SocraticKG: Knowledge Graph Construction via QA-Driven Fact Extraction

Anonymous ACL submission

Abstract

Constructing Knowledge Graphs (KGs) from unstructured text provides a structured framework for knowledge representation and reasoning, yet current LLM-based approaches struggle with a fundamental trade-off: factual coverage often leads to relational fragmentation, while premature consolidation causes information loss. To address this, we propose SocraticKG, an automated KG construction method that introduces question-answer pairs as a structured intermediate representation to systematically unfold document-level semantics prior to triple extraction. By employing 5W1H-guided QA expansion, SocraticKG captures contextual dependencies and implicit relational links typically lost in direct KG extraction pipelines, providing explicit grounding in the source document that helps mitigate implicit reasoning errors. Evaluation on the MINE benchmark demonstrates that our approach effectively addresses the coverage-connectivity trade-off, achieving superior factual retention while maintaining high structural cohesion even as extracted knowledge volume substantially expands. These results highlight that QA-mediated semantic scaffolding plays a critical role in structuring semantics prior to KG extraction, enabling more coherent and reliable graph construction in subsequent stages.

1 Introduction

As large language models (LLMs) are widely used in knowledge-intensive applications, concerns surrounding factual reliability, interpretability, and grounding have become more pronounced (Ji et al., 2023; Huang et al., 2025). While Retrieval-Augmented Generation (RAG) addresses these concerns by anchoring models to external sources, it often struggles with fragmented contexts and shallow integration of complex facts (Lewis et al., 2020; Gao et al., 2023). In response, Knowledge Graphs (KGs) have re-emerged as a complementary solution, providing a structured and verifiable backbone for explicit knowledge representation and reasoning (Pan et al., 2023; Rajabi and Etminani, 2024). However, the reliance on manual curation has historically limited the availability of domain-specific KGs, thereby motivating growing interest in automated construction methods that can scale to diverse and large-scale text sources (Ren et al., 2024).

Recent advances in LLMs have enabled more semantically grounded approaches to knowledge graph construction, moving beyond rule-based pattern matching toward methods that leverage neural reasoning to interpret unstructured text (Zhu et al., 2024). Current approaches address the construction challenge through different strategies. Some methods focus on capturing explicit factual mentions in a single pass, extracting triples directly from text (Cabot and Navigli, 2021; Shang et al., 2022; Zhang and Soh, 2024). Others adopt consolidation-centric strategies, organizing extracted facts around pre-identified entity structures to improve graph coherence (Zhong and Chen, 2021; Ye et al., 2022a; Wei et al., 2023; Mo et al., 2025).

However, these approaches face a persistent challenge: fully externalizing the narrative logic of source documents into structured graphs. The resulting knowledge graphs often struggle with a fundamental tension between factual coverage and structural coherence. Graphs may contain many facts but remain fragmented with weak semantic connectivity, or they may be well-organized yet incomplete, having filtered out contextual nuances that do not conform to predefined structures. At the core of this challenge lies the difficulty of balancing comprehensive information extraction with meaningful connectivity across the graph.

To address this limitation, we draw inspiration from how humans naturally process and organize information from text. Rather than attempting to extract structured knowledge in a single step, human comprehension is fundamentally interrogative: readers construct understanding by progressively

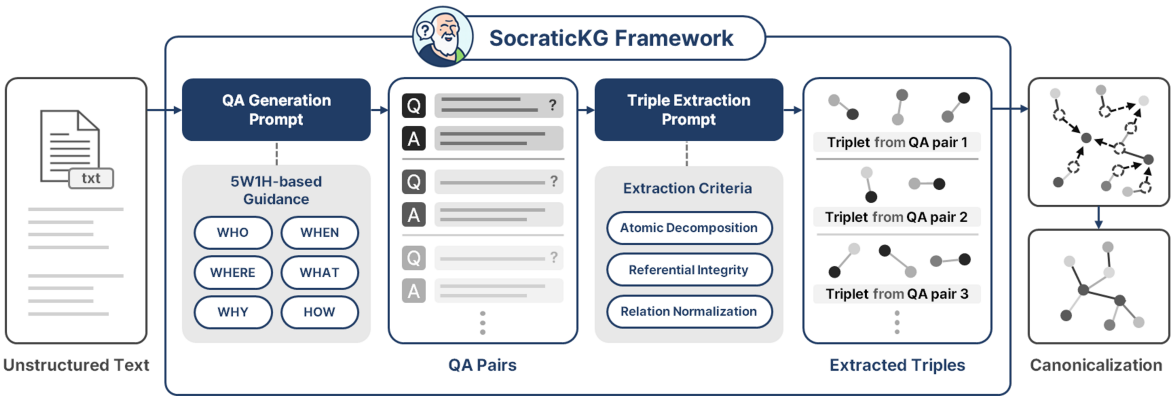

Figure 1: The overall architecture of the SocraticKG framework. Given unstructured text, the method first generates atomic QA pairs through 5W1H-guided questioning, then extracts triples from these QA pairs, and finally canonicalizes the triples to produce a cohesive knowledge graph.

clarifying salient concepts through active inquiry (Graesser and Person, 1994; Ambrose et al., 2010). This process of interrogative learning serves as a natural scaffold for organizing complex information. Question-Answering (QA), in particular, facilitates focused attention and explicit articulation of relationships that might otherwise remain implicit in direct extraction (Wu et al., 2020).

Building on this insight, we propose SocraticKG (SoKG), a method that treating QA not merely as a retrieval mechanism, but as a structured intermediate representation that systematically unfolds document-level semantics prior to graph construction (FitzGerald et al., 2018; Cohen et al., 2023). SoKG employs a structured interrogative framework based on the 5W1H framework (*who*, *what*, *when*, *where*, *why*, and *how*) to generate document-grounded QA pairs that capture key concepts, relationships, and contextual dependencies. This QA-mediated expansion articulates implicit connections and contextual nuances in explicit natural language format. The resulting intermediate representation facilitates more consistent and complete triple extraction by providing well-defined semantic units rather than requiring simultaneous resolution of semantics and structure. These extracted triples are then unified through a canonicalization process (Mo et al., 2025) that resolves surface-form variations and consolidates the graph into a coherent structure.

We evaluate our proposed method on the MINE (Measure of Information in Nodes and Edges) benchmark (Mo et al., 2025), a recently proposed benchmark designed to measure factual recoverability from automatically constructed knowledge graphs. Our results demonstrate that SocraticKG consistently outperforms state-of-the-art counterparts across multiple LLM backbones, achieving superior factual retention while producing more densely connected and less fragmented graphs. By reconciling factual coverage with structural coherence, SoKG provides a scalable approach for high-fidelity KG construction and more reliable structured reasoning.

