CAGE: COGNITIVE ATTRIBUTION GRAPHS FOR FAITHFUL INLINE CITATION GENERATION IN LONG-FORM QUESTION ANSWERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-form question answering increasingly relies on retrieved evidence to make LLM outputs verifiable, with inline citations tracing claims to source documents. However, existing systems often attach citations that are topically related but insufficient to support their claims. We identify *attribution ambiguity* as a structural challenge: end-to-end generation must implicitly resolve combinatorial claim–document assignments, obscuring evidential boundaries and increasing the risk of *evidence-boundary overrun*, where claims exceed cited support. To address this challenge, we propose CAGE (**C**ognitive **A**ttribution **G**raphs for Citation G**e**neration), a two-stage framework that introduces an explicit cognitive attribution map before answer generation. CAGE first trains a plug-and-play Cognitive Attribution Map Induction Model to construct answer-centered support subgraphs, aligning each semantic answer unit with supporting documents through explicit relations. A Structured Citation Reasoning Model then realizes these units as sentence-level claims with map-aligned citations. Experiments on ASQA, ELI5, and ExpertQA show that CAGE achieves best performance, demonstrating the effectiveness of attribution-space contraction and map-guided citation generation.

1 INTRODUCTION

Large language models (LLMs) can generate fluent long-form responses for knowledge-intensive tasks, yet hallucinations and unreliable source attribution undermine the verifiability and trustworthiness of their outputs Farquhar et al. (2024); Yan et al. (2025); Chen et al. (2024). Retrieval-augmented generation (RAG) partially mitigates this issue by grounding generation in retrieved documents Pan et al. (2024; 2023), but document grounding alone does not establish whether each individual claim is supported by sufficient evidence. This limitation is especially consequential in downstream applications that consume LLM outputs without independent verification Zhang et al. (2026); Luo et al. (2025), making explicit claim-level attribution essential for trustworthy long-form generation.

Citation-grounded long-form generation requires claims to include citations to supporting documents Gao et al. (2022); Bohnet et al. (2022). Yet when claims and citations are generated jointly, the model must simultaneously decide what to state and what evidence to cite without explicit claim-specific support relations. Under weak, incomplete, or distributed evidence, this coupled process may produce claims that exceed available support while attaching topically related but insufficient citations Sun et al. (2025); Qian et al. (2026). We call this *evidence-boundary overrun*, where a claim exceeds what its cited documents can strictly justify.

The risk of such overrun is amplified by *attribution ambiguity*: retrieved documents may be relevant to the question without specifying which subset supports each individual claim. A document may provide only partial evidence, while different claims may require different document combinations. For a fixed set of m candidate claims and n retrieved documents, each claim may be assigned any non-empty document subset, yielding up to $(2^n - 1)^m$ possible citation assignments. This combinatorial space does not itself imply unsupported generation, but resolving it implicitly during decoding makes it difficult to distinguish topical relevance from sufficient support. Explic-

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

itly determining claim-specific support relations before final generation can therefore improve the verifiability of citation-grounded outputs and reduce the risk of *evidence-boundary overrun*.

Cognitive-map theories suggest that complex reasoning can benefit from transforming implicit relations among observations into explicit relational structures for structured inference Mark et al. (2020); Whittington et al. (2022); Tan et al. (2025). Drawing on this perspective, citation-grounded generation should externalize latent relations between retrieved evidence and semantic answer units as an explicit attribution structure before final decoding. Such a structure can exclude invalid evidence assignments, organize supporting documents around answer units, and make claim-specific support relations explicit. This also explains why stronger prompting or larger generators may still produce missing or incorrect citations: *improving model capacity alone may be insufficient without an explicit mechanism that externalizes claim–document relations and constrains unsupported assignments before generation.*

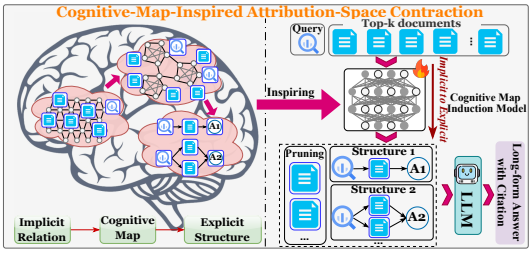


Figure 1: Motivation of CAGE. The human brain abstracts scattered observations into relational cognitive maps to organize and associate discrete information. CAGE simulates this workflow with a plug-and-play Cognitive Attribution Map Induction Model that converts retrieved documents into intermediate evidence structures, contracting the attribution space before LLM-based citation generation.

To this end, we introduce CAGE (**C**ognitive **A**ttribution **G**raphs for Citation G**e**neration), a two-stage framework for inline citation generation. As illustrated in Figure 1, CAGE first trains a plug-and-play *Cognitive Attribution Map Induction Model* to construct a cognitive attribution map from a question and its retrieved documents. The map consists of answer-centered support subgraphs that link answer units to their supporting documents and explicitly define the evidence boundary; unsupported documents are excluded, while the absence of supported answer units triggers evidence-driven abstention. In the second stage, a *Structured Citation Reasoning Model* reasons over the induced map, realizes its semantic answer nodes as coherent sentence-level claims, and generates inline citations aligned with the corresponding graph-specified document identifiers. CAGE externalizes answer–document attribution as validated support subgraphs, decoupling attribution from generation across heterogeneous backbones.

In summary, the main contributions are as follows:

- We identify *attribution ambiguity* as a structural cause of citation error in end-to-end generation, where unconstrained claim–document assignment induces a combinatorial citation-assignment space and allows topical relevance to substitute for evidential support.
- We propose CAGE, a two-stage framework that trains a plug-and-play Cognitive Attribution Map Induction Model and a Structured Citation Reasoning Model to explicitly determine semantic answer units and their supporting documents before final answer realization.
- We demonstrate best performance on ASQA, ELI5, and ExpertQA, and show that CAGE consistently improves proprietary LLMs, indicating that cognitive attribution map provide complementary gains beyond document filtering and enhance strong end-to-end systems.

2 Preliminaries

2.1 Task Definition: Inline Citation Generation

Given a question Q and top-k retrieved documents $\mathcal{D}_k = \{d_i\}_{i=1}^{k}$, inline citation generation requires an LLM $\mathcal{M}$ to produce a long-form answer grounded in specific source documents. We formalize the output as

$$\mathcal{Y} = \{(s_j, c_j)\}_{j=1}^{m} \leftarrow \mathcal{M}(Q, \mathcal{D}_k), \tag{1}$$

where s_j is an individual factual statement and $c_j \subseteq \mathcal{D}_k$ is the citation set supporting it, m is the number of claims. If $\mathcal{D}_k$ lacks sufficient evidence for Q, the model should perform evidence-driven abstention by returning an explicit refusal.

2.2 Cognitive Attribution Map

Given a query Q and $\mathcal{D}_k = \{d_1, \ldots, d_k\}$, the map $\mathbf{G}$ is defined as a collection of answer-centered support subgraphs:

$$\mathbf{G} = \{\mathcal{G}_r\}_{r=1}^{m}, \qquad 0 \leq m. \tag{2}$$

Each subgraph represents one semantic answer unit:

$$\mathcal{G}_r = (\mathcal{V}_r, \mathcal{E}_r; \rho_r), \qquad \mathcal{V}_r = \{Q\} \cup \mathcal{D}_r \cup \{a_r\}, \tag{3}$$

where $\mathcal{D}_r \subseteq \mathcal{D}_k$ is a non-empty supporting document set, a_r is a canonical semantic answer unit supported by $\mathcal{D}_r$, and ρ_r describes their support relation. The answer node specifies the content to be conveyed without fixing its final wording or position in the long-form answer.

A document node $d_i = (\mathrm{id}, \mathrm{name}, \mathrm{content})$ stores its citation ID, title, and evidence text. The query node follows the same schema, with $\mathrm{id}(Q) = \mathtt{Q}$, $\mathrm{name}(Q)$ containing the query, and $\mathrm{content}(Q)$ a fixed instruction. An answer node

$$a_r = \big(\mathrm{id}(a_r), \mathrm{name}(a_r), \mathrm{content}(a_r)\big), \tag{4}$$

uses a unique unit ID, the fixed name "Answer Unit", and its canonical semantics as content. The edge set is

$$\mathcal{E}_r = \{(Q, d_i), (d_i, a_r) \mid d_i \in \mathcal{D}_r\}, \tag{5}$$

encodes evidence selection through query-to-document edges and evidential support through document-to-answer edges. A complete subgraph example is provided below:

Example graph 1:

```
Nodes: id:Q name: Who wrote The Sound of Silence? content: Reasoning the graph with node A
based on the node starting with d.
  id:d1 name: The Sound of Silence content: "The Sound of Silence" is a song by Simon &
Garfunkel. It was written by Paul Simon.
  id:A1 name: Answer Unit content: Paul Simon wrote "The Sound of Silence".
Edges: nodes:Q inputs: outputs: d1
  nodes:d1 inputs: Q outputs: A1
  nodes:A1 inputs: d1 outputs: N/A
subgraph_description: The document supports the answer by stating that "The Sound of Silence"
was written by Paul Simon.
```

Attribution-space contraction. The graph induces the retained document set $\mathcal{D}^+ = \bigcup_{r=1}^{m} \mathcal{D}r$, with $n = |\mathcal{D}^+|$, and specifies the attribution target

$$\alpha_{\mathbf{G}}(a_r) = \{\mathrm{id}(d_i) \mid d_i \in \mathcal{D}_r\}. \tag{6}$$

Thus, if a generated claim $s\pi(r)$ semantically realizes a_r, its target citation set is $c^*_{\pi(r)} = \alpha_{\mathbf{G}}(a_r)$. Documents outside $\mathcal{D}^+$ are excluded from generation; if $m = 0$, CAGE abstains.

Without an attribution map, each of the m claims may select any non-empty subset of the n retained documents, yielding up to

$$|\Omega_{\mathrm{unstr}}| = (2^n - 1)^m. \tag{7}$$

Once induced, the attribution map $\mathbf{G}$ fixes one citation set for each answer node:

$$\Omega_{\mathbf{G}} = \big\{\big(\alpha_{\mathbf{G}}(a_1), \ldots, \alpha_{\mathbf{G}}(a_m)\big)\big\}, \qquad |\Omega_{\mathbf{G}}| = 1. \tag{8}$$

Thus, conditioned on the induced map, CAGE reduces the downstream citation-assignment space from up to $(2^n - 1)^m$ unconstrained choices to one graph-specified target, while attribution is resolved during map induction.

3 Methodology

This section presents CAGE, a two-stage framework for citation-grounded long-form answer generation. *Validated Cognitive Attribution Synthesis* (VCAS) constructs validated cognitive attribution map and map-aligned answers, which supervise a *Cognitive Attribution Map Induction Model* (CMI) for constructing answer-centered support subgraphs and a Structured Citation Reasoning Model (SCR) for map-conditioned answer generation.

