SAGE-KG: A Comprehensive Evaluation of Agentic KG Generation

Yashrajsinh Chudasama*, Deepanshi Acharya*, Arpit Rana*
*Dhirubhai Ambani University, Gandhinagar, India

I. Introduction

Large language models (LLMs) excel in reasoning and generation tasks but are prone to hallucinations when knowledge is missing or outdated. Augmenting LLMs with structured external knowledge, such as knowledge graphs (KGs), can improve factual reliability. Existing KG construction methods, however, often rely on proprietary models, predefined schemas, or domain-specific fine-tuning, limiting scalability and accessibility.

We present SAGE-KG, an agentic framework that uses small, open-source language models to automatically convert unstructured text into high-quality knowledge graphs without supervision or fixed schema requirements. The framework orchestrates three sequential agents—fact extractor, entity-relation schema planner, and triplet creator—to produce highly accurate and complete (subject, predicate, object) triplets across diverse domains. For downstream evaluation, we integrate a retrieval mechanism combining query-conditioned seed selection, multi-hop graph traversal, and cross-encoder reranking to provide context for grounded question answering. Extensive intrinsic and extrinsic evaluation demonstrates that SAGE-KG consistently outperforms existing graph-based RAG baselines on multi-hop QA benchmarks, achieving high factual precision, completeness, and reasoning performance, all without exposing source documents to the model.

II. Literature Review

Evaluating the quality of an automatically constructed knowledge graph requires a clear understanding of how well the extracted entities, relations, and triplets reflect the semantics of the underlying source text. To assess both intrinsic and extrinsic aspects of the generated graph, it was essential to review established methodologies, baseline models, and evaluation frameworks used in prior research. This literature review examines key approaches in knowledge graph construction, triplet extraction, and graph evaluation, providing the foundation upon which our evaluation strategy was designed.

A. Experimental Setup

1) Datasets: To ensure a fair and representative evaluation spanning a range of multi-hop reasoning styles, we sample 500 (question, answer, context) tuples from three widely used datasets bundled in the FlashRAG toolkit [38]: (i) **HotpotQA** [5], which emphasizes Wikipedia-based sentence-level supporting facts, (ii) **2WikiMultiHopQA** [10], which requires reasoning across multiple documents, and (iii) **MuSiQue** [12], which targets compositional inference over multistep fact chains. These datasets collectively emphasize both factual grounding and multi-hop completeness, making them well-suited for evaluating the quality of KG construction.

2) Baselines: We compare against three representative triplet extraction paradigms: (1) **Stanford OpenIE** [3], a classical linguistic pipeline; (2) **KGGen** [41], a state-of-the-art supervised LLM-based KG construction system using GPT-4o; and (3) **Zero-Shot GraphRAG** [21], a prompt-based triplet generator using Qwen2.5-14B. All baseline systems are run with the same preprocessing pipeline and hardware configuration for controlled comparison.

Our framework uses Qwen2.5 models (3B, 7B, 14B) [44] deployed through Ollama with `Q4_K_M` quantization [48] on an RTX 6000 Ada GPU [47]. Extraction is run deterministically with fixed decoding parameters (T=0, top-k=40, top-p=0.95). Chunking follows the LlamaIndex preprocessing pipeline (400 tokens, 50-token overlap) [46].

III. Methodology

A. Knowledge Multi-Hop Retrieval

Once the knowledge graph is populated, the retrieval phase maps a natural-language query q to a contextual triplet set through a structured operator $\mathcal{Q} : q \mapsto \mathcal{T}_q$. The process combines semantic subgraph localization, seed entity identification, and multi-hop reasoning by integrating complementary mechanisms from existing GraphRAG pipelines while preserving controllability and maintaining a tractable boundary for the search space. We refer to this enhanced retrieval framework as **Enhanced Graph Traversal (EGT)**, highlighting its adaptation and augmentation of prior GraphRAG methods for context-aware, multi-hop reasoning. The overall workflow of the retrieval and answer generation is depicted in Figure **??**.

1) Subgraph Selection via Semantic Similarity: To restrict retrieval to semantically relevant regions of the graph, we first operate at the chunk level. Given a query q, a dense embedding $\mathbf{e}_q = \phi_{\text{query}}(q) \in \mathbb{R}^{1024}$ is computed and compared against precomputed chunk embeddings $\mathbf{e}_c$ via cosine similarity:

$$\text{sim}_{\cos}(\mathbf{e}_q, \mathbf{e}_c) = \frac{\mathbf{e}_q \cdot \mathbf{e}_c}{||\mathbf{e}_q||_2 \cdot ||\mathbf{e}_c||_2}. \quad (1)$$

The top-k ranked chunks form the candidate region $\mathcal{C}_{\text{sel}}$, thereby reducing the graph to a focused subspace prior to finer entity reasoning. This step draws methodological inspiration

from HippoRAG [25], but differs in that it operates on structured triplet-derived embeddings rather than fallback chunk reconstructions.

2) Seed Entity Selection: Within $\mathcal{C}_{\text{sel}}$, entities most relevant to the query are identified using a hybrid selection strategy that balances semantic and lexical cues:

$$\mathcal{V}_{\text{vec}} = \arg\text{topk}_{n \in \mathcal{V}_{\mathcal{C}_{\text{sel}}}} \text{sim}_{\cos}(\mathbf{e}_q, \mathbf{e}_n)$$
$$\mathcal{V}_{\text{bm25}} = \arg\text{topk}_{n \in \mathcal{V}_{\mathcal{C}_{\text{sel}}}} \text{BM25}(q, n.\text{name})$$

The final seed set is the deduplicated union:

$$\mathcal{V}_{\text{seeds}} = \mathcal{V}_{\text{vec}} \cup \mathcal{V}_{\text{bm25}}, \quad |\mathcal{V}_{\text{seeds}}| \leq 10.$$

Embeddings from BERT-based [7] models capture conceptual similarity (e.g., "tariff cap" vs. "price ceiling"), whereas BM25 retains exact lexical matching, which is especially useful when dealing with domain-specific terms or proper names.

