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

Zhining Liu, Rana Ali Amjad, Ravinarayana Adkathimar, Tianxin Wei, Hanghang Tong

University of Illinois Urbana-Champaign, Amazon Science

ACL 2025

Presenter: NGUYEN Tan Minh

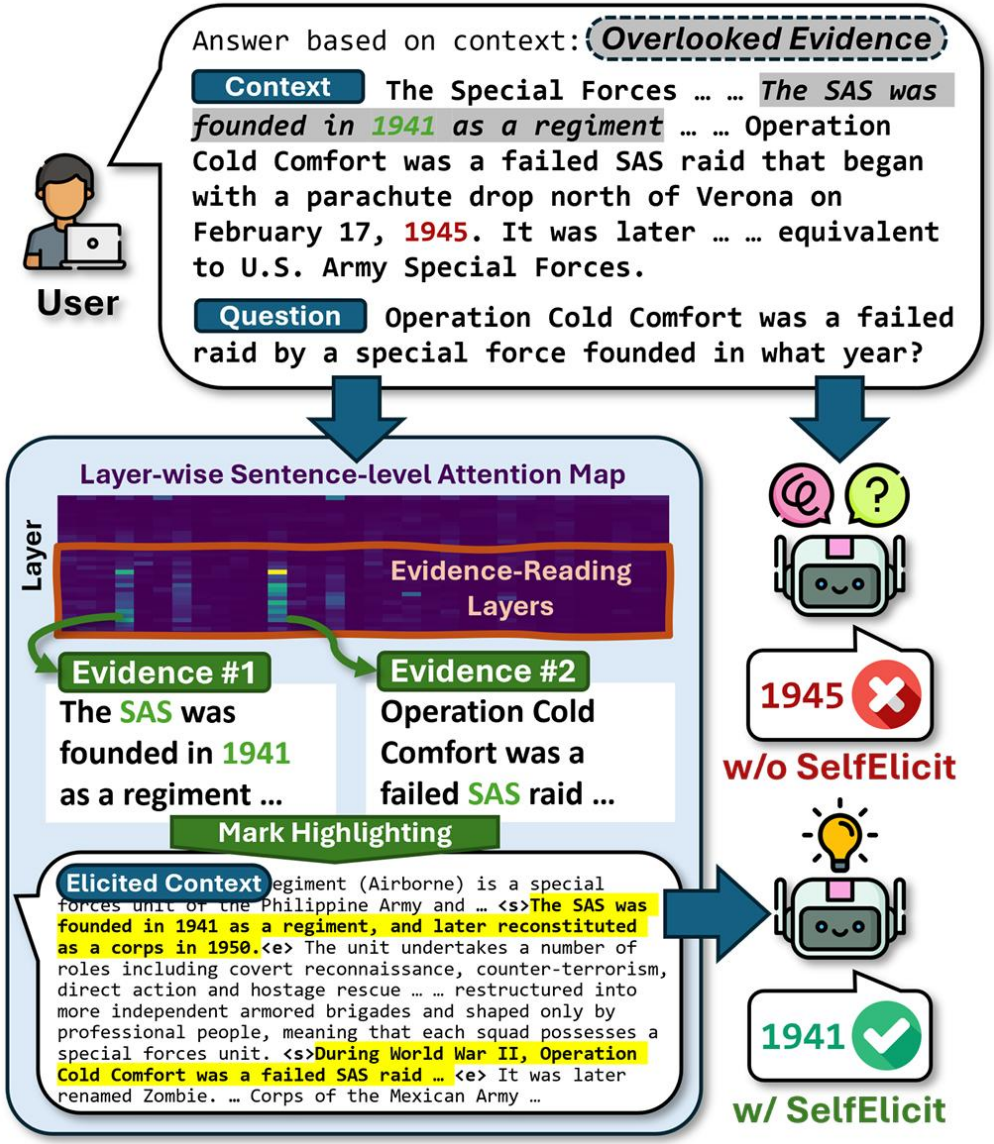
Paper is available at *2025.acl-long.448.pdf

Factual Accuracies & Hallucination

Retrieval-augmented generation (RAG) achieves remarkable capabilities.

A significant limitation remains:

- Fail to leverage supporting facts within context
- Lead to factual hallucination