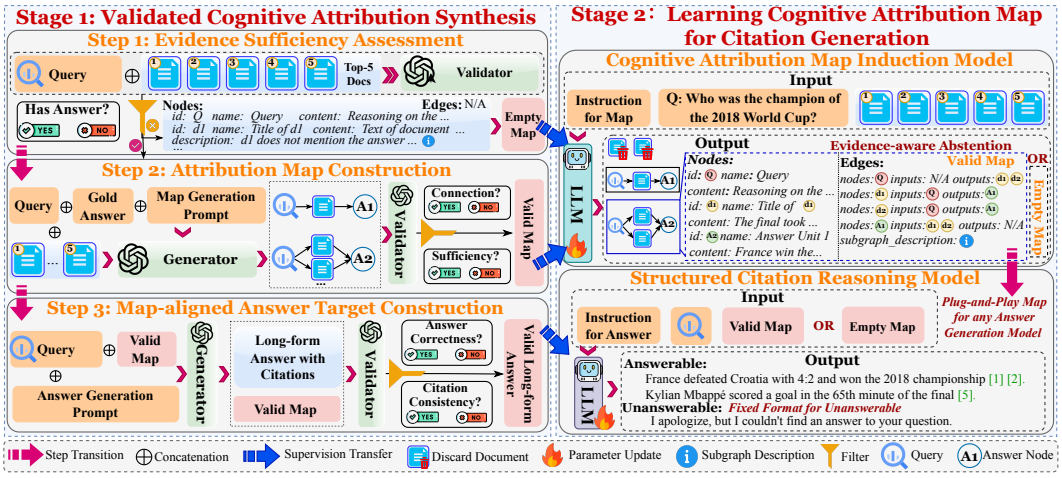


Figure 2: Overview of CAGE. Stage 1 Validated Cognitive Attribution Synthesis constructs validated cognitive attribution map and map-aligned answer supervision, which are used to train the Cognitive Attribution Map Induction Model and the Structured Citation Reasoning Model, respectively, in Stage 2.

3.1 Validated Cognitive Attribution Synthesis (VCAS)

This section introduces how VCAS constructs supervision through two LLM-instantiated roles: a generator and a validator. As shown in Figure 2, VCAS converts query–document instances into two aligned supervision sets: one for CMI and the other for SCR.

Since LLMs generate token sequences, a non-empty cognitive attribution map $\mathbf{G} = \{\mathcal{G}r\}_{r=1}^{m}$ is represented by a canonical textual encoding:

$$z_{\text{map}} = \text{Enc}_{\text{map}}(\mathbf{G}) = \text{Enc}_V(\mathcal{V}(\mathbf{G})) \oplus \text{Enc}_E(\mathcal{E}(\mathbf{G})) \oplus \text{Enc}_\rho(\boldsymbol{\rho}(\mathbf{G})), \tag{9}$$

where $\mathcal{V}(\mathbf{G})$, $\mathcal{E}(\mathbf{G})$, and $\boldsymbol{\rho}(\mathbf{G})$ denote its nodes, edges, and descriptions, respectively. The encoding preserves the map structure while serving as an autoregressive prediction target.

Step 1: Evidence Sufficiency Assessment. Given a query Q and its top-k retrieved documents $\mathcal{D}_k = \{d_1, \ldots, d_k\}$, the validator determines whether the retrieved documents contain enough evidence to support at least one query-responsive answer unit:

$$\exists a,\ \exists \mathcal{D}' \subseteq \mathcal{D}_k, \qquad \mathcal{D}' \neq \emptyset, \qquad \mathcal{D}' \models_Q a. \tag{10}$$

Here, $\mathcal{D}' \models_Q a$ means that $\mathcal{D}'$ provides sufficient evidence for a under Q.

Unsupported instances are assigned a document preserving empty map:

$$\mathbf{G}^*_\emptyset = (\mathcal{V}^*_\emptyset, \mathcal{E}^*_\emptyset; \rho^*_\emptyset), \mathcal{V}^*_\emptyset = d_i, \quad \mathcal{E}^*_\emptyset = \emptyset, \quad z^*_{\text{map},\emptyset} = \text{Enc}_{\text{map}}(\mathbf{G}^*_\emptyset). \tag{11}$$

Here, $\rho^*_\emptyset$ explains the evidence insufficiency. The empty map retains all candidate document nodes but contains no answer nodes, selected supporting-document assignments, or support edges. These instances train CMI to identify unsupported questions without constructing spurious answer units. The seed questions are sampled from the training sets of FRONT Huang et al. (2024a) and ASQA Stelmakh et al. (2022) and the number of documents was set to 5, consistent with prior study Song et al. (2025).

Step 2: Attribution Map Construction. For each supported instance, the generator $\mathcal{M}_{\text{gen}}$ receives $(Q, \mathcal{D}_k, \mathcal{P}_{\text{map}})$ and generates a candidate cognitive attribution map through its canonical textual encoding:

$$\overline{z}_{\text{map}} = \mathcal{M}_{\text{gen}}(Q, \mathcal{D}_k, \mathcal{P}_{\text{map}}) = \text{Enc}_{\text{map}}(\overline{\mathbf{G}}), \qquad \overline{\mathbf{G}} = \{\overline{\mathcal{G}}_r\}_{r=1}^{\bar{m}}. \tag{12}$$

where $\mathcal{P}_{\text{map}}$ is a specific prompt for map generation.

The validator checks each candidate subgraph for structural connectivity and evidential support:

$$\mathbb{V}_Q(\overline{\mathcal{G}}_r) = \mathbb{I}\left[\text{Connected}(\overline{\mathcal{G}}_r) \wedge \overline{\mathcal{D}}_r \models_Q \overline{a}_r\right]. \tag{13}$$

Here, $\text{Connected}(\overline{\mathcal{G}}_r)$ requires $\overline{\mathcal{D}}_r \neq \emptyset$ and every document node to lie on a complete support path from the query node to the semantic answer node:

$$\forall d_i \in \overline{\mathcal{D}}_r, \qquad (Q, d_i) \in \overline{\mathcal{E}}_r \ \wedge \ (d_i, \overline{a}_r) \in \overline{\mathcal{E}}_r. \tag{14}$$

Thus, all query, document, and answer nodes participate in the support structure, with no isolated nodes. The relation $\overline{\mathcal{D}}_r \models_Q \overline{a}_r$ indicates that the assigned documents collectively provide sufficient evidence for the semantic answer node.

The accepted subgraphs are assembled into the validated cognitive attribution map:

$$\mathbf{G}^* = \{\overline{\mathcal{G}}_r \in \overline{\mathbf{G}} \mid \mathbb{V}_Q(\overline{\mathcal{G}}_r) = 1\}. \tag{15}$$

A supported instance is discarded if $\mathbf{G}^* = \emptyset$. Otherwise, the validated map is encoded as the CMI target:

$$z^*_{\text{map}} = \text{Enc}_{\text{map}}(\mathbf{G}^*). \tag{16}$$

Together with unsupported instances encoded as document preserving empty maps, these targets form the CMI supervision set $\mathcal{S}_{\text{CMI}}$, which contains 7.8K samples.

Step 3: Map-aligned Answer Target Construction. Since each retained document node contains its document ID, title, and textual content, the encoded map provides both textual evidence and citation identifiers for answer generation. The generator receives $(Q, z^*_{\text{map}}, \mathcal{P}_{\text{ans}})$ and constructs a map-aligned long-form answer with inline citations:

$$\mathcal{Y}^* = \{(s^*_j, c^*_j)\}_{j=1}^m = \mathcal{M}_{\text{ans}}(Q, z^*_{\text{map}}, \mathcal{P}_{\text{ans}}), \tag{17}$$

where s^*_j is a sentence-level claim and c^*_j is its citation set, $\mathcal{P}_{\text{ans}}$ is a prompt for map-aligned answer generation.

The answer target must satisfy a one-to-one alignment between graph nodes and generated claims. Since the final answer order need not follow the graph order, correctness is defined by the existence of a permutation π such that

$$s^*_{\pi(r)} \equiv a^*_r, c^*_{\pi(r)} = \alpha_{\mathbf{G}^*}(a^*_r), \forall r \in \{1, \ldots, m\}. \tag{18}$$

Thus, every generated claim must realize exactly one validated answer node, and its citation set must match the citation target specified by the corresponding answer-centered support subgraph.

The validator retains only answers that are correct with respect to the validated answer nodes and whose citation assignments are consistent with the graph-specified mapping. The accepted outputs form $\mathcal{S}_{\text{SCR}}$, containing 2.3K instances. The prompt of $\mathcal{P}_{\text{map}}$ and $\mathcal{P}_{\text{ans}}$ can be found in Appendix.

3.2 Learning Cognitive Attribution Map for Citation Generation

CAGE factorizes citation-grounded generation through the textual encoding of a cognitive attribution map:

$$\begin{aligned} &p_{\theta_g,\theta_a}(z_{\text{map}}, \mathcal{Y} \mid Q, \mathcal{D}_k, \mathcal{P}_{\text{map}}, \mathcal{P}_{\text{ans}}) \\ &\quad = p_{\theta_g}(z_{\text{map}} \mid Q, \mathcal{D}_k, \mathcal{P}_{\text{map}})\, p_{\theta_a}(\mathcal{Y} \mid Q, z_{\text{map}}, \mathcal{P}_{\text{ans}}). \end{aligned} \tag{19}$$

Inference follows the same two-stage decomposition:

$$\hat{z}_{\text{map}} = \arg\max_z p_{\theta_g}(z \mid Q, \mathcal{D}_k, \mathcal{P}_{\text{map}}), \hat{\mathcal{Y}} = \arg\max_{\mathcal{Y}} p_{\theta_a}(\mathcal{Y} \mid Q, \hat{z}_{\text{map}}, \mathcal{P}_{\text{ans}}). \tag{20}$$

3.2.1 Cognitive Attribution Map Induction Model

CMI implements the first stage of CAGE by learning to generate the textual encoding of a cognitive attribution map. Given a query Q and its top-k retrieved documents $\mathcal{D}_k$, CMI predicts

$$\hat{z}_{\text{map}} = f_{\theta_g}(Q, \mathcal{D}_k, \mathcal{I}_{\text{map}}). \tag{21}$$