In summary, we make the following contributions in this work:

- We propose SocraticKG, a QA-mediated method for knowledge graph construction that formalizes question-answering as a semantic scaffold for unfolding document narratives and explicitly articulating implicit connections prior to structural extraction.
- We introduce 5W1H-guided QA expansion as a systematic approach for surfacing latent dependencies typically overlooked in direct extraction, thereby improving factual coverage while reducing implicit reasoning errors.
- We demonstrate that our approach mitigates structural fragmentation and information loss, achieving superior factual retention and recoverability across various LLMs.

2 Related Work

2.1 Direct Triple Extraction

Knowledge Graph (KG) construction has evolved from conventional Open Information Extraction (OpenIE) (Etzioni et al., 2008; Fader et al., 2011) to modern approaches that extract triples directly via LLMs (Cabot and Navigli, 2021; Bi et al., 2024; Zhang and Soh, 2024). While OpenIE is

constrained by surface linguistic patterns (Niklaus et al., 2018), such direct extraction methods leverage LLM reasoning capabilities to bridge semantic gaps without explicit intermediate representations.

However, this direct extraction approach often limits the model to capturing surface-level, explicit mentions while overlooking the latent logical ties that bind them. As noted by Zhu et al. (2024); Meher et al. (2025), this approach often yields shallow factual coverage, often producing fragmented subgraphs that lack the connectivity required for effective graph-based reasoning.

2.2 Consolidation-Centric Strategies

To address the fragmentation issues, various pipelines emphasize structural coherence through post-extraction consolidation. These entity-first approaches organize extracted facts by first identifying key entities, then structuring relations around this pre-established entity framework. GraphRAG (Edge et al., 2024) builds a global index of entities and relationships partitioned into hierarchical communities for query-focused summarization, whereas KGGen (Mo et al., 2025) emphasizes clustering-based canonicalization of entities and relations to produce compact and reusable knowledge graphs. Similarly, CLARE (Henry and Gong, 2025) anchors its relational extraction on initial entity identification to ensure semantic precision within consolidated text.

Despite their effectiveness in organizing triples, these consolidation-focused strategies strategies can act as a representational bottleneck (Ye et al., 2022b). When entity sets are fixed early in the pipeline, relations or contextual dependencies that do not conform to the initial entity structure may be excluded. This sequencing effectively prioritizes structural utility over factual density, potentially under-representing the document's latent relations.

2.3 Transform-Then-Extract Approaches

To reduce extraction complexity, various approaches employ a two-stage process: first transforming raw text through intermediate representations, then extracting triples from the transformed output. Common transformation strategies include coreference resolution to handle referential expressions (Manning et al., 2014; Cetto et al., 2018) and syntactic sentence decomposition to simplify complex structures (Niklaus et al., 2019; Niklaus, 2022). CoDe-KG (Anuyah et al., 2025), for instance, leverages human-guided prompt intervention to incorporate these transformation tasks, ensuring structural clarity prior to extraction.

These transformation-based approaches operate primarily at the sentence level, focusing on local syntactic normalization rather than document-level semantic organization. While effective for resolving surface-level ambiguities within individual sentences, they do not systematically capture cross-sentence dependencies or contextual relationships that span the document. This limits their ability to externalize the broader narrative structure and global semantics required for comprehensive knowledge graph construction.

2.4 QA for Knowledge Extraction

Question-answering has been widely used to elicit structured information from text (Levy et al., 2017; Li et al., 2019; Du and Cardie, 2020), by leveraging the cognitive process of interrogative inquiry, which facilitates the construction of situation models (Chi et al., 1989; Graesser and Person, 1994). Recent extraction methods, such as StoryNet (Nagireddy, 2021) and ChatIE (Wei et al., 2023), incorporate QA-driven prompting as a core component of their extraction pipelines.

However, these approaches treat QA pairs as transient artifacts, generating and consuming them within a single extraction pass, without formalizing them as an intermediate representation for organizing document-level semantics. As a result, they lack systematic question generation and struggle to surface implicit relational and contextual dependencies prior to triple extraction.

While recent work has explored QA as an intermediate step for interpretable knowledge construction (Aneja et al., 2025), it primarily emphasizes retrieval utility through factual restatement, rather than semantic organization. Collectively, these gaps suggest that formalizing QA as a structured intermediate representation provides a more robust foundation for construction, particularly when systematic inquiry is used to proactively externalize latent relational and causal dependencies.

3 Methods

SoKG introduces QA pairs as a structured intermediate representation for LLM-based KG construction. Rather than prompting LLMs to extract triples directly from raw text, our approach first decomposes the document into explicit QA pairs that resolve contextual dependencies and referential am-

biguities in natural language. These QA pairs are then mapped to atomic triples and unified through canonicalization to produce the final KG.

3.1 5W1H-Guided QA Generation

This stage transforms document into a collection of discrete, self-contained QA pairs. To ensure comprehensive coverage of the factual content in the text, we design a prompt strategy based on two core principles: systematic questioning and contextual independence (detailed prompt in Appendix B.1).

Detailed Questioning via 5W1H We leverage the 5W1H framework to guide systematic question formulation. The LLM generates multiple questions spanning all six categories and diverse aspects of the document. As a result, the resulting QA pairs capture both surface-level entities and complex dependencies, including causal rationales (*why*) and procedural details (*how*).

Contextual Independence To ensure each QA pair functions as a standalone unit, we instruct the LLM to generate answers that are fully understandable without referencing the original source text. Specifically, the model is required to replace pronouns (*e.g., it*, *they*) with their explicit entity names, resolving referential ambiguities. This constraint prevents information loss when each QA pair is processed individually in the triple extraction phase.

3.2 Triple Extraction from QA

This stage transforms the QA pairs into structured triples by treating each pair as an independent extraction unit. By operating on these logically self-contained units allows the extraction process to focus on well-defined semantic boundaries, reducing errors common in direct extraction from long, complex texts. To achieve this, the LLM is instructed to follow three specific constraints (detailed prompt in Appendix C.1).

Atomic Decomposition The model decomposes each QA pair into separate, atomic triples, capturing fine-grained facts from both the inquiry and the response to maximize factual richness.

Entity Clarity All entities are expressed as specific noun phrases, and any triple containing ambiguous pronouns is discarded. This ensures that every extracted fact is self-contained and grounded in clear evidence.

Simplified Relations Predicates are distilled into concise verb phrases to reduce surface-form variations, facilitating subsequent canonicalization. The model is instructed to skip extraction if the relationship remains ambiguous.