Self-guided Evidence Eliciting

Leverage LLMs internal representation to identify relevant evidence

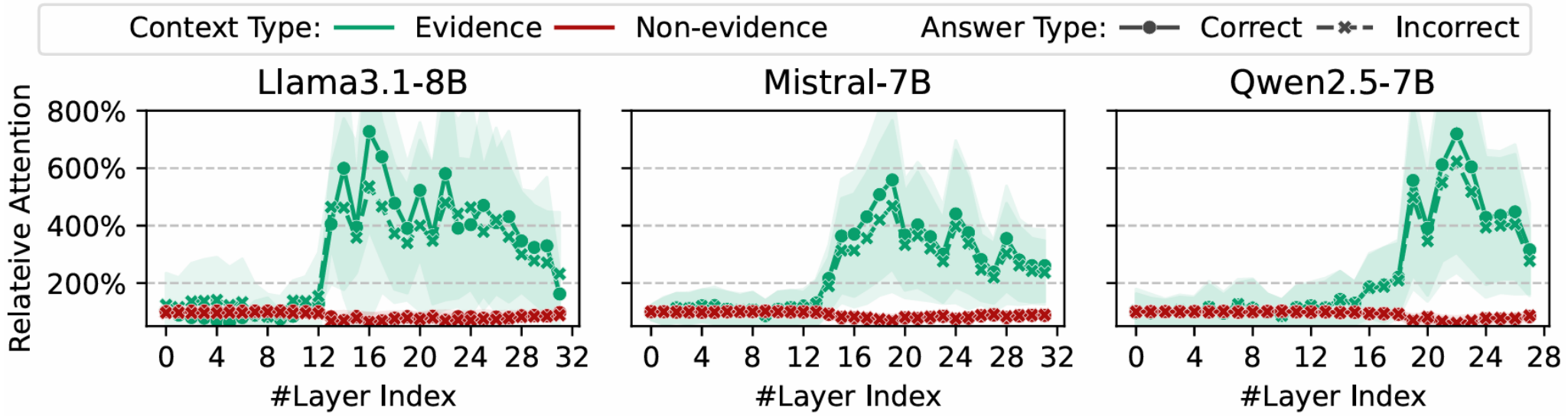
- Let a context-based QA: $\tau(c, q)$, c: $\{s_1, s_2, \ldots, s_m\}$
- Sentence-level attention vector $\bar{\boldsymbol{a}} \in \mathbb{R}^m$ is obtained from token level attention a^l as follows

$$\bar{\mathbf{a}}^{(\ell)} := [\bar{a}_1^{(\ell)}, \bar{a}_2^{(\ell)}, \cdots, \bar{a}_m^{(\ell)}] \in \mathbb{R}^m,$$

$$\text{where} \quad \bar{a}_i^{(\ell)} = \frac{1}{|t_{\mathbf{s}_i}^{\text{end}} - t_{\mathbf{s}_i}^{\text{start}} + 1|} \sum_{j=t_{\mathbf{s}_i}^{\text{start}}}^{t_{\mathbf{s}_i}^{\text{end}}} \mathbf{a}_j^{(\ell)}. \quad (2)$$

Self-guided Evidence Eliciting

Deeper layers pay greater attention to evidence in the context



Self-guided Evidence Eliciting

Sentence evidence score e_i

$$\mathbf{e} := [e_1, e_2, \cdots, e_m] \in \mathbb{R}^m,$$
$$\text{where} \quad e_i = \frac{1}{|\mathcal{L}_{ER}|} \sum_{\ell \in \mathcal{L}_{ER}} \bar{a}_i^{(\ell)} \qquad (3)$$

Predict evidence sentences $\mathcal{S}_{SE}$

$$\mathcal{S}_{\text{SE}} = \{\mathbf{s}_i | \mathbf{s}_i \in \mathcal{S}; e_i \geq \alpha \cdot \max(\mathbf{e})\}, \qquad (4)$$

Contextual Evidence Highlighting

- Place text markers <start_important> and <end_important> before and after each evidence sentence.
- Update task instruction to guide LLM towards highlighted sentences

SELFELICIT Prompt Template τ_{SEQA}

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

SelfElicit Procedure

SelfElicit is an inference, training-free method

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

$$\mathbf{e} := [e_1, e_2, \cdots, e_m] \in \mathbb{R}^m,$$
$$\text{where} \quad e_i = \frac{1}{|\mathcal{L}_{ER}|} \sum_{\ell \in \mathcal{L}_{ER}} \bar{a}_i^{(\ell)} \quad (3)$$

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

Algorithm 1 SELFELICIT

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

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

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

Experiments

Metrics: EM, Token-level F1

Datasets

Table 5: Dataset description and statistics.

Dataset	#ContextTokens Avg.	#ContextTokens Max	#Samples	Source
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

Experiments

SelfElicit Setup:

- Evidence Layer$\mathcal{L}_{ER}$: last 50% layers
- Evidence threshold $\alpha = 0.5$

Baselines:

- Chain-of-thought
- FullElicilit
- PromptElicit

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}`

Main Results

- FullElicit does not provide meaningful improvements

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

Main Results

- FullElicit does not provide meaningful improvements
- CoT shows inconsistency among LLM backbone

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

Main Results

- FullElicit does not provide meaningful improvements
- CoT shows inconsistency among LLM backbone
- **SelfElicit** significantly & consistently improves the performance

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

Main Results

SelfElicit highlights relevant evidence

	(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 Cathe	**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. ✓