SCR Model	Method	ASQA(610 answerable, 338 unanswerable)				ExpertQA(682 answerable, 1487 unanswerable)				ELI5(207 answerable, 793 unanswerable)				Average
		EM^{Fl}_{AC}	$F1_{GR}$	$F1_{GC}$	**TRUST**	EM^{Fl}_{AC}	$F1_{GR}$	$F1_{GC}$	**TRUST**	EM^{Fl}_{AC}	$F1_{GR}$	$F1_{GC}$	**TRUST**	
Qwen-2.5 -0.5B	ICL†	20.96	47.19	0.34	22.83	21.42	38.71	0.44	20.19	13.73	33.14	0.37	15.75	19.59
	PostCite†	8.55	50.84	8.23	22.54	13.32	48.08	5.60	22.33	9.87	27.10	4.10	13.69	19.52
	PostAttr	8.55	50.84	2.23	20.54	13.32	48.08	1.49	20.96	9.87	27.10	0.68	12.55	18.02
	FRONT†	42.83	39.15	45.87	42.62	18.27	24.05	34.62	25.65	13.74	17.29	27.95	19.66	29.31
	T-ALIGN†	50.59	61.28	52.40	54.76	18.16	63.31	35.07	38.85	13.68	60.79	22.72	32.40	42.00
	CAGE	**56.90**	**69.89**	**70.49**	**65.76**	**26.59**	**71.45**	**64.13**	**54.06**	**21.91**	**61.30**	**42.73**	**38.65**	**52.82**
	Δ	6.31↑	8.61↑	18.09↑	11.00↑	5.17↑	8.14↑	29.06↑	15.21↑	8.17↑	0.51↑	14.78↑	6.25↑	10.82↑
Qwen-2.5 -1.5B	ICL†	50.55	41.74	6.69	32.99	30.67	26.09	6.89	21.22	20.56	17.78	4.99	14.44	22.88
	PostCite†	16.36	52.46	15.40	28.07	22.22	48.66	16.92	29.27	15.63	26.71	5.17	15.84	24.39
	PostAttr†	16.36	52.46	4.45	24.42	22.22	48.66	13.15	28.01	15.63	26.71	0.62	14.32	22.25
	FRONT†	57.74	41.36	55.70	51.60	29.15	24.60	50.22	34.66	19.57	17.29	37.70	24.85	37.04
	T-ALIGN†	52.68	62.38	66.81	60.62	25.06	68.38	51.44	48.29	19.03	57.91	31.63	36.19	48.37
	CAGE	**59.27**	**69.81**	**80.67**	**69.91**	**32.05**	**71.68**	**77.34**	**60.35**	**28.52**	**61.47**	**43.14**	**44.38**	**58.21**
	Δ	1.53↑	7.43↑	13.86↑	9.29↑	2.90↑	3.30↑	25.90↑	11.37↑	8.95↑	3.56↑	5.44↑	8.19↑	9.84↑
Qwen-2.5 -3B	ICL†	37.72	51.36	51.72	46.93	35.14	49.65	42.67	42.49	29.12	46.31	34.34	36.59	42.00
	PostCite†	9.58	35.30	10.94	18.61	0	40.66	0	13.55	21.73	48.49	7.56	25.93	19.36
	PostAttr†	9.58	35.30	36.29	27.06	0	40.66	0	13.55	21.73	48.49	1.31	23.84	21.48
	FRONT†	55.15	44.01	62.72	53.96	25.67	29.86	44.48	33.34	18.69	25.37	37.40	27.15	38.15
	T-ALIGN†	55.19	63.76	78.64	65.86	20.97	65.79	60.25	49.0	22.52	**64.38**	42.01	42.97	52.61
	CAGE	**59.98**	**70.08**	**83.09**	**71.05**	**33.68**	**71.75**	**78.39**	**61.27**	**30.69**	61.63	**48.55**	**46.96**	**59.76**
	Δ	4.79↑	6.32↑	4.45↑	5.19↑	0.54↑	5.96↑	18.14↑	12.27↑	1.57↑	-2.75↓	6.54↑	3.99↑	7.15↑
Qwen-2.5 -7B	ICL†	58.94	54.34	75.46	62.91	**36.33**	42.28	56.09	44.9	28.27	37.13	44.13	36.51	48.11
	PostCite†	27.52	45.93	4.19	25.88	25.58	54.9	13.77	31.42	21.82	22.23	7.03	17.03	24.78
	PostAttr†	27.52	45.93	17.92	30.46	25.58	54.9	12.46	30.98	21.82	22.23	0.96	15.00	25.48
	FRONT†	**64.58**	60.08	58.27	60.98	32.41	55.56	67.35	51.77	28.27	54.14	56.61	46.34	53.03
	T-ALIGN†	55.04	66.22	83.57	68.28	25.57	69.16	62.7	52.48	24.30	**63.79**	47.02	45.04	55.27
	CAGE	60.98	**70.09**	**85.32**	**72.12**	33.14	**71.61**	**80.04**	**61.60**	**31.32**	61.31	**53.44**	**48.70**	**60.81**
	Δ	-3.60↓	3.87↑	1.75↑	3.84↑	-3.19↓	2.45↑	12.69↑	9.12↑	3.05↑	-2.48↓	6.42↑	2.36↑	5.54↑
Qwen3.5-9B	CAGE	60.21	70.08	87.87	**72.72**	33.74	71.70	80.95	**62.13**	29.83	61.55	57.79	**49.72**	**61.52**

Table 1: Main results on ASQA, ExpertQA, and ELI5. Qwen3.5-9B is fixed as CMI and Qwen-2.5 models serve as SCR, except the last row, which uses separate Qwen3.5-9B models for both stages. The answerable setting and † results follow Song et al. (2025). Δ is the gain over the strongest baseline; **Average** is mean **TRUST**. LLaMA results are in the Appendix A.2.1.

Here, $\hat{z}_{\text{map}}$ denotes the generated textual encoding of the cognitive attribution map, $\mathcal{I}_{\text{map}}$ is a task specific instruction for CMI. When well formed, it specifies an induced graph-structured cognitive attribution map $\hat{\mathbf{G}}$. If the generated encoding corresponds to an empty map, CMI indicates that no answer unit can be supported by the retrieved documents. Otherwise, the encoded map specifies answer-centered support subgraphs following the schema in Eq. 3, including answer nodes, supporting document nodes, support edges, and descriptions.

Thus, CMI learns to decide what can be supported, abstract supported answer units, and assign document evidence to these units before any final answer sentence is decoded. The graph-structured map remains the intermediate reasoning object, while its textual encoding is the sequence generated and optimized by the LLM.

CMI is trained to generate the target map encoding z^*_{map} constructed by VCAS in Eq. 16. For non-empty maps, this encoding preserves the typed nodes, support edges, and support descriptions of $\mathbf{G}^*$. For empty maps, it produces a schema-preserving empty-map encoding that contains no evidence-supported answer nodes, supporting document nodes, or support edges, but records why the retrieved documents fail to support the question.

Let $X_g = (Q, \mathcal{D}_k, \mathcal{I}_{\text{map}})$, CMI is optimized by autoregressive maximum likelihood:

$$\mathcal{L}_{\text{CMI}} = -\mathbb{E}_{\mathcal{S}_{\text{CMI}}} \sum_{t=1}^{|z^*_{\text{map}}|} \log P_{\theta_g}\left(z^*_{\text{map},t} \mid z^*_{\text{map},<t}, X_g\right). \tag{22}$$

where the expectation is taken over $(X_g, z^*_{\text{map}}) \sim \mathcal{S}_{\text{CMI}}$.

At inference time, CMI generates the map encoding $\hat{z}_{\text{map}}$ and passes it to SCR. If the encoding specifies an empty map, it preserves the attribution-map schema and records why the retrieved documents fail to support the question, but contains no concrete evidence-supported answer nodes, supporting

document nodes, or support edges. Otherwise, it provides the encoded support subgraphs, retained document content, and map-specified citation targets for citation reasoning.

3.2.2 STRUCTURED CITATION REASONING MODEL

SCR implements the second stage of CAGE by reasoning over the encoded cognitive attribution map and realizing the underlying map as a citation-grounded long-form answer. Given the Q and z_{map}, SCR predicts the final answer:

$$\hat{\mathcal{Y}} = f_{\theta_a}\left(Q, z_{\text{map},}, \mathcal{I}_{\text{ans}}\right). \tag{23}$$

Here, $\mathcal{I}_{\text{ans}}$ is an instruction for SCR and $\hat{\mathcal{Y}} = \{(\hat{s}_j, \hat{c}_j)\}_{j=1}^{m}$, where $\hat{s}_j$ is a generated claim and $\hat{c}_j$ is its citation set.

SCR performs three map-aligned operations: realizing answer nodes as sentence-level claims, aligning each claim with citation IDs specified by the support-subgraph document assignments, and organizing the cited claims into a coherent long-form answer. Since the output order need not follow the support-subgraph order, correctness is defined by a permutation π satisfying

$$s_{\pi(r)} \equiv a_r, \quad c_{\pi(r)} = \alpha_{\mathbf{G}}(a_r), \quad \forall r \in \{1, \ldots, m\}, \tag{24}$$

where $\alpha_{\mathbf{G}}(a_r)$ denotes the citation identifiers of the document nodes assigned to answer node a_r in $\mathbf{G}$.

For autoregressive training, the map-aligned answer itself serves as the target sequence, written as claim–citation units:

$$\mathcal{Y} = \oplus_{j=1}^{m}[s_j; c_j]. \tag{25}$$

Let $X_a = (Q, z_{\text{map}}, \mathcal{I}_{\text{ans}})$, SCR is optimized by autoregressive maximum likelihood over the final answer sequence:

$$\mathcal{L}_{\text{SCR}} = -\mathbb{E}_{(X_a, \mathcal{Y}) \sim \mathcal{S}_{\text{SCR}}} \sum_{t=1}^{|\mathcal{Y}|} \log P_{\theta_a}\left(\mathcal{Y}_t \mid \mathcal{Y}_{<t}, X_a\right). \tag{26}$$

This objective trains SCR to generate cited claims aligned with the map-specified document IDs encoded in z_{map}.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Datasets. We evaluate our method on three representative long-form QA benchmarks for inline citation generation: ASQA Stelmakh et al. (2022), ELI5 Fan et al. (2019), and ExpertQA Malaviya et al. (2024).

Baselines. We compare CAGE with six representative baselines: *ICL*, *PostCite*, *PostAttr* Gao et al. (2023), *FRONT* Huang et al. (2024a), *T-ALIGN* Song et al. (2025), and *Ground-GRPO* Sim et al. (2025), covering in-context generation, post-hoc attribution, preference learning, grounded answering, and GRPO-based alignment. We exclude START Huang et al. (2024b), LongCite Zhang et al. (2025), and SelfCite Chuang et al. (2025) due to differences in reproducibility, task formulation, and evaluation protocol.

LLMs. We evaluate CAGE with `Qwen-2.5`, `Qwen-3` series and `Qwen-3.5` as SCR models. In CMI, we use Qwen3.5-9B by default. We further compare CAGE with `GPT-5.5` and `Claude-Sonnet-4.6`. The implementation details can be found in Appendix A.1.

Metrics. We adopt TRUST Song et al. (2025) as the overall metric, which averages answer calibration $\text{EM}_{\text{AC}}^{F1}$, refusal behavior F1_{RG}, and citation groundedness F1_{CG}. Here, $\text{EM}_{\text{AC}}^{F1}$ measures correctness on answerable questions while accounting for the model's decision to answer, F1_{RG} evaluates answering/refusal decisions over answerable and unanswerable questions, and F1_{CG} measures citation precision and recall using the TRUE model following prior work Gao et al. (2023); Song et al. (2025). The final score is:

$$\textbf{TRUST} = \frac{1}{3}\left(\text{EM}_{\text{AC}}^{F1} + \text{F1}_{\text{RG}} + \text{F1}_{\text{CG}}\right). \tag{27}$$

Here, the details of metric can be found in the Appendix A.7.