3.3 Graph Construction from Triples

The final stage unifies discrete triples into a cohesive graph structure. Since extraction occurs across independent QA units, the raw set often contains redundant or synonymous mentions for the same concept. To resolve these redundancies, we adopt the canonicalization procedure from Mo et al. (2025), which combines embedding-based clustering with LLM-based refinement.

The canonicalization process is performed independently on entities and relations through a cluster-then-refine process. First, semantic embeddings are generated for all unique entities and relations using a text embedding model. To narrow the search space, these embeddings are partitioned into clusters of a manageable size for entities and relations respectively via K-means clustering. Within each cluster, the top-k potential matches for each anchor are identified by balancing dense semantic similarity with sparse lexical overlap (BM25). Finally, synonyms and abbreviations are resolved by an LLM, which maps these variants to a single representative entity or relation to consolidate fragmented triples into a cohesive, canonicalized graph.

4 Experiments

To validate the effectiveness of SoKG, we utilized the MINE benchmark (Mo et al., 2025). MINE was designed to quantify the information gap between raw text and its graph representation by measuring how much source information is recoverable. The benchmark comprises 100 diverse articles, each paired with 15 verified atomic facts, providing a rigorous evaluation framework across 1,500 independent factual instances. Following the benchmark protocol, we constructed one KG per article and evaluated each graph from two complementary perspectives: factual retention and structural characteristics.

4.1 Evaluation Metrics

Factual Retention Score As the primary metric, we measured the proportion of ground-truth facts successfully recovered from the constructed KGs.

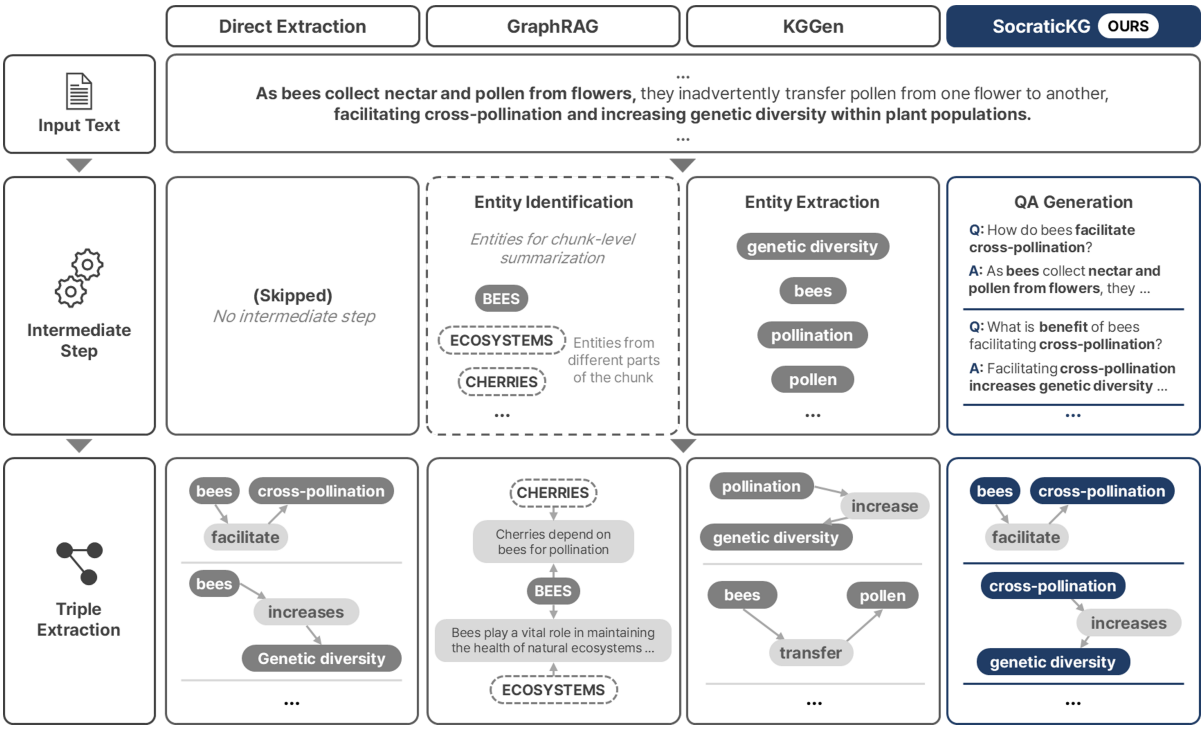

Figure 2: Comparison of extraction pipelines using an example output from Gemini-2.5-flash-lite. While baseline pipelines often miss the syntactic connection in complex sentences, failing to recover the causal link between bees and genetic diversity, SoKG leverages QA-driven reasoning to explicitly reconstruct the intermediate concept. As a result, SoKG successfully recovers the complete causal chain (bees → cross-pollination → genetic diversity), whereas baselines tend to simplify or fragment this relationship.

Following the MINE benchmark protocol, we retrieved a local subgraph for each fact, consisting of the top-8 nodes most semantically similar to the target statement and their 2-hop neighbors. An LLM-judge then determined whether the fact was logically supported by the retrieved subgraph context. The score represents the percentage of verifiable facts, reflecting how well the graph preserves information from the source text for downstream tasks such as retrieval and reasoning.

Structural Cohesion and Density To analyze the organization and coherence of the KGs, we investigated:

- **Average Degree (Deg):** The average number of unique neighboring nodes per node, capturing the local connectivity density of the graph (Barabási, 2013). It reflects how many distinct entities a node is connected to, irrespective of relation direction. Following standard practice for undirected graphs, we compute the average degree as

$$\text{Deg} = \frac{2E}{N},$$

where N denotes the number of nodes and E the number of edges.

- **Triple Count (#Tri):** The total number of atomic facts externalized in the graph.

- **Normalized Fragmentation Index (NFI):** Motivated by the notion of graph fragmentation as the decomposition of a network into disconnected components (Borgatti, 2003), we define a component-based metric as:

$$\text{NFI} = \frac{C-1}{N-1},$$

where C denotes the number of connected components and N is the total number of nodes ($N \geq 2$). This formulation normalizes fragmentation to the unit interval $[0, 1]$, where 0 corresponds to a fully connected graph ($C = 1$) and 1 indicates a completely fragmented network ($C = N$).

4.2 Comparative Analysis Design

KG Construction Methods We compared SoKG against three representative approaches in

Method	Qwen-2.5	GPT-4o-mini	GPT-4o	Gemini-2.5	Claude-4
Direct Extraction	66.5	68.5	78.1	84.6	86.8
GraphRAG	59.7	49.5	49.3	48.5	52.3
KGGen	56.7	44.3	66.4	62.5	69.1
SoKG (w/o 5W1H)	67.1	80.5	83.5	85.6	94.6
SoKG (Ours)	**73.4**	**83.9**	**89.3**	**87.7**	**96.3**

Table 1: Comparison of factual retention scores (%) on the MINE benchmark. SoKG consistently achieves the highest performance across all evaluated models. The vanilla variant (*i.e.,* SoKG w/o 5W1H) shows how the 5W1H scaffold captures procedural and causal facts to improve factual consistency even on smaller models like Qwen-2.5.

the current landscape of LLM-based KG construction. To ensure a valid comparison, we selected the comparative methods that operated in autonomous and open-domain settings without predefined schemas or human intervention. Figure 2 summarizes the procedures of these methods.