3) Graph Traversal and Multi-Hop Reasoning: From the seed set, reasoning proceeds through controlled graph traversal to uncover relevant relational structure. We integrate two complementary traversal modes:

a) Multi-hop Neighborhood Expansion.: For each seed $v \in \mathcal{V}_{\text{seeds}}$, a bounded-depth exploration retrieves relational neighborhoods by collecting all paths originating from v with depth at most 3:

$$\mathcal{P}_{\text{multihop}}(v) = \{(s,p,o) \mid (s,p,o) \in \text{path}(v),\ |\text{path}(v)| \leq 3\}.$$

This captures implicit supporting context and cross-chunk continuity, capped at $\tau_{\text{hop}} = 25$ triplets per seed for tractability [29].

b) Structured Reasoning Path Discovery.: Inspired by PathRAG [37], shortest-path reasoning is invoked when a query implies connections between disjoint entities or factual clusters:

$$\mathcal{P}_{\text{reasoning}} = \{(s,p,o) \mid (s,p,o) \in \pi(v_s, v_t),\ |\pi| \leq 3\},$$

where $\pi(v_s, v_t)$ denotes a constrained path between candidate source and target nodes. This step surfaces chains that span documents or policy clauses, which are otherwise missed by purely local traversal.

The final retrieval set $\mathcal{T}_q$ is the union of multi-hop expansions and reasoning paths, passed downstream to ranking/reranking and ultimately answer synthesis. This hybrid design maintains both factuality and completeness, achieving retrieval behavior that is faithful to the underlying graph semantics rather than the raw text corpus.

B. Answer Generation with LLM

The final stage converts the retrieved relational evidence into a natural-language answer grounded in knowledge graph-derived context. Let $\mathcal{T}_{\text{raw}} = \mathcal{P}_{\text{multihop}} \cup \mathcal{P}_{\text{reasoning}}$ denote the union of traversal-derived triplets. Since traversal is performed on an undirected representation for completeness, triplets are first normalized and deduplicated using canonical (subject, predicate, object) keys to produce a unique evidence pool.

1) Cross-Encoder Reranking: To retain only the most query-relevant evidence, we apply cross-encoder reranking on the candidate triplet set. Each triplet t is serialized into textual form and scored jointly with the query using the cross-encoder model `ms-marco-MiniLM-L-12-v2` [35].

$$\text{score}_{\text{cross}}(q,t) = \text{CrossEncoder}(q, \text{serialize}(t)) \in [0,1]. \quad (2)$$

The top-κ triplets ($\kappa = 50$ in our implementation) according to this relevance score form the final contextual knowledge set $\mathcal{T}_q$. This step filters out traversal noise and ensures that downstream reasoning is both compact and semantically aligned with the information need.

2) LLM-Guided Answer Synthesis: The filtered context $\mathcal{T}_q$ is then passed to a lightweight LLM for controlled answer generation:

$$\text{answer} = \text{LLM}(q, \mathcal{T}_q \mid \text{prompt}). \quad (3)$$

The prompt template explicitly instructs the model to treat triplets as ground truth and prohibits unsupported inference. Unlike fully parametric QA, which relies on latent memory, this formulation enables explicit grounding in the constructed knowledge graph—thereby reducing hallucinations and ensuring auditability. The LLM functions solely as a surface realizer, while correctness is guaranteed by the evidence selection pipeline.

This separation of retrieval and generation enforces traceability: each component of the final answer can be linked back to a concrete triplet and, through stored provenance identifiers, ultimately to the source document.

IV. Experiments

A. Intrinsic Evaluation

1) Evaluation Metrics: Since gold-standard triplets are not provided in these datasets and manual scoring is not scalable for large graph extractions, we adopt the LLM-as-Judge protocol [14], [23], which has shown high agreement with expert annotators in factual assessment tasks.

We report three intrinsic metrics: (i) **Factual Precision**: whether each extracted triplet is directly supported by the source text ; Triplets are rated on a 0–10 scale:

- 10: Fully accurate
- 8–9: Mostly correct
- 6–7: Partially incorrect
- 4–5: Several inaccuracies
- 1–3: Mostly incorrect
- 0: Completely false

(ii) **Completeness**: whether the extracted graph captures all salient facts present in the text; Triplets are scored on a 0–10 scale according to the following rubric:

- 10: Fully captures all ground truth facts with possibly helpful relevant detail.
- 8–9: Covers most facts clearly with minor omissions or additional context that does not contradict.
- 6–7: Captures some key facts but misses several points or adds moderately extraneous/non-contradictory information.

TABLE I: Intrinsic Evaluation Scores Across Datasets (1-10 scale). **Bold** indicates best performance.

Model	HotpotQA						MuSiQue						2WikiMultiHopQA					
	GPT-4o-mini			Gemini-2.0-flash			GPT-4o-mini			Gemini-2.0-flash			GPT-4o-mini			Gemini-2.0-flash		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
OpenIE	6.77	7.77	7.24	6.78	6.68	6.73	6.90	7.70	7.28	6.94	6.82	6.88	6.80	7.70	7.22	6.84	6.86	6.85
Zero-shot GraphRAG	**8.93**	8.39	8.65	8.43	**8.31**	8.15	8.91	8.38	8.64	8.40	7.81	8.10	8.89	8.26	8.56	**8.31**	7.81	8.01
KGGen	8.78	**9.07**	**8.92**	**8.80**	7.96	**8.55**	**9.10**	8.50	8.79	**8.79**	7.96	**8.35**	8.10	7.70	7.89	7.82	7.28	7.54
SAGE-KG:q-3b	6.80	6.70	6.75	6.41	6.20	6.30	6.80	7.00	6.90	6.44	6.29	6.36	6.40	6.60	6.50	6.08	5.97	6.02
SAGE-KG:q-7b	6.60	6.80	6.70	7.81	7.71	7.76	8.20	8.20	8.20	7.80	7.73	7.76	7.90	8.00	7.95	7.56	7.48	7.52
SAGE-KG:q-14b	8.90	8.80	8.85	8.35	8.22	8.28	8.90	**8.90**	**8.90**	8.35	**8.31**	8.33	**8.90**	**8.80**	**8.85**	8.28	**8.15**	**8.21**

- 4–5: Partial coverage with many omissions or questionable additional information.
- 1–3: Contains little of the ground truth facts.
- 0: No relevant facts are present or the answer is misleading.

(iii) **F1-based Average Quality**, representing the harmonic mean.

