European Cup Winners' Cup and the 28th European Cup Winners' Cup Final. The final was held at Stade de la Meinau in Strasbourg, France. Mechelen won the match 1–0 thanks to a goal by Piet den Boer. The 1961 European Cup Winners' Cup Final was a football match contested between Fiorentina of Italy and Rangers of Scotland. It was the final of the 1960–61 European Cup Winners' Cup the first UEFA Cup Winners' Cup final. It was the only time that the final was played over two legs. The first leg was played at Ibrox Stadium, Glasgow and the second leg at the Stadio Comunale in Florence. It was Rangers first European final and in doing so became the first British team to reach the final of a European football competition. It was Fiorentina's second European final having previously reached the 1957 European Cup final. The 1987 European Cup Winners' Cup Final was a football match contested between Ajax of Netherlands and Lokomotive Leipzig of East Germany. It was the final match of the 1986–87 European Cup Winners' Cup and the 27th European Cup Winners' Cup Final. The final was held at Olympic Stadium in Athens, Greece. Ajax won the match 1–0 with a 20th-minute header from Marco van Basten. Lars Lunde (born 21 March 1964) is a Danish former professional football player, who played in the striker position. Lunde got his breakthrough with Brøndby IF in 1983, and he made his debut for the Denmark national football team in October 1983. He was sold to Young Boys Bern in Switzerland, before moving to German club Bayern Munich in 1986. He was a part of the Bayern team which won the German Bundesliga championship in 1987, and he came on as a late substitute when Bayern lost the 1987 European Cup Final to FC Porto. He played the last of his three matches for the Danish national team in April 1987, before leaving Bayern during the 1987–88 season. He went on to play for a number of smaller clubs, ending his career with FC Baden in Switzerland. The 1978 European Cup Final was an association football match between Liverpool of England and Club Brugge of Belgium on 10 May 1978 at Wembley Stadium, London, England (the venue was decided in Bern by the UEFA Executive Committee on 20 September 1977). It was the final match of the 1977–78 season of Europe's premier cup competition, the European Cup. Liverpool were the reigning champions and were appearing in their second European Cup final. Club Brugge were appearing in their first European Cup final. The two sides had met once before in European competition, when they contested the 1976 UEFA Cup Final, which Liverpool won 4–3 on aggregate. The 1985 European Cup Winners' Cup Final was a football match contested between Everton of England and Rapid Wien of Austria. It was the final match of the 1984–85 European Cup Winners' Cup and the 25th European Cup Winners' Cup Final. The final was held at Feijenoord Stadion in Rotterdam, Netherlands, on 15 May 1985. Everton, which dominated throughout, won the match 3–1 thanks to goals by Andy Gray, Trevor Steven and Kevin Sheedy. Everton were unable to defend the trophy: as league champions they would have entered the 1985–86 European Cup, but they were not permitted to play in either competition following the actions of rival Liverpool fans at the Heysel Stadium, which saw all English clubs banned from European competitions.</p>	<p>Question: Which team did Lars Lunde play for when defeated for the 1987 European Cup Final?</p> <p>True Answer: Bayern Munich</p> <p>BASE: FC Porto. ✗</p> <p>+SelfElicit: Bayern Munich. ✓</p>
Multi-hop Reasoning	<p>Mount Barker Junction railway station is a disused station on the Adelaide to Wolseley line serving the South Australian city of Mount Barker. Mount Gambier railway station was the junction station for the Naracoorte–Millicent and Mount Gambier–Heywood lines in the South Australian city of Mount Gambier. Frewville is a small suburb in the South Australian city of Adelaide. It is three kilometres south-east of Adelaide's central business district (CBD). The 2006 South Australian Super League was the first season of the South Australian Super League, the new top division of association football in South Australia, replacing the South Australian Premier League, which became the second division. It was also the first year that football in South Australia was run by the Football Federation of South Australia, which replaced the South Australian Soccer Federation. The season came down to a final round relegation battle between White City Woodville and Adelaide Olympic. Olympic lost 3–1 at Modbury while White City went down 1–0 away to Cumberland. This sent Olympic down to play in the Premier League in 2007. Adelaide City won the title with games to spare after being runaway leaders, finishing the season unbeaten. Norwood is a suburb of Adelaide, about 4 km east of the Adelaide city centre. The suburb is in the City of Norwood Payneham & St Peters, the oldest South Australian local government municipality, with a city population over 34,000. Whyalla railway station was the terminus station of the Whyalla line serving the South Australian city of Whyalla. Walter Frank Giffen (20 September 1861 in Norwood – 28 June 1949 in Adelaide) was an Australian cricketer who played in 3 Tests between 1887 and 1892. He was the brother of the great all-rounder George Giffen. The City of Burnside is a local government area with an estimated population of 44,300 people in the South Australian city of Adelaide. Burnside was founded in August 1856 as the District Council of Burnside, and was classed as a city in 1943. It is named after the property of an early settler and stretches from the Adelaide Parklands into the Adelaide foothills. It is bounded by Adelaide, Adelaide Hills Council, Campbelltown, Mitcham, Norwood Payneham and St Peters and Unley. The city has an area of 27.53 km². Glenunga is a small southern suburb of 2,539 people in the South Australian city of Adelaide. It is located five kilometres southeast of the Adelaide city centre. The name Glenunga is taken from an Aboriginal language "unga" meaning near and "glen" because of its proximity to Glen Osmond (see Manning's places of South Australia by Geoffrey H. Manning published in 1990). Bounded on the north by Windsor Road, the east by Portrush Road, the south-west by Glen Osmond Road and the west by Conyngham Street, the leafy suburb forms a rough triangular layout. It is close by to other Burnside council suburbs of Toorak Gardens and Glenside. Collina is a suburb of the Australian city of Griffith in the Riverina region of New South Wales. The suburb is in the City of Griffith local government area. Collina is 4 km northwest of the Griffith city centre and reflects the city's rapid growth in the early 2000s.</p>	<p>Question: Walter Giffen is from a suburb of which South Australian city?</p> <p>True Answer: Adelaide</p> <p>Base: Norwood. ✗</p> <p>+SelfElicit: Adelaide. ✓</p>