- **Direct Extraction:** A single-pass extraction strategy where triples are generated directly from raw text (Appendix C.2). For fair comparison, we apply the identical canonicalization procedure used in KGGen and SoKG to consolidate the extracted triples. It serves as a primary benchmark for the LLM's implicit reasoning capability without the benefit of intermediate semantic scaffolding.
- **GraphRAG:** A prominent solution across industry and academia for global, query-focused entity indexing. We utilize Microsoft's official implementation for hierarchical community detection and aggregation, providing a benchmark against the widely adopted text-summary-based method.
- **KGGen:** A recent state-of-the-art method focusing on entity-centric extraction and structural consolidation. It serves as a primary comparative method for evaluating factual retention and structural cohesion in open-domain.
- **SoKG (w/o 5W1H):** An ablated variant of SoKG that retains QA pairs as its intermediate representation but replaces the 5W1H-guided inquiry with generic QA. This design isolates the contribution of the 5W1H-guided scaffold to evaluate its impact.
- **SoKG (Ours):** Our proposed approach utilizing 5W1H-guided QA generation to systematically construct KGs from source documents. Unless otherwise specified, SoKG refers to this complete implementation.

Evaluation across LLMs To assess robustness across varying LLM architectures and scales, we evaluated the selected KG construction methods on five LLMs: GPT-4o, GPT-4o-mini, Gemini-2.5 (Gemini-2.5-Flash-Lite), Qwen-2.5 (Qwen2.5-7B-Instruct), and Claude-4 (Claude-4-Sonnet).

4.3 Implementation Details

For all LLMs, we set the decoding temperature to 0 to ensure reproducibility, except for GraphRAG, which follows the default stochastic configuration of its official implementation.

We adopted the canonicalization and factual retention evaluation protocol proposed by Mo et al. (2025). For the semantic clustering mentioned in Section 3.3, we partitioned entities and relations into clusters containing at most 128 elements. For the identification of potential matches, we set the candidate retrieval size to $k = 16$, which defines the number of top-ranked duplicates evaluated by the LLM. All embedding-based processes used the all-MiniLM-L6-v2 model, and factual verification was performed via an LLM-as-a-judge protocol using GPT-4o.

5 Results and Discussion

5.1 Factual Retention Performance

Table 1 summarizes the factual retention performance on the MINE benchmark. Across all compared methods and evaluated LLMs, SoKG consistently achieves the highest scores, peaking at 96.3% with Claude-4.

The comparison between Direct Extraction and SoKG (w/o 5W1H) illustrates the benefit of introducing QA as an intermediate representation. Even without 5W1H guidance, SoKG outperforms Direct Extraction on all LLM models. This advantage stems from decomposing documents into discrete, self-contained QA pairs prior to triples extraction.

Method	Qwen-2.5			GPT-4o-mini			GPT-4o			Gemini-2.5			Claude-4		
	N	E	Deg	N	E	Deg	N	E	Deg	N	E	Deg	N	E	Deg
Direct Extraction	21.7	17.3	1.60	33.5	28.1	1.69	33.9	27.4	1.62	58.4	64.1	2.20	46.4	40.8	1.77
GraphRAG	19.8	19.0	**2.00**	11.2	10.2	1.84	11.3	9.70	1.75	15.4	17.7	2.35	14.6	16.2	2.20
KGGen	28.1	22.1	1.56	19.3	16.7	1.75	33.2	28.9	1.74	38.1	43.2	2.23	57.2	58.9	2.07
SoKG (w/o 5W1H)	28.0	25.4	1.81	49.2	50.5	2.06	51.9	49.1	1.89	58.0	67.8	2.34	84.2	94.5	2.25
SoKG (Ours)	**34.9**	**34.1**	1.96	**57.9**	**62.2**	**2.16**	**62.3**	**60.5**	**1.95**	**65.7**	**80.8**	**2.47**	**104.2**	**128.4**	**2.48**

Table 2: Topological characteristics averaged over the 100 articles in the MINE benchmark. N, E, and Deg denote the mean count of Nodes, Edges, and Average Degree per graph, respectively. SoKG consistently expands the knowledge scale while maintaining high connectivity density across all backbones.

Method	Qwen-2.5		GPT-4o-mini		GPT-4o		Gemini-2.5		Claude-4	
	NFI	#Tri	NFI	#Tri	NFI	#Tri	NFI	#Tri	NFI	#Tri
Direct Extraction	0.162	1,955	0.145	3,100	0.172	2,941	0.038	7,315	0.127	4,417
GraphRAG	0.084	1,981	**0.038**	1,076	**0.083**	1,009	0.036	1,848	0.067	1,590
KGGen	0.187	2,375	0.091	1,942	0.112	3,089	0.030	5,301	0.052	6,391
SoKG (w/o 5W1H)	0.106	2,871	0.059	5,646	0.092	5,345	0.034	7,875	0.056	10,511
SoKG (Ours)	**0.078**	**3,958**	0.047	**7,069**	0.086	**6,627**	**0.023**	**9,612**	**0.039**	**14,849**

Table 3: Comparison of graph fragmentation averaged over the 100 articles (NFI; lower is better) and total extracted information volume summed over the 100 articles (#Tri). The results demonstrate that SoKG effectively resolves the trade-off between knowledge coverage and structural connectivity, maintaining high graph cohesion even as the volume of extracted facts increases.

Notably, both GraphRAG and KGGen underperform Direct Extraction in terms of factual retention. GraphRAG prioritizes hierarchical community structures and query-focused summarization over comprehensive fact preservation, resulting in lower coverage of atomic facts. KGGen's entity-first bottleneck similarly leads to fact omission when initial entity identification fails, showing inconsistent performance across models.

In contrast, SoKG with 5W1H guidance further enhances performance by systematically surfacing procedural and causal dimensions. This interrogative framework ensures that latent dependencies are explicitly captured, maintaining high factual consistency regardless of the underlying LLM model's inherent reasoning capacity.

To further validate our triple extraction strategy, we conducted additional experiments in Appendix A. By isolating the impact of the QA scaffold from the extraction strategy, these studies reveal that entity-first approaches persist as a performance bottleneck even when applied to QA-preprocessed inputs.