Let $\mathcal{T}(c)$ denote the extracted triplets from chunk c. Precision is defined as:

$$\mathcal{P}_f(t) = \text{LLM}_{\text{judge}}(t, c; \pi_{\text{precision}}), \tag{4}$$

where the judge verifies factual support without hallucination. the completeness is estimated by checking whether each ground-truth fact $f \in \mathcal{F}^*(c)$ is recoverable from the extracted KG:

$$\mathcal{R}_f(f) = \text{LLM}_{\text{judge}}(f, \mathcal{T}(c); \pi_{\text{recall}}). \tag{5}$$

The overall factual quality score is given by:

$$F1_f = 2 \cdot \frac{\mathcal{P}_f \cdot \mathcal{R}_f}{\mathcal{P}_f + \mathcal{R}_f}.$$

To reduce model-specific bias, we use both GPT-4o-mini [27] and Gemini 2.0 Flash [49] as independent judges and report scores averaged across both [45].

V. Experimental Results

In this section, we first report the results of the intrinsic evaluation, which analyzes the structural and qualitative characteristics of the constructed knowledge graphs. Table **??** presents a comparison of knowledge graph statistics across HotpotQA, MusiQue, and 2WikiMultiHopQA datasets for several KG construction methods, including OpenIE, Zero-shot GraphRAG, KGGen, and our SAGE-KG variants (Qwen2.5-3B, 7B, 14B). The table reports the total number of nodes, relations, and triplets, capturing the completeness of extraction, relational diversity, and semantic coverage. SAGE-KG graphs, particularly with larger models, achieve comparable or higher structural richness than prior methods while maintaining a more compact and semantically coherent representation. Average construction times for SAGE-KG were 266.8 minutes (3B), 262.8 minutes (7B), and 608.2 minutes (14B) across all datasets, with occasional variability in the 3B model on complex data. Overall, these results highlight that SAGE-KG effectively balances graph density, relational diversity, and construction efficiency relative to existing approaches.

Table I further summarizes the intrinsic evaluation scores across the three datasets under both GPT-4o-mini and Gemini 2.0 Flash judges.

On **HotpotQA**, which emphasizes sentence-level factual grounding, KGGen achieves the highest F1 (8.92) under GPT-4o-mini, primarily due to its superior recall. Zero-shot extraction achieves slightly higher precision (8.93) but suffers from lower recall, resulting in a marginally lower overall F1 score. Notably, our SAGE:q-14B model demonstrates competitive performance (8.85 F1), closely matching KGGen while maintaining a balanced precision–recall tradeoff, indicating effective agentic triplet extraction even with constrained architectural scaling.

For **MuSiQue**, which tests compositional multi-step reasoning, KGGen again leads in higher F1; however, SAGE:q-14B narrows the gap considerably (8.90 F1 on GPT-4o-mini) and surpasses lower-capacity SAGE variants, highlighting that model scaling substantially benefits compositional inference within the agentic framework.

On **2WikiMultiHopQA**, representing cross-document reasoning, SAGE:q-14B achieves the highest or tied F1 (8.85), outperforming KGGen (7.89) and maintaining stable scores across both judges. Zero-shot extraction excels marginally in precision (8.89) but suffers from reduced recall, which limits its multi-hop reasoning capacity. This demonstrates that SAGE's agentic knowledge graph construction produces more complete and coherent triplets capable of supporting cross-document inference.

Across judges, KGGen exhibits higher variance, particularly in recall for 2WikiMultiHopQA, whereas SAGE:q-14B shows smaller cross-judge differences, suggesting more robust factual grounding. The progression from SAGE:q-3B to q-14B confirms that the agentic extraction pipeline benefits from larger model capacity, improving both precision and recall while ensuring stable performance in multi-hop and cross-document scenarios.

Overall, these results indicate that while traditional structured LLM pipelines like KGGen perform well on localized extraction, the SAGE-KG framework provides a *scalable, agentic, and robust solution* that excels in complex multi-hop reasoning, achieving a favorable balance between precision,

TABLE II: MINE Dataset Benchmark Results

Dataset	SAGE-KG (ours)			Baseline Methods		
	q-3B	q-7B	q-14B	KGGen	GraphRAG	OpenIE
MINE	72.06	84.76	**92.95**	66.07	47.80	29.84

Fig. 1: Distribution of MINE scores for GraphRAG, OpenIE, KGGen, and SAGE Method (3B, 7B, 14B). Dotted vertical lines show average performance. SAGE-KG (Qwen2.5:14B) scored 92.95% on average, outperforming KGGen (66.07%), GraphRAG (47.80%), and OpenIE (29.84%).

recall, and judge-consistent factual quality.

A. Evaluation using the MINE Benchmark

To further assess the factuality and completeness of our extracted knowledge graphs, we evaluate on the MINE (Measure of Information in Nodes and Edges) benchmark [41]. MINE consists of manually verified facts designed for fine-grained KG evaluation, penalizing both omissions and hallucinations, thereby providing a stringent test of relational accuracy.

Table II presents the factual capture performance of our agentic extraction pipeline against representative baselines: OpenIE, GraphRAG, and KGGen. Across all model sizes, our approach consistently surpasses the baselines. In particular, the Qwen2.5:14B variant achieves **92.95%** factual capture, markedly higher than KGGen (66.07%), GraphRAG (47.80%), and OpenIE (29.84%).

Overall, these results demonstrate that the SAGE-KG framework not only excels in multi-hop intrinsic evaluations but also achieves state-of-the-art performance on highly controlled, manually verified benchmarks, confirming its ability to extract semantically rich and accurate knowledge graphs without overgeneration.

VI. Extrinsic Evaluation

Beyond intrinsic factual quality, we evaluate the downstream utility of our extracted knowledge graphs in multi-hop question-answering (QA) tasks. The goal is to measure whether SAGE-KG triplets, when integrated into a retrieval-augmented generation (RAG) pipeline, improve the accuracy, relevance, and overall quality of LLM responses.

A. Experimental Settings

1) Baselines: We benchmark our approach against three representative categories:

- **Triplet-extraction methods (SAGE-KG, OpenIE, KGGen, Zero-Shot GraphRAG).** To ensure a fair comparison, all triplets are stored as graph databases and queried through a unified Enhanced Graph Traversal (EGT) system, since these methods lack an integrated retrieval component.
- **Microsoft GraphRAG.** Both local and global variants are reproduced following official implementations. The local variant performs retrieval within community-level subgraphs, whereas the global variant leverages multi-level graph summarization to enable cross-cluster reasoning [34].
- **Standard RAG.** A conventional text-embedding pipeline is implemented using chunk-level retrieval with the `BAAI/bge-large-en-v1.5` encoder. Candidate passages are selected via cosine similarity and then fed to the LLM for answer generation [9].