4.2 Experimental Results and Analysis

4.2.1 Main Results

The experimental results are displayed in Table 1 and the experimental analyses are listed as follows:

(1) CAGE consistently achieves the best TRUST scores across Qwen-2.5 SCR scales, improving over the strongest baseline by 10.82%, 9.84%, 7.15%, and 5.54% from 0.5B to 7B. The largest gain appears in attribution grounding, with $F1_{GC}$ improving by up to 29.06% on ExpertQA. These results support that graph-specified evidence paths reduce attribution ambiguity by contracting the attribution space.

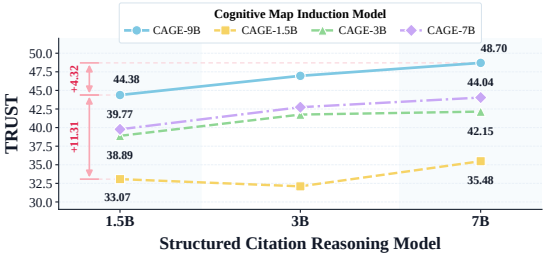


Figure 3: Effect of model scale on TRUST performance. CAGE-9B uses Qwen3.5-9B in the CMI stage, while CAGE-1.5B/3B/7B use Qwen-2.5 models of the corresponding sizes for map induction.

(2) Across model–dataset settings, CAGE delivers consistent gains in answer calibration, refusal behavior, and citation grounding, with particularly strong improvements in $F1_{GC}$, reaching 70.5% on ASQA with Qwen-2.5-0.5B and 78.4% on ExpertQA with Qwen-2.5-3B. These results suggest that the graph-guided intermediate representation better organizes evidence for attribution without compromising overall response quality.

(3) Finally, Table 1 shows that scaling the SCR model yields only marginal gains once a high-quality graph is available. Qwen3.5-9B achieves the highest average TRUST score of 61.52%, but improves over the Qwen-2.5-7B SCR model by only 0.71%. This suggests that reliable support graphs, rather than model scaling, are the primary driver of evidence alignment and attribution quality.

4.2.2 Effect of Model Scale in CAGE

As shown in Figure 3, scaling the SCR model improves TRUST, but increasing map induction capacity yields larger gains. With CAGE-9B as the induction model, scaling SCR from 1.5B to 7B improves TRUST by 4.32%. In contrast, replacing same-size induction models with CAGE-9B yields gains of 11.31%, 5.21%, and 4.66% for 1.5B, 3B, and 7B SCR models, respectively. This suggests that TRUST is more constrained by map induction quality, as stronger induction better establishes evidence–answer alignments and contracts the attribution space before generation.

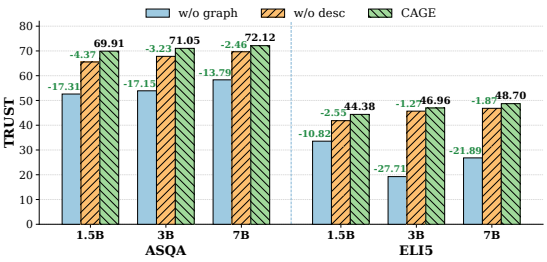


Figure 4: Ablation study. We fix the structured citation reasoning model size (1.5B/3B/7B) and ablate the map induction model, including "*w/o graph*" (no structure, documents retained) and "*w/o desc*" (no graph descriptions).

4.2.3 Ablation Study

We conduct ablation experiments for the CMI stage, as reported in Figure 4. Specifically, *w/o graph* the explicit support edges while retaining the same selected documents and subgraph descriptions, and *w/o desc* removes the textual descriptions associated in the subgraph.

(1) In the *w/o graph* setting, when holding evidence selection and the textual answer plan fixed. TRUST drops from 69.91% to 52.60% on ASQA with Qwen-2.5-1.5B, a 17.31% decrease. This suggests that explicit claim–document assignments contribute beyond the selected evidence and textual descriptions.

(2) In the *w/o desc* setting, performance also declines, but less severely than removing the graph structure; for example, TRUST on ASQA with Qwen-2.5-1.5B drops from 69.91% to 65.54%. This suggests that textual descriptions help the generator interpret the semantic role of each support unit, while the main attribution-space contraction still comes from the graph structure.

4.2.4 Analysis of CMI under Different SCR Backbones

Table 2 evaluates CAGE from two perspectives. (1) CAGE outperforms strong end-to-end baselines: on ASQA, it improves TRUST over Claude-4.6 and Ground-GRPO by 7.21% and 6.94%, respectively, supporting the benefit of attribution-space contraction beyond stronger backbones or preference optimization alone. (2) CMI also shows plug-and-play generality across heterogeneous SCR backbones, improving both Claude and GPT models without additional training; on ExpertQA, CAGE with Claude-4.6 raises TRUST from 55.33% to 66.43%. These results demonstrate CMI's value as a reusable attribution module across model families and scales.

Method	ASQA				ExpertQA			
	$\mathrm{EM}^{\mathrm{F1}}_{\mathrm{AC}}$	$\mathrm{F1}_{\mathrm{RG}}$	$\mathrm{F1}_{\mathrm{CG}}$	**TRUST**	$\mathrm{EM}^{\mathrm{F1}}_{\mathrm{AC}}$	$\mathrm{F1}_{\mathrm{RG}}$	$\mathrm{F1}_{\mathrm{CG}}$	**TRUST**
Ground-GRPO$_{4B}$	47.22	62.88	81.94	64.01	20.56	67.14	59.37	49.02
CMI$_{9B}$+SCR$_{Qwen3\text{-}4B}$	59.50	66.54	86.82	$70.95_{6.94\uparrow}$	36.27	62.22	78.81	$62.41_{13.39\uparrow}$
GPT-5.5	66.18	61.54	70.69	66.14	46.74	54.46	49.78	50.33
CMI$_{9B}$+SCR$_{GPT\text{-}5.5}$	61.41	70.09	92.32	$74.61_{8.47\uparrow}$	32.88	71.74	89.15	$64.59_{14.26\uparrow}$
Claude-4.6	70.46	64.56	73.22	69.41	47.66	64.54	53.80	55.33
CMI$_{9B}$+SCR$_{Claude\text{-}4.6}$	65.29	70.09	94.48	$76.62_{7.21\uparrow}$	36.70	71.74	90.86	$66.43_{11.10\uparrow}$

Table 2: Backbone capability and proprietary model comparison on ASQA and ExpertQA. 9B denotes Qwen3.5-9B. We additionally fine-tune Qwen3-4B for the SCR stage to enable a direct comparison with Ground-GRPO.

5 Related Work

Existing methods for inline citation generation can be categorized into parameter-updating and prompt-based approaches. *Parameter-updating* methods train models on citation-oriented supervision. Early work constructs statement–document attribution data with NLI models or advanced LLMs and improves citation quality through supervised fine-tuning or preference optimization (Ye et al., 2024; Huang et al., 2024a; Li et al., 2024). Later studies explore self-improving data synthesis (Huang et al., 2024b), coarse-to-fine construction for long-context generation (Zhang et al., 2025), context-ablation-based preference optimization (Chuang et al., 2025), and grounding from parametric or external knowledge (Khalifa et al., 2024; Shen et al., 2025).

Prompt-based methods keep the backbone LLM frozen and use prompting or decomposition to elicit citation behavior. ALCE (Gao et al., 2023) prompts LLMs with retrieved documents, while later work adopts chain-of-thought, search-based expansion, or query/document selection (Ji et al., 2024; Li & Ng, 2025; Hirsch et al., 2025). Attribute First, then Generate (Slobodkin et al., 2024) organizes selected evidence spans into sentence-level plans before generation. In contrast, CAGE models attribution as an explicit intermediate reasoning structure by inducing validated answer–document support graphs, enabling generation to follow structured attribution constraints rather than evidence planning alone.

6 Conclusion and Limitations

In this paper, We propose CAGE, a two-stage framework that models answer–evidence attribution through a cognitive attribution map before generation, reducing attribution ambiguity and improving claim-level citation reliability. Experiments on ASQA, ELI5, and ExpertQA show consistent gains across open-source and proprietary LLMs, demonstrating its plug-and-play utility without generator-specific retraining or architectural modifications. Despite these gains, map induction fails to preserve supporting evidence for 15.42% of instances on average, limiting downstream recovery. Future work will improve map induction and reduce cross-stage error propagation.

REFERENCES

Bernd Bohnet, Vinh Q Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, et al. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint arXiv:2212.08037*, 2022.

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. Complex claim verification with evidence retrieved in the wild. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Association for Computational Linguistics, 2024.

Yung-Sung Chuang, Benjamin Cohen-Wang, Zejiang Shen, Zhaofeng Wu, Hu Xu, Xi Victoria Lin, James R Glass, Shang-Wen Li, and Wen-Tau Yih. Selfcite: Self-supervised alignment for context attribution in large language models. In *International Conference on Machine Learning*, pp. 10839–10858. PMLR, 2025.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3558–3567, Florence, Italy, 2019. Association for Computational Linguistics.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, N. Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. Rarr: Researching and revising what language models say, using language models. *ArXiv*, abs/2210.08726, 2022. URL https://arxiv.org/abs/2210.08726.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488. Association for Computational Linguistics, 2023.

Eran Hirsch, Aviv Slobodkin, David Wan, Elias Stengel-Eskin, Mohit Bansal, and Ido Dagan. Laquer: Localized attribution queries in content-grounded generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15355–15370, 2025.

Lei Huang, Xiaocheng Feng, Weitao Ma, Yuxuan Gu, Weihong Zhong, Xiachong Feng, Weijiang Yu, Weihua Peng, Duyu Tang, Dandan Tu, et al. Learning fine-grained grounded citations for attributed large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 14095–14113, 2024a.

Lei Huang, Xiaocheng Feng, Weitao Ma, Liang Zhao, Yuchun Fan, Weihong Zhong, Dongliang Xu, Qing Yang, Hongtao Liu, and Bing Qin. Advancing large language model attribution through self-improving. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 3822–3836, 2024b.

Bin Ji, Huijun Liu, Mingzhe Du, and See-Kiong Ng. Chain-of-thought improves text generation with citations in large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):18345–18353, 2024.

Muhammad Khalifa, David Wadden, Emma Strubell, Honglak Lee, Lu Wang, Iz Beltagy, and Hao Peng. Source-aware training enables knowledge attribution in language models. In *First Conference on Language Modeling*, 2024.

Dongfang Li, Zetian Sun, Baotian Hu, Zhenyu Liu, Xinshuo Hu, Xuebo Liu, and Min Zhang. Improving attributed text generation of large language models via preference learning, 2024. URL https://arxiv.org/abs/2403.18381.