5.2 Graph Scale and Connectivity

The superior factual retention shown in Table 1 raises a critical question: is SoKG simply extracting more triples, or is it building a fundamentally better graph? To address this, we examine graph scale and connectivity in Table 2.

SoKG significantly expands graph scale while maintaining or improving connectivity density across all evaluated LLMs. In contrast, Direct Extraction produces smaller graphs with lower connectivity, while GraphRAG generates compact structures that sacrifice comprehensive fact coverage for hierarchical organization. The comparison between SoKG (w/o 5W1H) and SoKG reveals that 5W1H guidance substantially increases the number of extracted entities and relations while enhancing connectivity density. This indicates that 5W1H systematically surfaces additional facts without fragmenting the graph structure.

Importantly, SoKG achieves higher connectivity than both Direct Extraction and KGGen despite using the same canonicalization procedure. This confirms that the structural advantage originates from the QA-mediated intermediate representation, enabling relevant evidence to co-locate within 2-hop neighborhoods and directly supporting the high fact recoverability in Table 1.

5.3 Factual Volume and Structural Cohesion

To further examine the relationship between knowledge coverage and graph fragmentation, we analyze triple counts and the NFI in Table 3. SoKG

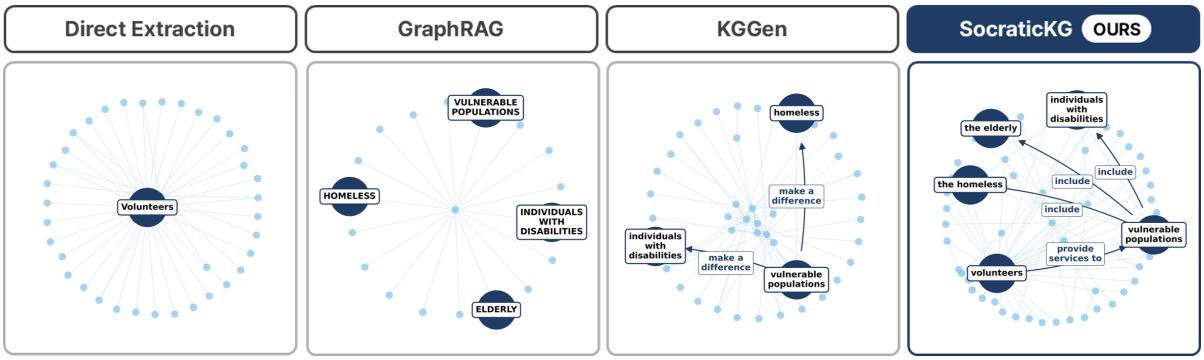

Figure 3: Comparison of extracted graphs for the example sentence: *"Volunteers provide essential services and support to vulnerable populations, such as the homeless, the elderly, and individuals with disabilities."* The nested relational path implied by this text (*Volunteers* → *Vulnerable Populations* → {*Homeless, Elderly, Individuals with disabilities*}) is emphasized to assess relational completeness. Specifically, nodes corresponding to this path are enlarged for clear visibility, connected by thick dark blue arrows to indicate the sequence of triples, while the remaining background graph elements are displayed in light blue.

substantially expands knowledge volume while simultaneously reducing fragmentation across all evaluated LLMs. While alternative methods either limit fact extraction (GraphRAG) or exhibit higher fragmentation (KGGen and Direct Extraction), SoKG extracts substantially more triples while maintaining lower NFI values.

The comparison between SoKG (w/o 5W1H) and SoKG (Ours) further illustrates the effectiveness of 5W1H guidance. Adding 5W1H consistently increases triple extraction volume while reducing or maintaining similar fragmentation levels. This pattern indicates that 5W1H not only surfaces additional facts but also enhances their integration into the graph structure.

5.4 Qualitative Analysis

The cases in Figures 2 and 3 provide concrete examples of how SoKG's interrogative process resolves the structural deficiencies and information loss observed in alternative methods.

Figure 2 illustrates SoKG's capacity to preserve logical coherence in complex participle phrases, such as *facilitating cross-pollination*. While other approaches produce fragmented or oversimplified triples, SoKG articulates mediating concepts to ensure a cohesive causal chain: *bees* → *cross-pollination* → *genetic diversity*.

Similarly, Figure 3 demonstrates how SoKG resolves relational fragmentation in nested entity structures. Alternative methods often omit key nodes or overlook entities like *the elderly*, whereas SoKG fully reconstructs the relational tree by identifying all key entities and linking them via precise predicates such as *provide services to* and *include*.

These examples demonstrate that QA-mediated semantic scaffolding, guided by 5W1H inquiry, systematically addresses both causal reconstruction and relational completeness, enabling more structured knowledge extraction.

6 Conclusion

We present SoKG, LLM-based KG construction method that uses QA pairs as a structured intermediate representation for document-level semantic expansion prior to triple extraction. By employing 5W1H-guided QA generation, SoKG resolves referential ambiguities and surfaces implicit relational dependencies, ensuring that subsequent structural mapping is grounded in explicit, contextualized entities rather than underspecified inferences.

Evaluation on the MINE benchmark demonstrates that SoKG achieves superior factual coverage while simultaneously improving structural cohesion across diverse LLMs. This performance stems from the QA-mediated scaffold, which systematically externalizes latent causal and relational dependencies that enhance graph connectivity even as the volume of extracted facts increases.

Our findings indicate that explicit semantic organization through QA generation is not merely an auxiliary preprocessing step but a critical component for maintaining graph fidelity in LLM-based construction. By addressing the inherent trade-off between factual coverage and structural connectivity, SoKG provides a more reliable foundation for document-grounded knowledge representation and structured reasoning.

Limitations

While SoKG enables the construction of dense knowledge graphs, the QA-mediated pipeline naturally involves higher token consumption and latency than direct triple extraction. We prioritize factual density over cost optimization, though efficiency remains a target for future refinement. Furthermore, as the graph quality depends on the reasoning depth of the underlying LLM, performance may vary in domains requiring highly specialized interrogative logic.

Regarding graph representation, our current use of binary triples may simplify multidimensional qualifiers (*e.g.,* temporal or spatial data) that could be more compactly encoded via n-ary relations. Finally, our evaluation focuses on factual recoverability through the MINE benchmark. While this aligns with our objective of preserving document-level semantics, other dimensions—such as schema-alignment and relation-type fidelity—are left as promising avenues for the community to explore as KG evaluation standards evolve.

Ethical Considerations

This study utilizes the publicly available MINE benchmark and LLMs. We acknowledge that the benchmark and underlying LLMs may possess inherent biases, which could be reflected in the constructed graphs. Additionally, automated extraction carries a risk of hallucinating facts not present in source documents. We recommend human verification and validation for applications in sensitive or high-stakes domains.