2) Evaluation Metrics: For each dataset, the constructed knowledge graphs provide contextual triplets during inference with GPT-4o-mini. We evaluate responses along three complementary dimensions: *Exact Match*, *Semantic Relevance*, and *Generation Evaluation*. The first two provide objective lexical and semantic comparisons to ground truth answers, while the third captures higher-level qualitative properties using LLM-based judgments.

a) Exact Match (E-M).: Measures strict textual correctness relative to the ground truth a^*:

$$\text{EM} = \mathbb{I}(a = a^*) \tag{6}$$

b) Semantic Relevance (S-R).: Captures semantically equivalent but lexically divergent responses using cosine similarity between sentence embeddings $\phi(\cdot)$ from `all-MiniLM-L6-v2`:

$$\text{SR} = \text{cosine}(\phi(a), \phi(a^*)). \tag{7}$$

c) Generation Evaluation (G-E).: GPT-4o-mini (G-Eo) and Gemini 2.0 Flash (G-Eg) independently assess responses across five dimensions—completeness, accuracy, knowledgeability, relevance, and logical coherence—each scored on $[1, 10]$. For a given model m, the aggregate score is:

$$\mathcal{S}_m = \frac{1}{5} \sum_{i=1}^{5} s_i \tag{8}$$

Results are reported separately for each judge to preserve distinct evaluation perspectives, and all scores are normalized to the $[1, 100]$ range.

G-E Dimensions and Rubrics: 1. Completeness. Does the answer cover all necessary facts? Multi-hop reasoning requires integrating evidence from multiple KG nodes.

Prompt (Exact Rubric Used):

```
"completeness": {
    "description": "whether the answer
        includes ALL important facts and
        distinct points from the ground truth,
         allowing consistent factual additions
        ",
    "rubric": (
        "Scoring Guide (0-10):"
        "10: Fully captures all Ground Truth
            facts with possibly helpful
            relevant detail."
        "8-9: Covers most facts clearly with
            minor omissions or some additional
             context that does not contradict
            ."
```

```
        "6-7: Captures some key facts but
            misses several points or adds
            moderately extraneous/non-
            contradictory info."
        "4-5: Partial coverage with many
            omissions or questionable
            additional info."
        "1-3: Contains little of the Ground
            Truth facts."
        "0: No relevant facts are present or
            answer is misleading."
    )
}
```

2. Accuracy. Are the facts strictly correct? This assesses factual consistency and ensures hallucinations are minimized.

Prompt (Exact Rubric Used):

```
"accuracy": {
    "description": "whether the answer is
        factually correct compared to ground
        truth, tolerating consistent
        elaborations",
    "rubric": (
        "Scoring Guide (0-10):"
        "10: Fully accurate; no factual errors
            ."
        "8-9: Mostly accurate with minor
            trivial errors or consistent
            additions."
        "6-7: Some factual inaccuracies or
            minor misinterpretations."
        "4-5: Several incorrect points."
        "1-3: Largely incorrect."
        "0: Completely false or unrelated."
    )
}
```

3. Knowledgeability. Evaluates whether the answer demonstrates domain understanding and correct reasoning depth.

Prompt (Exact Rubric Used):

```
"knowledgeability": {
    "description": "whether the answer shows
        accurate domain knowledge consistent
        with the ground truth, allowing
        relevant expansions",
    "rubric": (
        "Scoring Guide (0-10):"
        "10: Fully matches domain knowledge
            with clarity."
        "8-9: Mostly aligns with minor gaps or
             some relevant added detail."
        "6-7: Exhibits some understanding but
            also gaps."
        "4-5: Limited knowledge shown."
        "1-3: Minimal or incorrect domain
            knowledge."
        "0: No relevant domain knowledge."
    )
}
```

4. Relevance. Checks whether the answer stays fully on-topic with no unnecessary digressions.

Prompt (Exact Rubric Used):

```
"relevance": {
    "description": "whether the answer stays
        on-topic using only ground truth facts
         or consistent relevant information",
    "rubric": (
        "Scoring Guide (0-10):"
        "10: Entirely relevant and on-topic."
       "8-9: Mostly relevant; minimal off-
           topic content."
        "6-7: Some minor digressions."
        "4-5: Noticeable off-topic content."
        "1-3: Barely related."
        "0: Completely irrelevant."
    )
}
```

5. Logical Coherence. Evaluates clarity, flow, and whether reasoning is logically structured.

Prompt (Exact Rubric Used):

```
"logical_coherence": {
    "description": "whether the answer
        presents the ground truth facts
        clearly and logically, with possible
        well-integrated expansions",
    "rubric": (
        "Scoring Guide (0-10):"
        "10: Clear, well-structured, logically
             coherent."
        "8-9: Mostly clear with minor flow
            issues."
        "6-7: Some structure but less clear."
        "4-5: Poorly organized."
        "1-3: Very hard to follow."
        "0: Completely incoherent."
    )
}
```

B. Experimental Results

Table III summarizes the performance of our agentic SAGE-KG approach and baseline methods on three multi-hop QA datasets—HotpotQA, MuSiQue, and 2WikiMultiHopQA—across four evaluation metrics: Exact Match (E-M), Semantic Relevance (S-R), and two judge-based generation evaluation scores (G-Eo for GPT-4o-mini and G-Eg for Gemini 2.0 Flash).

Overall Observations. Naive LLM generation without KG grounding performs poorly across datasets, with low E-M and S-R scores, as it struggles to capture multi-hop dependencies. Many comparison-based questions further induce hallucinations or fact conflation, lowering accuracy. Standard RAG notably improves S-R and G-E metrics via embedding-based retrieval. Microsoft GraphRAG achieves moderate gains, with the global variant slightly outperforming the local one but remaining limited by coarse subgraph retrieval.