Junyi Li and Hwee Tou Ng. Think&cite: Improving attributed text generation with self-guided tree search and progress reward modeling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9928–9942, 2025.

Junyu Luo, Weizhi Zhang, Ye Yuan, Yusheng Zhao, Junwei Yang, Yiyang Gu, Bohan Wu, Binqi Chen, Ziyue Qiao, Qingqing Long, et al. Large language model agent: A survey on methodology, applications and challenges. *arXiv preprint arXiv:2503.21460*, 2025.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. ExpertQA: Expert-curated questions and attributed answers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3025–3045, Mexico City, Mexico, 2024. Association for Computational Linguistics.

Shirley Mark, Rani Moran, Thomas Parr, Steve W Kennerley, and Timothy EJ Behrens. Transferring structural knowledge across cognitive maps in humans and models. *Nature communications*, 11 (1):4783, 2020.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100. Association for Computational Linguistics, 2023.

Jeff Z. Pan, Simon Razniewski, Jan-Christoph Kalo, Sneha Singhania, Jiaoyan Chen, Stefan Dietze, Hajira Jabeen, Janna Omeliyanenko, Wen Zhang, Matteo Lissandrini, et al. Large language models and knowledge graphs: Opportunities and challenges. *Transactions on Graph Data and Knowledge*, 2023.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599, 2024. doi: 10.1109/TKDE.2024.3352100.

Pin Qian, Su Wang, Xiaoyuan Wang, Yihang Chen, Wenxuan Xu, Qiaolin Yu, Shuhuai Lin, Sipeng Zhang, Junxian You, and Xinpeng Wei. Relevant is not warranted: Evidence-force calibration for cited rag. *arXiv preprint arXiv:2605.28044*, 2026.

Jiajun Shen, Tong Zhou, Yubo Chen, Delai Qiu, Shengping Liu, Kang Liu, and Jun Zhao. Transparentize the internal and external knowledge utilization in llms with trustworthy citation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 17858–17877, 2025.

Shang Hong Sim, Tej Deep Pala, Vernon Toh, Hai Leong Chieu, Amir Zadeh, Chuan Li, Navonil Majumder, and Soujanya Poria. Lessons from training grounded llms with verifiable rewards, 2025. URL https://arxiv.org/abs/2506.15522.

Aviv Slobodkin, Eran Hirsch, Arie Cattan, Tal Schuster, and Ido Dagan. Attribute first, then generate: Locally-attributable grounded text generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3309–3344, 2024.

Maojia Song, Shang Hong Sim, Rishabh Bhardwaj, Hai Leong Chieu, Navonil Majumder, and Soujanya Poria. Measuring and enhancing trustworthiness of LLMs in RAG through grounded attributions and learning to refuse. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Iyrtb9EJBp.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. ASQA: Factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 8273–8288, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics.

Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. Redeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. In *International Conference on Learning Representations*, volume 2025, pp. 50250–50279, 2025.

Liwei Tan, Yidan Qiu, Lixin Qiu, Shuting Lin, Jinhui Li, Jiajun Liao, Yuting Zhang, Wei Zou, and Ruiwang Huang. The medial and lateral orbitofrontal cortex jointly represent the cognitive map of task space. *Communications Biology*, 8(1):163, 2025.

James CR Whittington, David McCaffary, Jacob JW Bakermans, and Timothy EJ Behrens. How to build a cognitive map. *Nature neuroscience*, 25(10):1257–1272, 2022.

Zhichao Yan, Jiapu Wang, Jiaoyan Chen, Xiaoli Li, Jiye Liang, Ru Li, and Jeff Z Pan. Atomic fact decomposition helps attributed question answering. *IEEE Transactions on Knowledge and Data Engineering*, 2025.

Xi Ye, Ruoxi Sun, Sercan Arik, and Tomas Pfister. Effective large language model adaptation for improved grounding and citation generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6237–6251, 2024.

Jiajie Zhang, Yushi Bai, Xin Lv, Wanjun Gu, Danqing Liu, Minhao Zou, Shulin Cao, Lei Hou, Yuxiao Dong, Ling Feng, et al. Longcite: Enabling llms to generate fine-grained citations in long-context qa. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5098–5122, 2025.

Jiajie Zhang, Xin Lv, Ling Feng, Lei Hou, and Juanzi Li. Chaining the evidence: Robust reinforcement learning for deep search agents with citation-aware rubric rewards. *arXiv preprint arXiv:2601.06021*, 2026.

A Appendix

AI Assistance Statement. In this paper, AI tools were used only for language polishing and the generation of training data.

A.1 Implementation details.

For the Validated Cognitive Attribution Synthesis stage, we use GPT-5.5 as both the validator and generator. We implement the training of the Cognitive Attribution Map Induction Model and Structured Citation Reasoning Model using LLaMA-Factory. Both components are optimized via full-parameter supervised fine-tuning with DeepSpeed ZeRO-3 and BF16 precision on 4 NVIDIA A800-80G GPUs. The graph induction model uses a per-device batch size of 6, while the citation reasoning model uses a per-device batch size of 8 with gradient accumulation of 2. We train both models for three epochs using AdamW with a learning rate of 1×10^{-5}, cosine learning-rate scheduling, and a warmup ratio of 0.1. The random seed is set to 42, and weight decay is set to zero for all experiments.

A.2 More Experimental results

A.2.1 Experimental results on LLaMA Families

Table 3 demonstrates that CAGE exhibits strong generalization across the LLaMA family. Despite substantial differences in the capacity of the SCR models, CAGE consistently achieves the highest overall TRUST performance on ASQA, ExpertQA, and ELI5, indicating that its effectiveness is not restricted to a particular model scale. Specifically, CAGE obtains average TRUST scores of 58.60%, 54.24%, 60.90%, and 59.36% with the 2–7B, 3.2–1B, 3.2–3B, and 3–8B models, respectively, outperforming the strongest baselines by 4.04%, 4.83%, 5.97%, and 3.04%. Moreover, CAGE achieves the best TRUST score across all 12 model–dataset combinations. Its TRUST scores remain consistently high on ASQA, ranging from 69.16 to 72.68, while reaching 50.09%–62.02% on the more challenging ExpertQA dataset and 43.47%–48.00% on ELI5.

The consistent improvements across both compact and relatively large LLaMA models suggest that the effectiveness of CAGE mainly arises from its structured attribution mechanism rather than increased model capacity. Particularly substantial gains are observed on difficult settings, such as ASQA with the 3.2–1B model and ExpertQA with the 3.2–3B model, where CAGE improves TRUST by 10.29% and 13.02%, respectively. Although CAGE does not achieve the best score

Size	Type	ASQA(610 answerable, 338 unanswerable)				ExpertQA(682 answerable, 1487 unanswerable)				ELI5(207 answerable, 793 unanswerable)				Average
		Truthfulness		AG	TRUST	Truthfulness		AG	TRUST	Truthfulness		AG	TRUST	
		$F1_{AC}$	$F1_{GR}$	$F1_{GC}$		$F1_{AC}$	$F1_{GR}$	$F1_{GC}$		$F1_{AC}$	$F1_{GR}$	$F1_{GC}$		
2-7b	ICL	0.00	26.28	0.00	8.76	0.00	41.01	9.25	16.84	0.50	46.71	5.04	15.57	13.72
	PostCite	0.07	35.23	0.00	11.77	4.85	44.27	5.23	18.12	1.86	44.98	13.80	17.29	15.73
	PostAttr	0.07	35.23	0.00	11.77	4.85	44.27	2.26	17.13	1.86	44.98	0.00	15.61	14.84
	FRONT	**60.47**	39.15	68.86	56.16	9.33	23.92	74.75	36.00	21.66	17.15	**52.72**	30.51	40.89
	T-ALIGN	52.48	66.12	**83.94**	67.51	25.03	67.91	62.46	51.8	22.54	**63.27**	47.35	44.39	54.57
	CAGE	58.59	**69.62**	83.34	**70.52**	**31.25**	**71.24**	**75.19**	**59.22**	**27.61**	61.44	49.15	**46.07**	**58.60**
	Δ	1.88↓	3.50↑	0.60↓	3.01↑	6.22↑	3.33↑	0.44↑	7.42↑	5.07↑	1.83↓	3.57↓	1.68↑	4.04↑
3.2-1b	ICL	35.95	50.94	9.96	32.28	21.55	32.83	9.04	21.14	12.87	27.10	5.23	15.07	22.83
	PostCite	0.59	50.22	0.24	17.02	5.48	49.1	2.67	19.08	2.04	50.88	1.02	17.98	18.03
	PostAttr	0.48	48.42	0.00	16.30	8.24	47.72	1.5	19.15	2.04	50.88	0.07	17.66	17.70
	FRONT	48.22	54.48	48.29	50.33	20.83	29.26	37.45	29.18	16.11	20.76	30.19	22.35	33.95
	T-ALIGN	38.64	58.61	**79.35**	58.87	20.32	64.87	62.1	49.1	13.20	59.35	**48.21**	40.25	49.41
	CAGE	**60.20**	**70.07**	77.19	**69.16**	**31.17**	**71.85**	**74.26**	**50.09**	**27.69**	**61.46**	41.25	**43.47**	**54.24**
	Δ	11.98↑	11.46↑	2.16↓	10.29↑	10.34↑	6.98↑	12.16↑	0.99↑	14.49↑	2.11↑	6.96↓	3.22↑	4.83↑
3.2-3b	ICL	2.04	27.98	53.95	27.99	33.5	51.21	38.37	41.03	18.55	55.56	30.70	34.94	34.65
	PostCite	31.03	56.59	2.29	36.87	25.68	38.11	5.29	23.03	18.12	25.14	4.44	15.90	25.27
	PostAttr	29.76	56.71	4.69	30.39	25.45	38.58	3.4	22.48	18.48	25.14	0.53	14.72	22.53
	FRONT	**63.19**	49.45	57.46	56.70	27.24	43.34	50.91	40.5	19.95	32.21	41.97	31.38	42.86
	T-ALIGN	59.82	66.38	84.21	70.14	11.72	56.93	78.35	49.0	18.33	**62.79**	**55.87**	45.66	54.93
	CAGE	60.45	**69.95**	**87.65**	**72.68**	**33.60**	**71.61**	**80.83**	**62.02**	**30.70**	61.39	51.92	**48.00**	**60.90**
	Δ	2.74↓	3.57↑	3.44↑	2.54↑	6.36↑	14.68↑	2.48↑	13.02↑	12.37↑	1.40↓	3.95↓	2.34↑	5.97↑
3-8b	ICL	3.01	28.58	86.50	39.36	2.82	42.5	69.46	38.26	0.00	44.23	0.00	14.74	30.79
	PostCite	32.98	53.31	28.01	38.10	14.06	50.08	7.09	23.74	20.80	45.88	8.06	24.91	28.92
	PostAttr	32.98	53.31	5.95	30.75	14.06	50.08	6.29	23.47	20.80	45.88	1.25	22.64	25.62
	FRONT	**62.25**	41.62	66.14	56.67	30.34	24.92	56.7	37.32	18.99	17.85	44.69	27.18	40.39
	T-ALIGN	53.94	65.49	**88.26**	69.23	27.36	67.07	70.11	54.85	20.81	**63.57**	50.24	44.87	56.32
	CAGE	59.39	**69.52**	85.78	**71.56**	**30.89**	**71.60**	**76.48**	**59.66**	**27.45**	61.38	**51.77**	**46.86**	**59.36**
	Δ	2.86↓	4.03↑	2.48↓	2.33↑	3.53↑	4.53↑	6.37↑	4.81↑	6.64↑	2.19↓	1.53↑	1.99↑	3.04↑

Table 3: Main results on ASQA, ExpertQA, and ELI5 under the TRUST evaluation framework. Qwen3.5-9B is used as the CMI model, while LLaMA families models of different scales are used as SCR models. Δ denotes the improvement of CAGE over the strongest baseline.