References

Susan A Ambrose, Michael W Bridges, Michele DiPietro, Marsha C Lovett, and Marie K Norman. 2010. *How learning works: Seven research-based principles for smart teaching*. John Wiley & Sons.

Kartikeya Aneja, Manasvi Srivastava, Subhayan Das, and Nagender Aneja. 2025. Interpretable question answering with knowledge graphs. *arXiv preprint arXiv:2510.19181*.

Sydney Anuyah, Mehedi Mahmud Kaushik, Sri Rama Krishna Reddy Dwarampudi, Rakesh Shiradkar, Arjan Durresi, and Sunandan Chakraborty. 2025. Automated knowledge graph construction using large language models and sentence complexity modelling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, page 15526–15550. Association for Computational Linguistics.

Albert-László Barabási. 2013. Network science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1987):20120375.

Zhen Bi, Jing Chen, Yinuo Jiang, Feiyu Xiong, Wei Guo, Huajun Chen, and Ningyu Zhang. 2024. Codekgc: Code language model for generative knowledge graph construction. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(3):1–16.

Stephen P Borgatti. 2003. *The key player problem.* na.

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. Rebel: Relation extraction by end-to-end language generation. In *Findings of the association for computational linguistics: emnlp 2021*, pages 2370–2381.

Matthias Cetto, Christina Niklaus, André Freitas, and Siegfried Handschuh. 2018. Graphene: a context-preserving open information extraction system. *arXiv preprint arXiv:1808.09463*.

Michelene TH Chi, Miriam Bassok, Matthew W Lewis, Peter Reimann, and Robert Glaser. 1989. Self-explanations: How students study and use examples in learning to solve problems. *Cognitive science*, 13(2):145–182.

William W Cohen, Wenhu Chen, Michiel De Jong, Nitish Gupta, Alessandro Presta, Pat Verga, and John Wieting. 2023. Qa is the new kr: Question-answer pairs as knowledge bases. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15385–15392.

Xinya Du and Claire Cardie. 2020. Event extraction by answering (almost) natural questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 671–683.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S Weld. 2008. Open information extraction from the web. *Communications of the ACM*, 51(12):68–74.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 1535–1545.

Nicholas FitzGerald, Julian Michael, Luheng He, and Luke Zettlemoyer. 2018. Large-scale qa-srl parsing. *arXiv preprint arXiv:1805.05377*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1).

Arthur C Graesser and Natalie K Person. 1994. Question asking during tutoring. *American educational research journal*, 31(1):104–137.

Ryan Henry and Jiaqi Gong. 2025. Clare: Context-aware, interactive knowledge graph construction from transcripts. *Information*, 16(10):866.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. *arXiv preprint arXiv:1706.04115*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Xiaoya Li, Fan Yin, Zijun Sun, Xiayu Li, Arianna Yuan, Duo Chai, Mingxin Zhou, and Jiwei Li. 2019. Entity-relation extraction as multi-turn question answering. *arXiv preprint arXiv:1905.05529*.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Dipak Meher, Carlotta Domeniconi, and Guadalupe Correa-Cabrera. 2025. Link-kg: Llm-driven coreference-resolved knowledge graphs for human smuggling networks. *arXiv preprint arXiv:2510.26486*.

Belinda Mo, Kyssen Yu, Joshua Kazdan, Joan Cabezas, Proud Mpala, Lisa Yu, Chris Cundy, Charilaos Kanatsoulis, and Sanmi Koyejo. 2025. Kggen: Extracting knowledge graphs from plain text with language models. *arXiv preprint arXiv:2502.09956*.

Srichakradhar Reddy Nagireddy. 2021. *StoryNet: A 5W1H-Based Knowledge Graph to Connect Stories*. University of Missouri-Kansas City.

Christina Niklaus. 2022. *From complex sentences to a formal semantic representation using syntactic text simplification and open information extraction*. Springer Nature.

Christina Niklaus, Matthias Cetto, André Freitas, and Siegfried Handschuh. 2018. A survey on open information extraction. *arXiv preprint arXiv:1806.05599*.

Christina Niklaus, Matthias Cetto, André Freitas, and Siegfried Handschuh. 2019. Transforming complex sentences into a semantic hierarchy. *arXiv preprint arXiv:1906.01038*.

Jeff Z. Pan, Simon Razniewski, Jan-Christoph Kalo, Sneha Singhania, Jiaoyan Chen, Stefan Dietze, Hajira Jabeen, Janna Omeliyanenko, Wen Zhang, Matteo Lissandrini, Russa Biswas, Gerard de Melo, Angela Bonifati, Edlira Vakaj, Mauro Dragoni, and Damien Graux. 2023. Large Language Models and Knowledge Graphs: Opportunities and Challenges. *Transactions on Graph Data and Knowledge*, 1(1):2:1–2:38.

Enayat Rajabi and Kobra Etminani. 2024. Knowledge-graph-based explainable ai: A systematic review. *J. Inf. Sci.*, 50(4):1019–1029.

Xubin Ren, Jiabin Tang, Dawei Yin, Nitesh Chawla, and Chao Huang. 2024. A survey of large language models for graphs. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6616–6626.

Yu-Ming Shang, Heyan Huang, and Xianling Mao. 2022. Onerel: Joint entity and relation extraction with one module in one step. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11285–11293.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, and 1 others. 2023. Chatie: Zero-shot information extraction via chatting with chatgpt. *arXiv preprint arXiv:2302.10205*.

Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. 2020. Corefqa: Coreference resolution as query-based span prediction. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 6953–6963.

Deming Ye, Yankai Lin, Peng Li, and Maosong Sun. 2022a. Packed levitated marker for entity and relation extraction. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 4904–4917.

Hongbin Ye, Ningyu Zhang, Hui Chen, and Huajun Chen. 2022b. Generative knowledge graph construction: A review. *arXiv preprint arXiv:2210.12714*.

Bowen Zhang and Harold Soh. 2024. Extract, define, canonicalize: An LLM-based framework for knowledge graph construction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9820–9836. Association for Computational Linguistics.

Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 50–61.

Yuqi Zhu, Xiaohan Wang, Jing Chen, Shuofei Qiao, Yixin Ou, Yunzhi Yao, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2024. Llms for knowledge graph construction and reasoning: Recent capabilities and future opportunities. *World Wide Web*, 27(5):58.