Triplet-Based RAG Methods. Among graph-based triplet extractors integrated with Enhanced Graph Traversal (EG), OpenIE provides modest improvements over naive generation, particularly in Semantic Relevance and G-E scores, but suffers from low E-M due to noisy or incomplete extractions. KGGen and Zero-Shot GraphRAG show stronger performance, with

TABLE III: Extrinsic Performance Across Question-Answering Datasets

Baselines	Query System	Model	HotpotQA				MuSiQue				2WikiMultiHopQA			
			E-M	S-R	G-Eo	G-Eg	E-M	S-R	G-Eo	G-Eg	E-M	S-R	G-Eo	G-Eg
Naive Generation	None	GPT-4o-mini†	10.20	37.37	51.60	46.60	0.20	22.21	33.84	27.42	10.00	34.96	34.86	33.46
	None	Qwen2.5-14B†	07.20	27.48	22.40	23.86	0.00	13.01	6.02	8.48	03.80	22.08	11.02	13.94
Standard RAG	Cosine Sim.§	GPT-4o-mini†	-	33.74	62.00	53.86	-	26.16	43.48	32.04	-	33.51	42.38	42.88
Microsoft GraphRAG	Global§	GPT-4o*	-	25.06	53.00	47.40	-	23.27	44.90	31.98	-	24.06	36.88	35.52
	Local§	GPT-4o*	-	28.76	50.90	44.92	-	07.69	44.50	36.10	-	30.86	36.76	32.72
OpenIE	EGT	-‡	19.00	50.98	55.60	52.64	08.00	39.70	45.42	39.90	14.40	48.30	45.02	44.92
KGGen	EGT	GPT-4o*	29.00	60.01	71.40	69.10	16.80	46.16	58.82	52.14	26.00	57.27	59.64	58.76
Zero-Shot GraphRAG	EGT	Qwen2.5-14B*	31.20	61.87	63.14	61.98	**24.00**	**57.21**	58.00	54.74	19.20	50.43	52.54	50.60
SAGE-KG (Ours)	EGT	Qwen2.5-3B*	20.00	51.40	57.10	54.26	10.00	46.79	55.42	49.20	17.20	46.20	44.82	42.78
	EGT	Qwen2.5-7B*	33.60	62.82	64.70	63.50	14.40	49.53	61.24	57.00	17.80	51.50	52.24	51.34
	EGT	Qwen2.5-14B*	**38.00**	**68.72**	**73.80**	**73.40**	17.20	51.20	**62.60**	**57.28**	**27.40**	**60.04**	**59.28**	**62.38**
	ToG-v1	Qwen2.5-14B*	38.00	66.91	67.04	67.02	23.40	54.55	51.44	49.46	44.60	64.02	57.24	58.02

Notes: The model is used for KG construction; ‡ indicates rule-based triplet generation. GPT-4o-mini is used to generate answers for all baselines except Naive Generation using Qwen2.5:14b. **G-Eo** denotes generation evaluation with GPT-4o-mini as the judge, and **G-Eg** denotes evaluation with Gemini 2.0 Flash. **EGT** refers to the Enhanced Graph Traversal module, **ToG-v1** represents the first version of the Think-on-Graph query system, and § indicates the native query baseline.

Zero-Shot GraphRAG achieving the highest S-R on MuSiQue (57.21) and competitive G-E scores, reflecting the effectiveness of prompt-based triplet extraction using large LLMs.

SAGE-KG Performance. Our method exhibits consistent improvements across all metrics and datasets as model size increases. SAGE-KG with Qwen2.5-14B achieves the highest E-M (38.00, 27.40 for HotpotQA and 2WikiMultiHopQA, respectively) and highest G-E scores (73.40–62.38) for all three datasets, indicating that agentic extraction coupled with EGT enables accurate multi-hop reasoning and produces high-quality, semantically coherent answers. Notably, SAGE-KG outperforms KGGen and Zero-Shot GraphRAG in both HotpotQA and 2WikiMultiHopQA, demonstrating its robustness in datasets that require multi-document integration.

Impact of Retrieval Optimization. Replacing EGT with the ToG-v1 (Think-on-Graph, Version-1) system further improves MuSiQue and 2WikiMultiHopQA performance (S-R of 64.02 and G-Eg of 58.02), demonstrating that enhanced graph traversal strategies can amplify the utility of agentically extracted KGs in downstream RAG tasks.

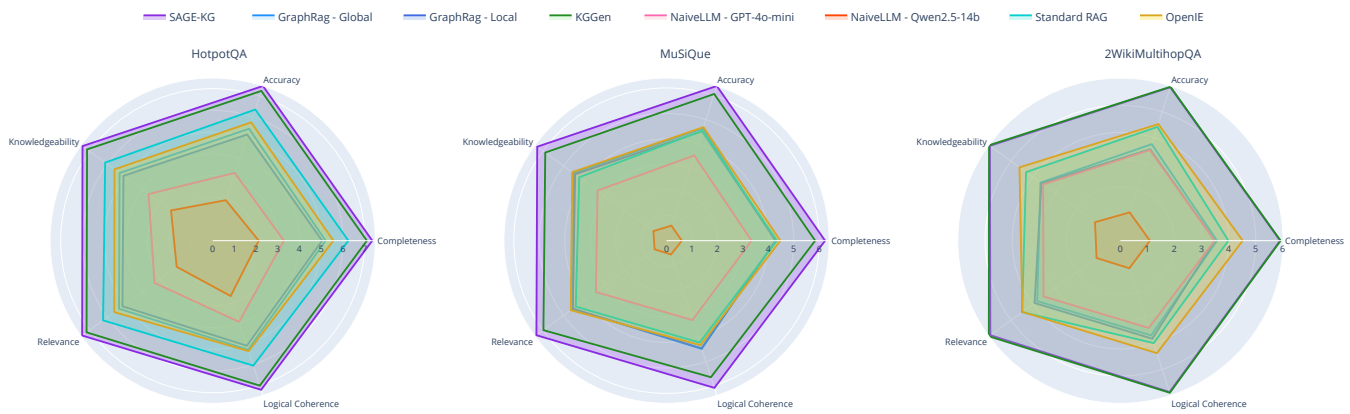
Scalability Across Model Sizes. The progressive improvement from Qwen2.5-3B to 14B confirms that larger LLMs benefit the SAGE-KG pipeline in terms of both extraction fidelity and downstream reasoning, aligning with the intrinsic evaluation trends observed earlier. Smaller models (3B, 7B) still provide competitive gains over existing baselines, making the approach practical in resource-constrained settings.

Qualitative Extrinsic Evaluation. Figure 2 presents the qualitative extrinsic evaluation of generated answers across five dimensions—completeness, accuracy, knowledgeability, relevance, and logical coherence—using two independent LLM-based judges (GPT-4o-mini and Gemini 2.0 Flash). Each radar chart corresponds to one of the three datasets (HotpotQA, MuSiQue, and 2WikiMultiHopQA) and compares our full SAGE-KG pipeline with existing baselines. Across all datasets, SAGE-KG consistently delivers higher-quality answers, demonstrating stronger reasoning fidelity and better grounding than competing methods.