Method	A-EBO↓	S-EBO↓	R-EBO↓
GPT-5.5	38.00	34.33	62.71
Claude-4.6	40.85	36.90	83.48
$CMI_{9B}+SCR_{Qwen\text{-}7B}$	14.69	13.11	39.28
$CMI_{9B}+SCR_{GPT\text{-}5.5}$	7.21	7.30	**15.49**
$CMI_{9B}+SCR_{Claude\text{-}4.6}$	**5.77**	**5.58**	18.94

Table 4: Evidence-boundary overrun on ASQA. A-EBO (Atomic-EBO) measures the proportion of unsupported atomic facts, S-EBO (EBO-Severity) their mean per-answer proportion, and R-EBO (Response-EBO) the proportion of answers containing any unsupported fact. Lower is better.

on every individual metric, it consistently provides the strongest aggregate TRUST performance. This result indicates that the cognitive attribution map helps balance answer correctness, response grounding, and citation quality by organizing claim-specific evidence before generation, thereby improving the robustness and cross-scale generalization of citation-grounded generation on LLaMA models.

A.3 Evidence-Boundary Overrun Evaluation

We evaluate evidence-boundary overrun (EBO) using the same 948 ASQA questions and the same top-5 retrieved documents for all methods. The evaluated outputs include Full CAGE, previously generated direct-answerer outputs, and newly generated outputs from the GPT-5.5 and Claude Sonnet 4.6 APIs. We use GPT-5.5 to decompose each long-form answer into atomic facts and apply the same TRUE model used for citation-groundedness evaluation to determine whether each fact is supported by the top-5 documents considered jointly. Refusal and empty responses are excluded from the EBO calculation.

Let N denote the number of valid responses, L_i the number of atomic facts in response i, $p_{i,\ell}$ its ℓ-th atomic fact, and $D_{i,5}$ the corresponding top-5 retrieved documents. We define $\mathrm{Entail}(p_{i,\ell}, D_{i,5}) \in \{0, 1\}$, where a value of 1 indicates that the atomic fact is supported by the document set.

Atomic-EBO. Atomic-EBO measures the proportion of atomic facts across all valid responses that are unsupported by the top-5 documents:

$$\text{Atomic-EBO} = 1 - \frac{\sum_{i=1}^{N} \sum_{\ell=1}^{L_i} \mathrm{Entail}(p_{i,\ell}, D_{i,5})}{\sum_{i=1}^{N} L_i}. \tag{28}$$

This metric is a fact-level micro-average. Responses containing more atomic facts therefore receive greater weight.

EBO-Severity. EBO-Severity first computes the proportion of unsupported atomic facts within each valid response and then averages it across responses:

$$\text{EBO-Severity} = \frac{1}{N} \sum_{i=1}^{N} \left(1 - \frac{1}{L_i} \sum_{\ell=1}^{L_i} \mathrm{Entail}(p_{i,\ell}, D_{i,5}) \right). \tag{29}$$

Each response receives equal weight, so this metric captures the average extent to which an answer exceeds its evidence boundary.

Response-EBO. Response-EBO measures the proportion of valid responses containing at least one unsupported atomic fact:

$$\begin{aligned} \text{Response-EBO} &= \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left[U_i > 0\right], \\ U_i &= \sum_{\ell=1}^{L_i} \left(1 - \mathrm{Entail}(p_{i,\ell}, D_{i,5})\right), \end{aligned} \tag{30}$$

This metric captures the occurrence of evidence-boundary overrun at the response level, regardless of the number of unsupported facts within the response. All three metrics are reported as percentages, and lower values indicate better evidence alignment.

CAGE substantially reduces evidence-boundary overrun. Following the atomic-fact decomposition paradigm of FActScore (Min et al., 2023), we use GPT-5.5 to decompose each valid long-form answer into atomic facts and apply TRUE to assess their support against the top-5 retrieved documents considered jointly. Unsupported facts are treated as evidence-boundary overruns. We report their overall proportion (Atomic-EBO), mean per-answer proportion (EBO-Severity), and occurrence rate across answers (Response-EBO).

As shown in Table 4, direct GPT-5.5 and Claude-4.6 yield Atomic-EBO scores of 38.00% and 40.85%, which fall to 7.21% and 5.77% after incorporating CMI with the corresponding SCR backbones. Similar reductions across EBO-Severity and Response-EBO show that stronger generation alone cannot ensure evidence-bounded answers. By constraining answer units to supporting evidence through a Cognitive Map, CAGE substantially reduces overrun at both the fact and answer levels.

A.4 Complete results of the Influence of CMI model across different SCR Backbones

Table 5 demonstrates that CAGE can effectively enhance diverse downstream LLMs, including both open-source and proprietary models, without requiring modifications to the underlying generators. Across ASQA, ELI5, and ExpertQA, CAGE consistently outperforms end-to-end baselines, achieving the best TRUST scores of 76.62% with Claude-4.6 on ASQA and 66.43% with Claude-4.6 on ExpertQA. Notably, when coupled with GPT-5.5 and Claude-4.6, CAGE improves TRUST by 7.21% and 11.10% on ASQA, and by 14.26% and 11.10% on ExpertQA, respectively. These gains are primarily attributed to stronger citation grounding, where CAGE achieves $\mathrm{F1}_{\mathrm{CG}}$ scores

of 94.48% on ASQA and 90.86% on ExpertQA with Claude-4.6. Unlike methods that rely on generator-specific optimization, CAGE separates evidence attribution from final response generation, allowing existing LLMs to leverage explicit support structures through a lightweight integration process. These results demonstrate that the induced attribution map provides a model-agnostic evidence alignment interface, enabling heterogeneous LLM backbones to produce more faithful and better-attributed responses while preserving their original generation capabilities. This property is particularly valuable for real-world applications built on proprietary LLM APIs, where model parameters and training pipelines are inaccessible.

A.5 Experimental results with same LLMs on CMI and SCR stage

Tables 6 and 7 demonstrate that CAGE remains effective even when the Cognitive Attribution Map Induction Model and Structured Citation Reasoning Model share the same backbone, indicating that the performance gains originate from explicit attribution modeling rather than backbone heterogeneity. Across different scales of Qwen and LLaMA models, CAGE and CAGE_{DPO} consistently improve TRUST over existing baselines, showing that the induced attribution graph provides additional reasoning capability beyond the original language models.

For Qwen models, the strongest CAGE configuration, using the 7B backbone, achieves TRUST scores of 67.14%, 58.32%, and 44.04% on ASQA, ExpertQA, and ELI5, respectively, while CAGE_{DPO} achieves its best TRUST scores of 70.47%, 62.23%, and 47.58% on ASQA, ExpertQA, and ELI5, respectively, with the 7B backbone. Similar trends are observed on LLaMA models, where CAGE obtains consistent gains, achieving TRUST scores of 63.35%, 66.49%, and 72.41% on ASQA with different model scales, and reaching 60.67% on ExpertQA with LLaMA-8B. Moreover, improvements are particularly evident in citation grounding, with CAGE variants achieving substantial gains in F1_{GC}, demonstrating stronger evidence alignment.

These results indicate that CAGE does not rely on additional model capacity or complementary backbones. Instead, the intermediate attribution graph explicitly organizes answer units and supporting evidence before generation, reducing attribution ambiguity and constraining citation selection. Therefore, even a single backbone model can benefit from the structured attribution process, validating the effectiveness and generality of CAGE as a model-independent framework for citation-grounded generation.

A.6 Experimental results when proprietary LLM as CMI model

Table 8 investigates the impact of different CMI models on CAGE performance while keeping the SCR models unchanged. The results show that stronger CMI models consistently improve the quality of induced attribution maps, leading to better citation-grounded generation across different SCR backbones. In particular, replacing GPT-5-mini with GPT-5.5 as the CMI model improves the average TRUST score across all datasets and SCR models, demonstrating that the quality of intermediate attribution structures directly affects downstream generation.

Quantitatively, GPT-5.5-based CAGE achieves substantial improvements over GPT-5-mini-based CAGE, especially on ASQA, where TRUST increases from 60.23–63.68% to 67.75–70.88%. Similar improvements are observed on ELI5 and ExpertQA, with notable gains in F1_{GC}, indicating that stronger CMI models mainly enhance evidence attribution and citation grounding rather than only answer generation. For example, with Qwen-2.5-7B as the SCR model, GPT-5.5 improves F1_{GC} from 72.48% to 82.66% on ASQA and from 47.69% to 64.08% on ExpertQA.

These results indicate that the CMI stage serves as a critical attribution bottleneck in CAGE. A stronger map induction model provides more accurate answer–evidence alignment, which enables the SCR model to generate better-supported claims regardless of its backbone architecture. Therefore, CAGE benefits not only from stronger generators but also from improving the quality of the intermediate cognitive attribution map.

A.7 METRICS.