A Ablation Studies

Prompt Archetype	Qwen-2.5		GPT-4o-mini		GPT-4o		Gemini-2.5		Claude-4	
	w/o 5W1H	Full	w/o 5W1H	Full	w/o 5W1H	Full	w/o 5W1H	Full	w/o 5W1H	Full
Role-Oriented (RO)	67.1	**73.4**	80.5	**83.9**	83.5	**89.3**	85.6	**87.7**	94.6	**96.3**
Procedural-Step (PS)	60.7	**64.5**	79.6	**83.0**	82.1	**87.1**	86.7	**87.3**	93.5	**95.8**
Instructional-Direct (ID)	65.3	**68.6**	77.9	**83.5**	79.7	**81.2**	83.9	**88.7**	91.3	**96.3**
Average	64.4	**68.8**	79.3	**83.5**	81.8	**85.9**	85.4	**87.9**	93.1	**96.1**

Table 4: Effectiveness of 5W1H-guided expansion across prompt archetypes. Scores represent factual retention (%) on the MINE benchmark. The results demonstrate that the 5W1H framework is a robust cognitive guide independent of stylistic framing.

Method	Qwen-2.5	GPT-4o-mini	GPT-4o	Gemini-2.5	Claude-4
KGGen	56.7	44.3	66.4	62.5	69.1
SoKG-EF	72.1	79.7	76.9	87.0	92.5
SoKG (Ours)	**73.4**	**83.9**	**89.3**	**87.7**	**96.3**

Table 5: Ablation study on input representation and triple extraction strategy. While SoKG-EF demonstrates the foundational impact of the QA scaffold, SoKG achieves peak performance by removing the entity-first bottleneck to maximize factual retention across all LLMs.

A.1 Robustness of Prompt Designs

To evaluate the contribution of the 5W1H framework, Table 4 compares the full 5W1H-integrated pipeline (Full) against the version omitting 5W1H-guided expansion (w/o 5W1H) across three distinct prompt archetypes:

- **Role-Oriented (RO):** Assigns a specific persona (e.g., Knowledge Archivist) and uses 5W1H as analytical lenses to guide deep exploration. This prompt design was adopted as the primary setting for our main experiments.
- **Procedural-Step (PS):** Defines a systematic workflow (Read $\rightarrow$ Segment $\rightarrow$ Generate) to ensure atomic factual extraction.
- **Instructional-Direct (ID):** Employs standard task-based instructions without complex role-play or multi-step procedures.

As shown in Table 4, the 5W1H framework provides a universal performance lift across all LLMs regardless of the underlying prompt structure. While the RO archetype generally yields the highest retention, peaking at 96.3% with Claude-4, even the more concise PS and ID templates show significant improvements once the interrogative scaffold is present. These results indicate that the 5W1H constraint functions as a fundamental cognitive guide that systematically surfaces procedural and causal dimensions.

A.2 The Entity-First Constraint

We evaluate the respective impacts of the QA scaffold and extraction strategy by comparing three configurations: KGGen, SoKG-EF (Entity-First), and SoKG. SoKG-EF incorporates both the extraction and consolidation logic of KGGen, applying this entity-centric pipeline to our QA-mediated scaffold. This setup allows us to evaluate the benefit of the scaffold independently while preserving the underlying entity-first logic.

As shown in Table 5, the superior performance of SoKG-EF over KGGen confirms that a QA scaffold effectively mitigates the complexity of raw text. However, SoKG's even greater success reveals that rigid entity-first filtering acts as a restrictive bottleneck, limiting the model's ability to capture full relational depth even when supported by a comprehensive QA scaffold.

Although intermediate stages introduce overhead, this rich interrogative structure justifies the investment by providing a dense semantic foundation for superior factual recoverability. Unlike isolated entity extraction which often incurs information loss through restrictive filtering, the QA-mediated scaffold preserves a richer semantic context. These results demonstrate that for structural refinement, a QA-driven approach offers a more systematic and inclusive foundation for knowledge construction than traditional entity-centric methods.

B QA Generation Prompt Details

B.1 Role-Oriented (RO), w/ 5W1H

```
## ROLE
You are a **Comprehensive Knowledge Archivist** who converts the [Full Document] into detailed,
    document-grounded QA pairs.

## OBJECTIVE
Extract as many meaningful Question-Answer pairs as possible from the document.
Use the 5W1H perspectives (Who, What, When, Where, Why, How) **as analytical lenses** to help you
    identify and expand potential questions, but do NOT restrict yourself to producing only
    5W1H-type questions.
Your goal is to maximize informational coverage, capturing every explicit fact, relation, event,
    definition, rationale, and process described in the document.

## INPUT
Full Document: "{document_text}"

## CONSTRAINTS
1. **Context-Independent**
   - Each QA must be self-contained and understandable without referencing the original text.
   - Replace pronouns with explicit entities.

2. **No Hallucination**
   - Use only facts explicitly stated in the document.

3. **Expansion-Oriented Thinking**
   - For each sentence or factual unit, consider the 5W1H perspectives as prompts to explore:
     - WHO is involved?
     - WHAT happened or is described?
     - WHEN did it occur?
     - WHERE did it occur?
     - WHY did it occur?
     - HOW was it carried out?
   - These perspectives are **guides** to inspire multiple possible QA pairs, even if they are
     implicit or only partially expressed.

4. **Coverage**
   - Extract all possible QA pairs that can be reasonably derived from the document.

## OUTPUT FORMAT
Return a JSON list of QA objects:

[
  {{"question": "...", "answer": "..."}},
  ...
]
```

B.2 Role-Oriented (RO), w/o 5W1H

```
## ROLE
You are a **Comprehensive Knowledge Archivist** who converts the [Full Document] into precise and
    meaningful QA pairs.

## OBJECTIVE
Extract as many high-quality Question-Answer pairs as needed to fully represent the document's
    explicit information.
Use the following analytical perspectives as guides to discover potential questions, but do NOT
    restrict your output to only these categories:

1. **Entities & Definitions** - Identify and clarify key terms, objects, roles, or concepts.
2. **Properties & Characteristics** - Extract notable features, attributes, components, or
    qualities.
3. **Events & Stated Facts** - Capture actions, processes, or explicit factual statements.
4. **Relationships & Dependencies** - Identify connections, comparisons, or dependencies between
    entities or ideas.

These perspectives are **guides for expanding coverage**, not mandatory categories.

## INPUT
Full Document: "{document_text}"

## CONSTRAINTS
1. **Context-Independent**
   - Each QA must be self-contained and understandable without referencing the original text.
   - Replace pronouns with explicit entities when needed.

2. **No Hallucination**
   - Use only facts explicitly stated in the document.

3. **Coverage without Inflation**
   - Extract all meaningful QA pairs that can be reasonably derived from the document.

## OUTPUT FORMAT
Return a JSON list:

[
  {{"question": "...", "answer": "..."}},
  ...
]
```