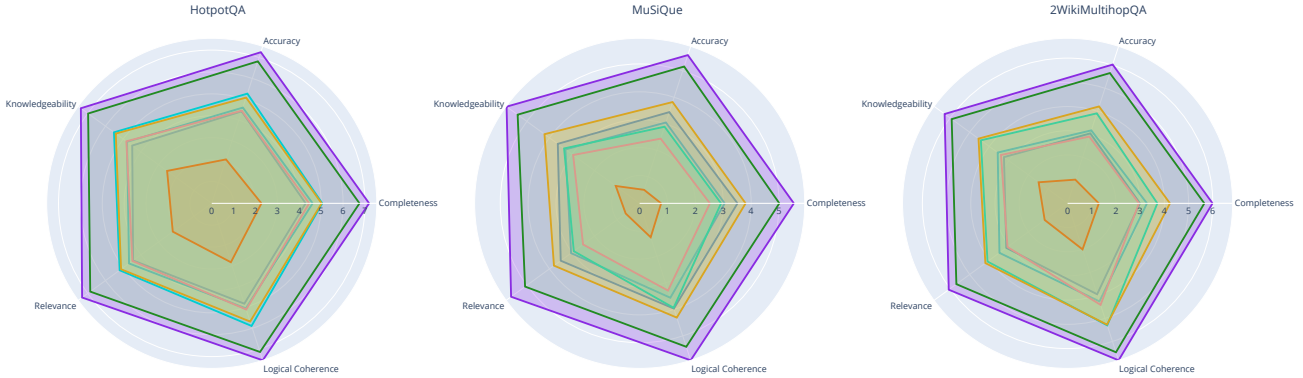
TABLE IV: Quadratic Weighted Cohen's Kappa for Different Methods Across Datasets

Methods	Datasets		
	HotpotQA	MusiQue	2WikiMultiHopQA
Baseline Methods			
Naive LLM - GPT	0.8365	0.8981	0.8480
Naive LLM - QWEN	0.8589	0.8792	0.8864
Standard RAG	0.8313	0.7657	0.7384
Microsoft GraphRAG - Global	0.8956	0.8523	0.8754
Microsoft GraphRAG - Local	0.9056	0.8691	0.8778
OpenIE	0.8937	0.8120	0.9154
KGGen	0.9248	0.8559	0.9234
Zero-Shot GraphRAG	0.9353	0.9086	0.9093
SAGE-KG (ours)			
SAGE-KG:q-3b	0.9082	0.8396	0.9084
SAGE-KG:q-7b	0.9359	0.8826	0.9108
SAGE-KG:q-14b	0.9355	0.8773	0.9355

Judge Agreement Analysis. To quantify the consistency between our two LLM-based judges (GPT-4o-mini and Gemini 2.0 Flash), we computed the quadratic weighted Cohen's Kappa across the five G-E evaluation metrics for each dataset (Table IV). The results indicate high inter-judge agreement for our agentic SAGE-KG models, with the 14B variant achieving κ scores of 0.9355 on 2WikiMultiHopQA and HotpotQA, and 0.8773 on MuSiQue. The agreement slightly decreases for smaller models (3B, 7B) and for non-agentic baselines, but remains substantial ($\kappa \geq 0.81$ in all cases), confirming that



(a) GPT-4o-mini evaluation



(b) Gemini 2.0 Flash evaluation

Fig. 2: Extrinsic evaluation of knowledge graphs based answers across five dimensions and three QA datasets using two LLM judges (GPT-4o-mini and Gemini 2.0 Flash).

judges provide reliable and consistent assessments of factual and reasoning quality across different extraction methods and datasets. This high level of concordance strengthens the validity of our extrinsic evaluation results.

In summary, the extrinsic evaluation demonstrates that our agentic SAGE-KG framework, leveraging instruction-conditioned triplet extraction and Enhanced Graph Traversal retrieval, consistently improves multi-hop question-answering performance across all datasets. The Qwen2.5-14B variant achieves the highest scores in exact match, semantic relevance, and generative evaluation metrics, outperforming both standard triplet-extraction baselines and advanced GraphRAG configurations. Importantly, the high inter-judge agreement, quantified via quadratic weighted Cohen's Kappa (up to 0.9355 for 2WikiMultiHopQA and HotpotQA), confirms the reliability and consistency of these assessments. Collectively, these findings validate that our method produces semantically rich and accurate knowledge graphs that effectively enhance downstream reasoning, achieving superior performance in both lexical fidelity and contextual reasoning tasks compared to existing state-of-the-art approaches.

VII. Conclusion

This work conducted a comprehensive intrinsic and extrinsic evaluation of SAGE-KG to assess the quality, reliability, and downstream usefulness of the generated knowledge graphs. Intrinsic evaluations demonstrated that SAGE-KG consistently produces high-precision, semantically coherent, and structurally complete triplets, outperforming existing extraction approaches across key measures such as factual accuracy, relation quality, and multi-hop consistency. These results confirm that the agentic extraction process yields knowledge graphs that are internally robust and faithful to the source text.

Extrinsic evaluations further validated the practical impact of these improvements. When used for multi-hop question answering, SAGE-KG enabled substantial gains across Exact Match, Semantic Relevance, and LLM-based Generation Evaluation metrics. Across all datasets, SAGE-KG surpassed baselines including KGGen, Microsoft GraphRAG, and OpenIE, demonstrating superior reasoning support and more effective retrieval of structured context. High inter-judge agreement, measured via quadratic weighted Cohen's Kappa, confirmed that the resulting answer quality is consistent, interpretable, and reliably assessed across independent judging models.

Together, these intrinsic and extrinsic results provide strong evidence that SAGE-KG produces knowledge graphs that are both structurally high-quality and functionally beneficial for downstream reasoning. The evaluation confirms SAGE-KG's ability to support accurate, coherent, and contextually grounded multi-hop question answering, establishing it as a reliable foundation for future work in knowledge-grounded generation and advanced retrieval-augmented reasoning.

Acknowledgment

This work emanates from research supported by a joint CSR grant from BSES Rajdhani Power Limited and BSES Yamuna Power Limited, New Delhi, under Grant Number CSR/BSES/A4-AR/SELC.