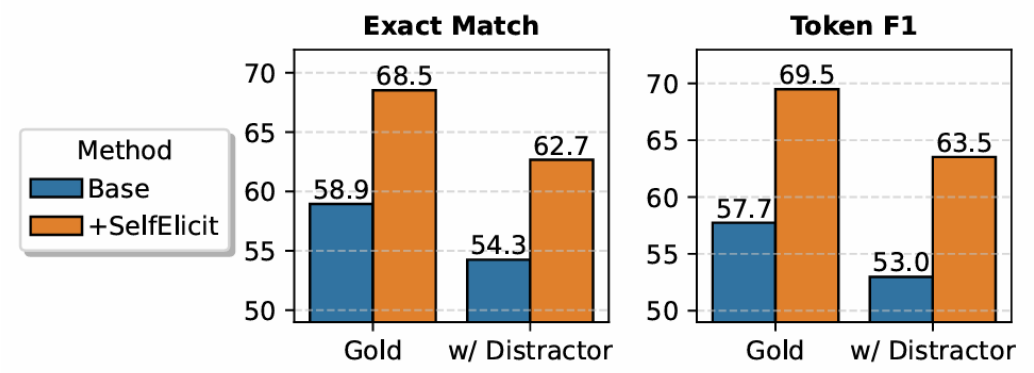
Table 3: SELFELICIT accurately identifies contextual evidence sentences across different datasets and models.

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

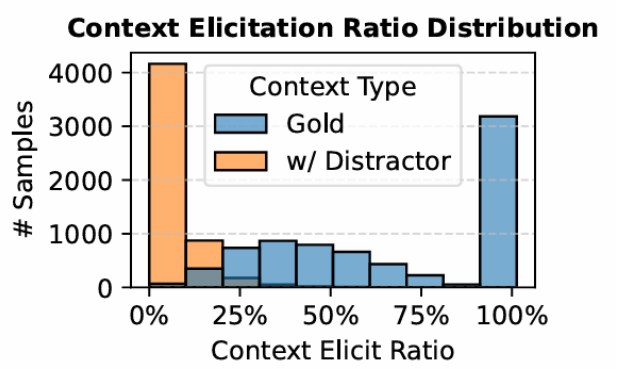
Analysis – Noisy context

SelfElicit is effective in context noise setting

- Noisier context, less highlighted sentences



(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.

Analysis – Choices of $\mathcal{L}_{ER}$

Model/Task-specific hyperparameters

Choosing last 50% of the layers achieves the best metrics

- Elicitation accuracy
- QA performance

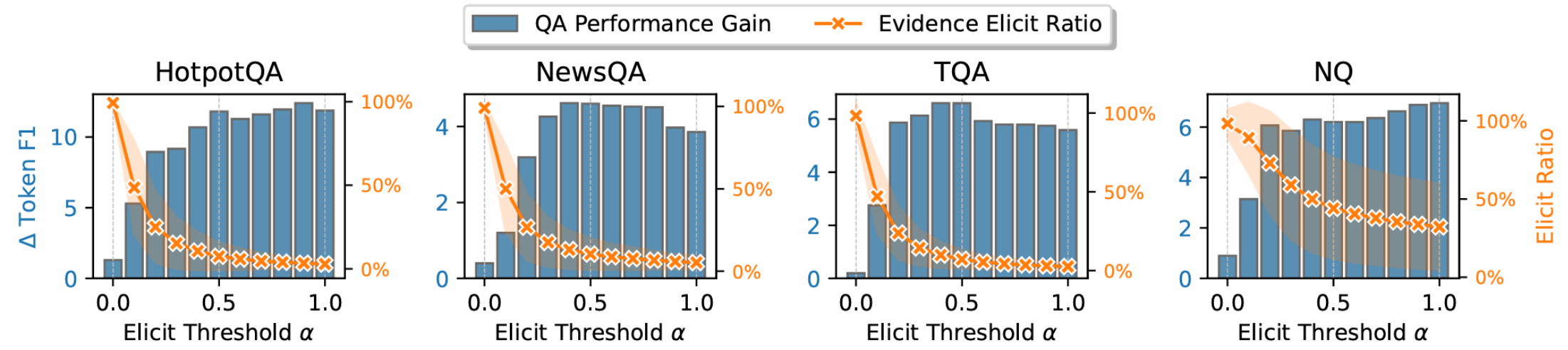
Table 4: Effect of different evidence-reading (ER) layer choices on elicit accuracy and QA performance.

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

Analysis – Choices of α

For simpler reasoning, best performance at $\alpha = 1$

Multi-hop reasoning , optimal $\alpha \approx 0.5$



Analysis – Highlighting v.s Filtering

- Filco: employs fine-tuned Llama-2 to identify evidence
- Filter: extractive summary of the original context
- Highlight: use special tokens to mark evidence

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

Conclusions

- SelfElicit leverages internal attention for automatic contextual evidence highlighting.
- SelfElicit improves the QA performance, factual faithfulness of LLMs.
- SelfElicit is efficient for inference, training-free, noise robustness.

Future work

- Adaptive threshold α based on LLM/data characteristics
- Extending SelfElicit to text generation tasks

Thank you for listening
Q&A