B.3 Procedural-Step (PS), w/ 5W1H

```
## ROLE
You are a **Document-Grounded QA Extractor**.

## OBJECTIVE
Convert the full document into high-coverage, explicit-fact QA pairs.

## PROCEDURE
1. Read the document end-to-end.
2. Segment into atomic factual units.
3. For each unit:
   - Generate QAs that capture all explicit information it contains.
   - When forming questions, view the unit through the 5W1H angles (Who, What, When, Where, Why,
     How) so that different aspects of the same fact can be covered.
4. Merge duplicates and keep the most precise wording.

## INPUT
Full Document: "{document_text}"

## CONSTRAINTS
- Context-Independent QAs only.
- No Hallucination.
- Prefer concise but complete answers.

## OUTPUT FORMAT
Return a JSON list:
[
  {{"question": "...", "answer": "..."}},
  ...
]
```

B.4 Procedural-Step (PS), w/o 5W1H

```
## ROLE
You are a **Document-Grounded QA Extractor**.

## OBJECTIVE
Convert the full document into high-coverage, explicit-fact QA pairs.

## PROCEDURE
1. Read the document end-to-end.
2. Segment into atomic factual units.
3. For each unit, generate QAs that capture all explicit information it contains.
4. Merge duplicates and keep the most precise wording.

## INPUT
Full Document: "{document_text}"

## CONSTRAINTS
- Context-Independent QAs only.
- No Hallucination.
- Prefer concise but complete answers.

## OUTPUT FORMAT
Return a JSON list:
[
  {{"question": "...", "answer": "..."}},
  ...
]
```

B.5 Instructional-Direct (ID), w/ 5W1H

```
Read the following document and generate question-answer pairs based on its content.
Generate as many high-quality questions as needed to cover the information explicitly stated in
    the document.
For the same piece of information, consider the 5W1H dimensions (Who, What, When, Where, Why, How)
    and generate separate questions whenever different aspects are supported by the text.
Do not stop at a single question if multiple 5W1H aspects can be identified.
If different parts of the document support different questions, include all of them.
Each question should be answerable using information explicitly stated in the document and written
    in a clear and self-contained manner.

Input Document:
"{document_text}"

Output Format:
Return a JSON list of objects in the following form:
[
  {{"question": "...", "answer": "..."}},
  ...
]
```

B.6 Instructional-Direct (ID), w/o 5W1H

```
Read the following document and generate question-answer pairs based on its content.
Generate as many high-quality questions as needed to cover the information explicitly stated in
    the document.
If different parts of the document support different questions, include all of them.
Each question should be answerable using information explicitly stated in the document and written
    in a clear and self-contained manner.

Input Document:
"{document_text}"

Output Format:
Return a JSON list of objects in the following form:
[
  {{"question": "...", "answer": "..."}},
  ...
]
```

C Triple Extraction Prompt Details

C.1 Triple Extraction from QA Pairs

```
## ROLE
You are a Semantic Knowledge Graph Builder.
Extract every structured triples (entity1, relation, entity2) from the Q&A pair, following the
    rules below.

## GOAL
From the question-answer pair, extract only useful, knowledge-ready triples that can serve as
    entries in a semantic knowledge graph.

## RULES
Extract clean (subject, relation, object) triples following the rules:

1. Split every stated or clearly implied fact into minimal triples; integrate question and answer
    context when needed.

2. Entities (entity1, entity2) must be short, concrete noun phrases.
   - No pronouns (this, that, it, its, these, those, etc.).
   - Entities must not be unresolved or reference-based pronouns (\eg those, they, someone,
    anyone, whoever); if such a pronoun appears, rewrite it into a specific, explicit noun phrase
    or skip the triple.
   - No clauses or relative clauses (no "who/that/which/what/as it ..." inside an entity).
   - No long gerund or sentence-like phrases. If a phrase contains a verb or clause marker,
    rewrite it into a concise noun concept or skip the triple.

3. Relations must be short, canonical verbs or verb phrases.
   - Express a single semantic link between the two entities (\eg causes, leads to, supports,
    believes, opposes).
   - Must be a compact predicate, not a sentence fragment.
   - No pronouns or clause markers inside the relation (no "its", "that", "as it", "what", etc.).
   - If the source uses an idiomatic or long expression, rewrite it into a simple canonical
    relation without pronouns or embedded clauses, or skip the triple.

4. Include a fact if it can be clearly rewritten into a concise, explicit triple that fits the
    rules above; otherwise skip it.

5. Output only concise, interpretable, knowledge-ready triples.

## INPUT
Q: {question}
A: {answer}

## OUTPUT FORMAT (JSON List)
- Return a list of JSON objects.
- Return [] if no valid triples exist.

[
  {{"entity1": "Specific_Noun", "relation": "precise_verb_phrase", "entity2": "Specific_Noun"}}
]
```

C.2 Triple Extraction from Raw Text (Direct Extraction)

```
## ROLE
You are a Semantic Knowledge Graph Builder.
Extract every structured triples (entity1, relation, entity2) from the text, following the rules
    below.

## GOAL
From the given text, extract only useful, knowledge-ready triples that can serve as entries in a
    semantic knowledge graph.

## RULES
Extract clean (subject, relation, object) triples following the rules:

1. Split every stated or clearly implied fact into minimal triples.

2. Entities (entity1, entity2) must be short, concrete noun phrases.
   - No pronouns (this, that, it, its, these, those, etc.).
   - Entities must not be unresolved or reference-based pronouns (\eg those, they, someone,
     anyone, whoever); if such a pronoun appears, rewrite it into a specific, explicit noun phrase
     or skip the triple.
   - No clauses or relative clauses (no "who/that/which/what/as it ..." inside an entity).
   - No long gerund or sentence-like phrases. If a phrase contains a verb or clause marker,
     rewrite it into a concise noun concept or skip the triple.

3. Relations must be short, canonical verbs or verb phrases.
   - Express a single semantic link between the two entities (\eg causes, leads to, supports,
     believes, opposes).
   - Must be a compact predicate, not a sentence fragment.
   - No pronouns or clause markers inside the relation (no "its", "that", "as it", "what", etc.).
   - If the source uses an idiomatic or long expression, rewrite it into a simple canonical
     relation without pronouns or embedded clauses, or skip the triple.

4. Include a fact if it can be clearly rewritten into a concise, explicit triple that fits the
    rules above; otherwise skip it.

5. Output only concise, interpretable, knowledge-ready triples.

## INPUT
Text: {document_text}

## OUTPUT FORMAT (JSON List)
- Return a list of JSON objects.
- Return [] if no valid triples exist.

[
  {{"entity1": "Specific_Noun", "relation": "precise_verb_phrase", "entity2": "Specific_Noun"}}
]
```