AI-Generated Content Acknowledgement

The authors acknowledge the use of AI-based tools (e.g., GPT models) for limited assistance in rephrasing manuscript text, improving code readability, and debugging. All intellectual contributions, analyses, and conclusions presented in this paper are the authors' own, and all AI-assisted content has been thoroughly reviewed and verified for accuracy and integrity.

References

[1] F. M. Suchanek, G. Kasneci, and G. Weikum, "YAGO: A Core of Semantic Knowledge," in *Proc. 16th Int. Conf. World Wide Web (WWW)*, 2007, pp. 697–706.

[2] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge," in *Proc. ACM SIGMOD Int. Conf. Management of Data*, 2008, pp. 1247–1250.

[3] G. Angeli, M. J. J. Premkumar, and C. D. Manning, "Leveraging Linguistic Structure For Open Domain Information Extraction," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2015, pp. 344–354.

[4] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural Architectures for Named Entity Recognition," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol. (NAACL-HLT)*, 2016, pp. 260–270.

[5] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, and C. D. Manning, "HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2018, pp. 2369–2380.

[6] S. Jia, Y. Xiang, X. Chen, K. Wang, and S. Wu, "Triple Trustworthiness Measurement for Knowledge Graph," in *Proc. 27th ACM Int. Conf. Inf. Knowl. Management (CIKM)*, 2018, pp. 393–402. [Online]. Available: https://doi.org/10.1145/3308558.3313586

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol. (NAACL-HLT)*, 2019, pp. 4171–4186.

[8] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, "Dense Passage Retrieval for Open-Domain Question Answering," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 6769–6781.

[9] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2020, pp. 9459–9474.

[10] X. Ho, A.-K. D. Nguyen, S. Sugawara, and A. Aizawa, "Constructing a Multi-hop QA Dataset for Comprehensive Evaluation of Reasoning Steps," in *Proc. 28th Int. Conf. Comput. Linguistics (COLING)*, 2020, pp. 6609–6625. [Online]. Available: https://aclanthology.org/2020.coling-main.580

[11] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, A.-C. N. Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab, and A. Zimmermann, "Knowledge Graphs," *ACM Comput. Surv.*, vol. 54, no. 4, pp. 1–37, Jul. 2021.

[12] H. Trivedi, N. Balasubramanian, T. Khot, and A. Sabharwal, "MuSiQue: Multihop Questions via Single-hop Question Composition," *Trans. Assoc. Comput. Linguistics*, vol. 10, pp. 539–554, 2022.

[13] Y. Zhang, X. Chen, J. Tang, and W. Zhang, "Evaluating Knowledge Graph Construction: Metrics and Benchmarks," *Inf. Process. Manag.*, vol. 59, no. 4, p. 102964, Jul. 2022.

[14] X. Li, Y. Wang, Z. Chen, J. Liu, and H. Wang, "Evaluating Knowledge Graph Extraction: Metrics and Benchmarks for Fact Extraction," in *Proc. ACM Web Conf. (WWW)*, 2022, pp. 1823–1832. [Online]. Available: https://doi.org/10.1145/3511998

[15] H. Lin, Y. Chen, and J. Zhang, "When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2023, pp. 9802–9822. [Online]. Available: https://aclanthology.org/2023.acl-long.546/

[16] A. Ji, C. Dhingra, D. Zaheer, W. W. Cohen, and Z. C. Lipton, "Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models," arXiv preprint arXiv:2309.01219, Sep. 2023. [Online]. Available: https://arxiv.org/abs/2309.01219

[17] M. Zha, Y. Wang, D. Chen, and H. Xu, "Extract, Define, Canonicalize: An LLM-based Framework for Knowledge Graph Construction," arXiv preprint arXiv:2307.15841, Jul. 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:268987666

[18] J. Zhao, W. Shao, Q. Zhu, Y. Liu, and W. Zhang, "A Survey of Large Language Models," arXiv preprint arXiv:2303.18223, Mar. 2023. [Online]. Available: https://arxiv.org/abs/2303.18223

[19] LlamaIndex Development Team, "LlamaIndex: A Data Framework for LLM Applications," 2023. [Online]. Available: https://llamaindex.ai. Accessed: Oct. 27, 2025.

[20] S. Xiao, Z. Liu, P. Zhang, and N. Muennighoff, "C-Pack: Packaged Resources To Advance General Chinese Embedding," arXiv preprint arXiv:2309.07597, Sep. 2023. [Online]. Available: https://arxiv.org/abs/2309.07597

[21] J. Wu, H. Liu, H. Wang, J. Chen, and Y. Zhang, "Zero-Shot and Few-Shot Learning With Knowledge Graphs: A Comprehensive Survey," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, 2023, doi: 10.1109/TNNLS.2023.3280101. [Online]. Available: https://ieeexplore.ieee.org/document/10144560

[22] H. Zhang, F. Zhu, L. Liu, Y. Cao, Y. Li, Z. Shao, and A. Bernstein, "KGreat: A Framework to Evaluate Knowledge Graphs via Downstream Tasks," in *Proc. 32nd ACM Int. Conf. Inf. Knowl. Management (CIKM)*, 2023, pp. 2094–2103. [Online]. Available: https://dl.acm.org/doi/10.1145/3583780.3615241

[23] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena," in *Proc. 37th Int. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2023, pp. 46595–46623. [Online]. Available: https://arxiv.org/abs/2306.05685

[24] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, and H. Wang, "Retrieval-Augmented Generation for Large Language Models: A Survey," arXiv preprint arXiv:2312.10997, Dec. 2023. [Online]. Available: https://arxiv.org/abs/2312.10997

[25] B. Jiang, Y. Xiao, K. Yan, J. Wang, and X. Chen, "HippoRAG: Neurobiologically Inspired Long-Term Memory for Large Language

Models," arXiv preprint arXiv:2405.14831, May 2024. [Online]. Available: https://arxiv.org/abs/2405.14831

[26] S. Zhang, Y. Li, X. Wang, J. Chen, and H. Liu, "AutoKG: Efficient Automated Knowledge Graph Generation for Large Language Models," arXiv preprint arXiv:2406.11638, Jun. 2024. [Online]. Available: https://arxiv.org/abs/2406.11638

[27] OpenAI, "GPT-4o System Card," OpenAI Technical Report, May 2024. [Online]. Available: https://openai.com/index/hello-gpt-4o/. Accessed: Oct. 27, 2025.