We use TRUST Song et al. (2025) as the overall metric, which consists of three dimensions: answer calibration $\mathrm{EM}^{F1}_{\mathrm{AC}}$, refusal scoring $\mathrm{F1}_{\mathrm{RG}}$, and $\mathrm{F1}_{\mathrm{CG}}$ computed using the TRUE[1] model. Following Song *et al.*, the test set is divided into answerable questions A_g and unanswerable questions A_{ug}. For $\mathrm{EM}^{F1}_{\mathrm{AC}}$, it is formalized as:

$$\mathrm{EM}^{\mathrm{F1}}_{\mathrm{AC}} = \frac{2\mathrm{EM}^{\alpha}_{\mathrm{AC}} \cdot \mathrm{EM}^{\beta}_{\mathrm{AC}}}{\mathrm{EM}^{\alpha}_{\mathrm{AC}} + \mathrm{EM}^{\beta}_{\mathrm{AC}}}, \tag{31}$$

where

$$\begin{aligned} \mathrm{EM}^{\alpha}_{\mathrm{AC}} &= \frac{1}{|A_r|} \sum_{q_i \in A_g \cap A_r} \mathrm{EM}^{q_i}_{\mathrm{AC}}, \\ \mathrm{EM}^{\beta}_{\mathrm{AC}} &= \frac{1}{|A_g|} \sum_{q_i \in A_g \cap A_r} \mathrm{EM}^{q_i}_{\mathrm{AC}}, \end{aligned} \tag{32}$$

and A_r denotes the set of answered questions. The refusal score $\mathrm{F1}_{\mathrm{RG}}$ is calculated as $\mathrm{F1}_{\mathrm{RG}} = \frac{1}{2}\left(\mathrm{F1}_{\mathrm{ref}} + \mathrm{F1}_{\mathrm{ans}}\right)$, where $\mathrm{F1}_{\mathrm{ref}} = \frac{2\mathrm{P}_{\mathrm{ref}} \cdot \mathrm{R}_{\mathrm{ref}}}{\mathrm{P}_{\mathrm{ref}} + \mathrm{R}_{\mathrm{ref}}}$, $\mathrm{F1}_{\mathrm{ans}} = \frac{2\mathrm{P}_{\mathrm{ans}} \cdot \mathrm{R}_{\mathrm{ans}}}{\mathrm{P}_{\mathrm{ans}} + \mathrm{R}_{\mathrm{ans}}}$, and

$$\begin{aligned} P_{\mathrm{ref}} = \frac{|\neg A_r \cap \neg A_g|}{|\neg A_r|}, R_{\mathrm{ref}} = \frac{|\neg A_r \cap \neg A_g|}{|\neg A_g|}, \\ P_{\mathrm{ans}} = \frac{|A_r \cap A_g|}{|A_r|}, R_{\mathrm{ans}} = \frac{|A_r \cap A_g|}{|A_g|}. \end{aligned} \tag{33}$$

For $\mathrm{F1}_{\mathrm{CG}}$, following Gao et al. (2023), we report the citation-grounded F1 score $\mathrm{F1}_{\mathrm{CG}} = \frac{2 \cdot \mathrm{CP} \cdot \mathrm{CR}}{\mathrm{CP} + \mathrm{CR}}$,: where CP and CR denote citation precision and recall, respectively. Finally, $\mathbf{TRUST} = \frac{1}{3}(\mathrm{F1}_{\mathrm{RG}} + \mathrm{EM}^{\mathrm{F1}}_{\mathrm{AC}} + \mathrm{F1}_{\mathrm{CG}})$.

B CASE STUDY

Figure 5 presents a case study illustrating how CAGE performs evidence-grounded reasoning through attribution graph construction. Given a complex question requiring multi-source information integration, CAGE first identifies multiple support subgraphs from retrieved documents, where each subgraph corresponds to a distinct answer unit with explicit evidence paths. In this example, the induced graph separates different aspects of the answer, including Anatoli Boukreev's death, Scott Fischer's death, and the total number of casualties, while linking each answer unit to its supporting documents. This decomposition prevents unsupported information mixing and enables precise claim-level citation generation. Compared with directly generating answers from retrieved documents, CAGE explicitly determines which evidence supports each answer unit before decoding, resulting in more faithful responses with well-aligned citations. Notably, CAGE does not simply select all death-related documents; instead, it distinguishes different evidence-supported answer units and attributes each claim to its corresponding source.

[1] https://huggingface.co/google/t5_xxl_true_nli_mixture

Method	G	A	EM^{F1}_{AC}	$F1_{RG}$	$F1_{CG}$	**TRUST**
ASQA						
Ground-GRPO	End-to-end		47.22	62.88	81.94	64.01
GPT-5.5	End-to-end		66.18	61.54	70.69	66.14
Claude-4.6	End-to-end		70.46	64.56	73.22	69.41
CAGE	9B	4B	59.50	66.54	86.82	70.95
CAGE	9B	4B-Inst.	61.07	68.34	88.10	72.50
CAGE	9B	4B-Think.	61.03	67.07	86.07	71.39
CAGE	9B	GPT-5.5	61.41	70.09	92.32	**74.61**
CAGE	9B	Claude-4.6	65.29	70.09	94.48	**76.62**
ELI5						
Ground-GRPO	End-to-end		20.32	61.85	51.71	44.62
GPT-5.5	End-to-end		42.76	45.81	40.44	43.00
Claude-4.6	End-to-end		37.33	61.03	46.91	48.42
CAGE	9B	4B	26.98	61.11	57.86	48.65
CAGE	9B	4B-Inst.	30.73	61.48	57.58	**49.92**
CAGE	9B	4B-Think.	27.48	59.03	58.26	48.26
ExpertQA						
Ground-GRPO	End-to-end		20.56	67.14	59.37	49.02
GPT-5.5	End-to-end		46.74	54.46	49.78	50.33
Claude-4.6	End-to-end		47.66	64.54	53.80	55.33
CAGE	9B	4B	36.27	62.22	78.81	62.41
CAGE	9B	4B-Inst.	36.61	73.35	80.41	63.46
CAGE	9B	4B-Think.	37.59	73.24	79.15	63.33
CAGE	9B	GPT-5.5	32.88	71.74	89.15	**64.59**
CAGE	9B	Claude-4.6	36.70	71.74	90.86	**66.43**

Table 5: Backbone capability and proprietary model comparison. G and A denote the map induction model and SCR model, respectively. 9B denotes Qwen3.5-9B; Q4B, 4B-Inst., and 4B-Think. denote Qwen3-4B, Qwen3-4B-Instruct-2507, and Qwen3-4B-Thinking-2507, respectively. CAGE fixes Qwen3.5-9B as G and evaluates Qwen3-4B variants as A, while gray rows denote end-to-end baselines.

Size	Type	ASQA(610 answerable, 338 unanswerable)				ExpertQA(682 answerable, 1487 unanswerable)				ELI5(207 answerable, 793 unanswerable)			
		Truthfulness		AG	TRUST	Truthfulness		AG	TRUST	Truthfulness		AG	TRUST
		$F1_{AC}$	$F1_{GR}$	$F1_{GC}$		$F1_{AC}$	$F1_{GR}$	$F1_{GC}$		$F1_{AC}$	$F1_{GR}$	$F1_{GC}$	
1.5B	ICL	50.55	41.74	6.69	32.99	30.67	26.09	6.89	21.22	20.56	17.78	4.99	14.44
	PostCite	16.36	52.46	15.40	28.07	22.22	48.66	16.92	29.27	15.63	26.71	5.17	15.84
	PostAttr	16.36	52.46	4.45	24.42	22.22	48.66	13.15	28.01	15.63	26.71	0.62	14.32
	FRONT	**57.74**	41.36	55.70	51.60	**29.15**	24.60	50.22	34.66	19.57	17.29	37.70	24.85
	T-ALIGN	52.68	62.38	66.81	60.62	25.06	**68.38**	51.44	48.29	19.03	**57.91**	31.63	36.19
	CAGE	50.71	**66.89**	**77.55**	**65.05**	24.80	52.79	53.97	43.85	21.98	43.31	33.89	33.07
3b	ICL	37.72	51.36	51.72	46.93	35.14	49.65	42.67	42.49	29.12	46.31	34.34	36.59
	PostCite	9.58	35.30	10.94	18.61	0	40.66	0	13.55	21.73	48.49	7.56	25.93
	PostAttr	9.58	35.30	36.29	27.06	0	40.66	0	13.55	21.73	48.49	1.31	23.84
	FRONT	55.15	44.01	62.72	53.96	25.67	29.86	44.48	33.34	18.69	25.37	37.40	27.15
	T-ALIGN	55.19	63.76	78.64	65.86	20.97	65.79	60.25	49.0	22.52	64.38	42.01	42.97
	CAGE	**55.78**	61.12	78.09	65.00	**31.71**	**68.17**	69.28	**56.38**	25.12	58.80	41.32	41.75
7b	ICL	58.94	54.34	75.46	62.91	36.33	42.28	56.09	44.9	28.27	37.13	44.13	36.51
	PostCite	27.52	45.93	4.19	25.88	25.58	54.9	13.77	31.42	21.82	22.23	7.03	17.03
	PostAttr	27.52	45.93	17.92	30.46	25.58	54.9	12.46	30.98	21.82	22.23	0.96	15.00
	FRONT	64.58	60.08	58.27	60.98	32.41	55.56	67.35	51.77	28.27	54.14	56.61	46.34
	T-ALIGN	55.04	**66.22**	83.57	68.28	25.57	69.16	62.7	52.48	24.30	63.79	47.02	45.04
	CAGE	60.24	59.03	82.15	67.14	34.78	69.00	71.19	58.32	28.84	54.48	48.80	44.04

Table 6: Experimental results of different model scales of Qwen on the ASQA, ExpertQA, and ELI5 datasets.

Size	Type	ASQA(610 answerable, 338 unanswerable)				ExpertQA(682 answerable, 1487 unanswerable)				ELI5(207 answerable, 793 unanswerable)			
		Truthfulness		AG	TRUST	Truthfulness		AG	TRUST	Truthfulness		AG	TRUST
		$F1_{AC}$	$F1_{GR}$	$F1_{GC}$		$F1_{AC}$	$F1_{GR}$	$F1_{GC}$		$F1_{AC}$	$F1_{GR}$	$F1_{GC}$	
2-7b	ICL	0.00	26.28	0.00	8.76	0.00	41.01	9.25	16.84	0.50	46.71	5.04	15.57
	PostCite	0.07	35.23	0.00	11.77	4.85	44.27	5.23	18.12	1.86	44.98	13.80	17.29
	PostAttr	0.07	35.23	0.00	11.77	4.85	44.27	2.26	17.13	1.86	44.98	0.00	15.61
	FRONT	60.47	39.15	68.86	56.16	9.33	23.92	74.75	36.00	21.66	17.15	52.72	30.51
	T-ALIGN	52.48	66.12	83.94	67.51	25.03	67.91	62.46	51.8	22.54	63.27	47.35	44.39
	CAGE	49.71	63.12	77.25	63.35	23.88	64.09	78.45	55.48	22.49	57.89	40.19	40.19
3.2-1b	ICL	35.95	50.94	9.96	32.28	21.55	32.83	9.04	21.14	12.87	27.10	5.23	15.07
	PostCite	0.59	50.22	0.24	17.02	5.48	49.1	2.67	19.08	2.04	50.88	1.02	17.98
	PostAttr	0.48	48.42	0.00	16.30	8.24	47.72	1.5	19.15	2.04	50.88	0.07	17.66
	FRONT	48.22	54.48	48.29	50.33	20.83	29.26	37.45	29.18	16.11	20.76	30.19	22.35
	T-ALIGN	38.64	58.61	79.35	58.87	20.32	64.87	62.1	49.1	13.20	59.35	48.21	40.25
	CAGE	43.97	58.77	67.51	56.75	24.59	64.84	62.41	50.62	13.13	56.54	36.56	35.41
3.2-3b	ICL	2.04	27.98	53.95	27.99	33.5	51.21	38.37	41.03	18.55	55.56	30.70	34.94
	PostCite	31.03	56.59	2.29	36.87	25.68	38.11	5.29	23.03	18.12	25.14	4.44	15.90
	PostAttr	29.76	56.71	4.69	30.39	25.45	38.58	3.4	22.48	18.48	25.14	0.53	14.72
	FRONT	63.19	49.45	57.46	56.70	27.24	43.34	50.91	40.5	19.95	32.21	41.97	31.38
	T-ALIGN	59.82	66.38	84.21	70.14	11.72	56.93	78.35	49.0	18.33	62.79	55.87	45.66
	CAGE	52.31	60.47	86.70	66.49	24.86	63.99	82.19	57.01	29.43	61.85	45.64	45.64
2-8b	ICL	3.01	28.58	86.50	39.36	2.82	42.5	69.46	38.26	0.00	44.23	0.00	14.74
	PostCite	32.98	53.31	28.01	38.10	14.06	50.08	7.09	23.74	20.80	45.88	8.06	24.91
	PostAttr	32.98	53.31	5.95	30.75	14.06	50.08	6.29	23.47	20.80	45.88	1.25	22.64
	FRONT	62.25	41.62	66.14	56.67	30.34	24.92	56.7	37.32	18.99	17.85	44.69	27.18
	T-ALIGN	53.94	65.49	88.26	69.23	27.36	67.07	70.11	54.85	20.81	63.57	50.24	44.87
	CAGE	61.38	69.93	85.92	72.41	31.39	71.15	79.46	60.67	27.45	61.38	51.77	46.86

Table 7: Experimental results of different model scales of LLaMA on the ASQA, ExpertQA, and ELI5 datasets.