[28] H. Luo, Y. Li, X. Zhu, J. Wang, and Z. Chen, "Synergizing Retrieval-Augmented Generation and Knowledge Reasoning: A Survey," arXiv preprint arXiv:2412.15909, Dec. 2024. [Online]. Available: https://arxiv.org/abs/2412.15909

[29] S. Vats, V. Kumar, A. Sharma, and R. Gupta, "HybridContextQA: A Hybrid Knowledge and Text Retrieval Framework for Conditional Question Answering," in *Proc. 1st Workshop Bridging Gap Between Human Automated Reasoning (BGHAR 2024), CEUR Workshop Proc.*, vol. 3853, 2024, pp. 1–12.

[30] Y. Tang, H. Zhang, J. Liu, W. Chen, and X. Wang, "A Survey on Multi-Hop Question Answering and Generation," arXiv preprint arXiv:2404.09380, Apr. 2024. [Online]. Available: https://arxiv.org/abs/2404.09380

[31] N. Liu, X. Chen, Y. Zhang, J. Wang, and H. Li, "Evaluating Graph-Augmented Retrieval-Augmented Generation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2024, pp. 5342–5358.

[32] S. Zhang, Z. Chen, R. Wang, Y. Li, L. Xu, Y. Cao, and Y. Zhang, "A Comprehensive Survey on Automatic Knowledge Graph Construction," *ACM Comput. Surv.*, vol. 56, no. 8, pp. 1–62, Apr. 2024. [Online]. Available: https://dl.acm.org/doi/full/10.1145/3618295

[33] J. L. Pan, S. Razniewski, and G. Weikum, "Unifying Large Language Models and Knowledge Graphs: A Roadmap," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 7, pp. 3580–3599, Jul. 2024.

[34] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, and J. Larson, "From Local to Global: A Graph RAG Approach to Query-Focused Summarization," arXiv preprint arXiv:2404.16130, Apr. 2024. [Online]. Available: https://arxiv.org/abs/2404.16130

[35] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2019, pp. 3982–3992. [Online]. Available: https://arxiv.org/abs/1908.10084

[36] CrewAI Inc., "CrewAI: Framework for Orchestrating Role-Playing, Autonomous AI Agents," 2024. [Online]. Available: https://www.crewai.com. Accessed: Oct. 27, 2025.

[37] J. Li, Y. Wang, X. Chen, Z. Liu, and H. Zhang, "PathRAG: Learning to Reason with Retrieved Paths," arXiv preprint arXiv:2410.10453, Oct. 2024. [Online]. Available: https://arxiv.org/abs/2410.10453

[38] J. Zhang, Y. Xiao, J. Dong, C. Zhou, Q. Zhang, and D. Yin, "FlashRAG: A Modular Toolkit for Efficient Retrieval-Augmented Generation Research," arXiv preprint arXiv:2405.13576, May 2024. [Online]. Available: https://arxiv.org/abs/2405.13576

[39] M. Abdin et al., "Phi-4 Technical Report," Microsoft Research Technical Report MSR-TR-2024-37, Dec. 2024. [Online]. Available: https://arxiv.org/abs/2412.08905

[40] Y. Xiao, J. Dong, C. Zhou, S. Dong, Q.-W. Zhang, D. Yin, X. Sun, and X. Huang, "GraphRAG-Bench: Challenging Domain-Specific Reasoning for Evaluating Graph Retrieval-Augmented Generation," arXiv preprint arXiv:2501.02404, Jan. 2025. [Online]. Available: https://arxiv.org/abs/2501.02404

[41] B. Mo, K. Yu, J. Kazdan, P. Mpala, L. Yu, C. Kanatsoulis, and S. Koyejo, "KGGen: Extracting Knowledge Graphs from Plain Text with Language Models," arXiv preprint arXiv:2502.09956, Feb. 2025. [Online]. Available: https://arxiv.org/abs/2502.09956

[42] Y. Tian, Z. Yang, X. Du, J. Chen, and H. Wang, "AutoSchemaKG: Autonomous Knowledge Graph Construction through Dynamic Schema Induction from Web-Scale Corpora," arXiv preprint arXiv:2501.23628, Jan. 2025. [Online]. Available: https://arxiv.org/abs/2501.23628

[43] H. Luo, Y. Li, X. Zhu, J. Wang, Z. Chen, and Y. Zhang, "HopRAG: Multi-Hop Reasoning for Logic-Aware Retrieval-Augmented Generation," arXiv preprint arXiv:2502.12442, Feb. 2025. [Online]. Available: https://arxiv.org/abs/2502.12442

[44] Qwen Team, "Qwen2.5: A Party of Foundation Models," Qwen Technical Report, Sep. 2024. [Online]. Available: https://qwenlm.github.io/blog/qwen2.5/. Accessed: Oct. 27, 2025.

[45] A. Shah, R. Kumar, and P. Singh, "Can GPT-4o Mini and Gemini 2.0 Flash Predict Fine-Grained Fashion Product Attributes? A Zero-Shot Analysis," arXiv preprint arXiv:2507.09950, Jul. 2025. [Online]. Available: https://arxiv.org/abs/2507.09950

[46] LlamaIndex Development Team, "SentenceSplitter - LlamaIndex Documentation," LlamaIndex Docs, 2024. [Online]. Available: https://docs.llamaindex.ai/en/stable/api_reference/node_parsers/sentence_splitter/. Accessed: Oct. 27, 2025.

[47] NVIDIA Corporation, "NVIDIA RTX 6000 Ada Generation," NVIDIA Product Specification, 2024. [Online]. Available: https://www.nvidia.com/en-in/design-visualization/rtx-6000/. Accessed: Oct. 27, 2025.

[48] G. Gerganov, "llama.cpp: Inference of Meta's LLaMA Model in Pure C/C++," GitHub Repository, 2023. [Online]. Available: https://github.com/ggerganov/llama.cpp. Accessed: Oct. 27, 2025.
Note: Ollama utilizes the quantization methods (Q_K_M formats) developed in the llama.cpp project for efficient model inference.

[49] Google DeepMind, "Gemini 2.0: Our New AI Model for the Agentic Era," Google AI Blog, Dec. 2024. [Online]. Available: https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/. Accessed: Oct. 27, 2025.