Dataset	SCR Model	GPT-5-mini as CMI				GPT-5.5 as CMI			
		EM^{F1}_{AC}	$F1_{GR}$	$F1_{GC}$	TRUST	EM^{F1}_{AC}	$F1_{GR}$	$F1_{GC}$	TRUST
ASQA	LLaMA3.2-1B	63.15	52.13	65.40	60.23	67.79	61.71	73.76	67.75
	LLaMA3.2-3B	65.49	52.36	73.20	63.68	68.52	61.71	81.01	70.42
	Qwen-2.5-3B	64.02	52.13	71.76	62.64	67.84	61.71	78.99	69.51
	Qwen-2.5-7B	65.15	52.58	72.48	63.40	68.26	61.71	82.66	70.88
ELI5	LLaMA3.2-1B	23.39	43.15	22.93	29.82	24.99	40.70	42.94	36.21
	LLaMA3.2-3B	25.03	43.32	31.04	33.13	28.95	40.99	56.19	42.04
	Qwen-2.5-3B	25.34	43.24	27.81	32.13	29.55	41.09	53.83	41.49
	Qwen-2.5-7B	25.41	43.24	30.15	32.93	30.26	41.09	55.08	42.14
ExpertQA	LLaMA3.2-1B	33.66	53.64	37.69	41.67	32.17	45.01	53.18	43.45
	LLaMA3.2-3B	37.88	55.18	48.49	47.18	36.82	45.03	64.75	48.87
	Qwen-2.5-3B	38.53	55.08	46.56	46.72	35.38	45.08	62.94	47.80
	Qwen-2.5-7B	39.19	55.18	47.69	47.35	37.54	45.08	64.08	48.90

Table 8: Effect of CMI model quality on CAGE performance with different SCR backbones. GPT-5-mini and GPT-5.5 are used as CMI models, while finetuned Qwen and LLaMA models serve as SCR models.

Case Study

Question
Who died in the book *Into Thin Air*?

Retrieved Documents

D1
Many public criticisms at Krakauer concerning the accuracy of each man's account of what happened on the mountain during the 1996 climbs. **Boukreev was killed in 1997 in an avalanche during a winter ascent of Annapurna in Nepal.**

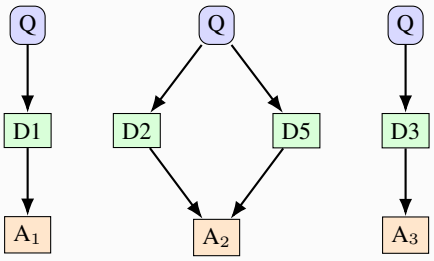
D2
Journalist Jon Krakauer, on assignment from *Outside* magazine, was in a party led by guide Rob Hall that lost four climbers on the south side; he afterwards published the bestseller Into Thin Air (1997), which related his experience. **Anatoli Boukreev was a guide in Scott Fischer's party, which lost Scott Fischer, but no clients.**

D3
Into Thin Air details the author's experience at the 1996 Mount Everest disaster, **in which eight climbers were killed and several others were stranded by a rogue storm.**

D4
Adventure Consultants was the guiding company led by Rob Hall.Members of his party died on the way down. The disaster became very well known, with ten million people reading about it in *Into Thin Air*. In 1996, Hall also employed two Sherpa people who managed to survive.

D5
The storm led to a number of deaths, **including both head guides (Rob Hall and Scott Fischer)**. Krakauer clarified his initial statements—especially those regarding the death of Andy Harris—in *Into Thin Air*.

Support Graph



Edges:
subgraph0
nodes:Q inputs: N/A outputs: D1
nodes:D1 inputs: Q outputs: A1
nodes:A1 inputs: D1 outputs: N/A
subgraph_description: The document supports the candidate answer by confirming that Anatoli Boukreev died in 1997 during a winter ascent of Annapurna.
subgraph1
nodes:Q inputs: N/A outputs: D2,D5
nodes:D2 inputs: Q outputs: A2
nodes:D5 inputs: Q outputs: A2
nodes:A2 inputs: D2,D5 outputs: N/A
subgraph_description: The documents support that Scott Fischer died in the 1996 Everest disaster, as they mention his party and the tragic event.
subgraph2
nodes:Q inputs: N/A outputs: D3
nodes:D3 inputs: Q outputs: A3
nodes:A3 inputs: D3 outputs: N/A
subgraph_description: The document supports the candidate answer by stating that eight climbers were killed during the 1996 Mount Everest disaster.

Generated Answer
Anatoli Boukreev, who served as a guide on the 1996 Mount Everest expedition described in Jon Krakauer's *Into Thin Air*, died in 1997 during a winter ascent of Annapurna in Nepal [1]. Scott Fischer, a guide in the 1996 Mount Everest disaster, died during the 1996 Everest disaster while leading his own expedition [2][5]. Eight climbers died during the 1996 Mount Everest disaster, as documented in Jon Krakauer's book *Into Thin Air* [3].

Figure 5: A case study illustrating graph-based reasoning for multi-source information integration. Subgraph 0 identifies Anatoli Boukreev's death, Subgraph 1 captures Scott Fischer's death supported by two documents, and Subgraph 2 provides the total death toll of eight climbers.

Prompt for Cognitive Attribution Map Construction

```
  Important rules for subgraph construction:
  1. Each subgraph corresponds to exactly one answer node: subgraph0 -> A1, subgraph1 ->
     A2, subgraph2 -> A3, and so on. Never reuse the same A node across multiple subgraphs
     .
  2. Each subgraph expresses one semantically independent answer aspect. If different
     documents support different facts or angles of the answer, split them into separate
     subgraphs with separate A nodes.
  3. Only group multiple T nodes into the same subgraph when they jointly support the same
      single answer aspect (as T1 and T5 both support who won the 2018 World Cup in
     subgraph0).
  The most important: 4. When all documents fail to support answering questions, an empty
     chart should be output. In the empty chart, the input and output of each document
     node are empty, and Reason (Within 20 tokens) should provide an explanation for each
     document indicating that it is not supported.
  Empty edges is:

  Edges:
  nodes:Q       inputs:         outputs:
  nodes:D1      inputs:         outputs:
  reason: ...
  nodes:D2      inputs:         outputs:
  reason: ...
  nodes:D3      inputs:         outputs:
  reason: ...
  nodes:D4      inputs:         outputs:
  reason: ...
  nodes:D5      inputs:         outputs:
  reason: ...
  For example, the input is:
  Who was the champion of the 2018 World Cup?
  Document 1: France defeated Croatia 4-2 in the 2018 final.
  Document 2: Kylian Mbappé contributed 2 goals in the final, playing a significant role
     in France's championship victory.
  Document 3: It discusses the 2022 World Cup final

  Document 4: It discusses the semi-finals of the 2018 World Cup
  Document 5: France was the champion of the 2018 World Cup
  We want the model output(one subgraph, one A node, one independent answer aspect.):
  Edges
  subgraph0
  nodes:Q       inputs: N/A     outputs: D1,D5
  nodes:D1      inputs: Q       outputs: A1
  nodes:D5      inputs: Q       outputs: A1
  nodes:A1      inputs: D1, D5  outputs: N/A
  subgraph_description: Document 1 and Document 5 jointly directly support the answer,
     explicitly stating that France was the champion of the 2018 World Cup. Document 1
     provides details of the final score, while Document 5 directly states the ownership
     of the championship

Based on a given single question node Q and several document title nodes D (D1--D5),
     construct one or more subgraphs containing answer node A through reasoning.  The
     question should be treated as answerable if at least one document node contains
     explicit statements or clear indicators that directly answer the question, even if

     other documents are irrelevant.  When the question involves multiple entities, multi-
     hop reasoning, or requires information decomposition, multiple subgraphs can be
     generated, but they must maintain logical clarity and semantic independence. The A
     node in the generated edges should be inferred from the supporting D node(s). The
     description of the A node should indicate that all its input D nodes individually
     support the A node, or collectively contribute to the generation of the A node. D
     nodes that are irrelevant to the A node should not be included as inputs to the A
     node.
  {user_input}
```

Prompt for Long-form Answer Generation on the Map

```
Generate the final answer from the retrieved documents, using the graph only to decide
    citations.

  Graph schema:
  - Q is the user question.
  - D nodes are retrieved documents. D1 maps to citation [1], D2 maps to citation [2], and
     so on.
  - Each subgraph is one citation plan. It lists the D documents that should support one
    answer sentence.
  - subgraph_description is only a hint about why those D documents were selected; do not
    copy it as the answer.

  Rules:
  - If at least one evidence path is provided, answer the question. Do not refuse.


  - For each subgraph, write exactly one independent answer sentence grounded only in the
    documents used by that
  subgraph.
  - Do not merge subgraphs, even if they appear to support the same answer.
  - Keep each sentence short and focused on one factual claim.
  - Cite that sentence with exactly the numeric citations for the D nodes in that subgraph
    , unless a listed document
  does not support the sentence.
  - Use document text, not subgraph_description text, to decide the factual content of the
     sentence.
  - Every answer sentence must end with numeric citations in this exact form: sentence
    text [1].
  - Do not put citations on a separate line or at the start of a sentence.
  - Do not cite documents that are not listed for the evidence paths.
  - Output the independent answer sentences as one concise paragraph.

  Question:
  {question}


  Evidence paths extracted from the graph:
  Path {i}:
  - Citation nodes for one answer sentence: {[citation_ids]}
  - Selection hint, not answer text: {subgraph_description}

  Referenced documents:
  [{citation_id}] {document_title}
  {document_text}

  Original graph structure:
  {graph_input}
```