The Knowledge Engineering Review

RESEARCH ARTICLE

Open Access

https://doi.org/10.48130/ker-0026-0017

The Knowledge Engineering Review **2026**, 41: e016

AutoQG: automated evidence-traceable QA via ontology-guided KG

Xiaoli Shen[1], Guilin Qi[1*], Peng Zhang[2], Dingsheng Yang[1], Yang Yang[3], Muyan Yang[1], Haowen Guo[1] and Huikang Hu[1]

[1] *School of Computer Science and Engineering, Southeast University, Nanjing, Jiangsu 211189, China*
[2] *School of Life Science and Technology, Southeast University, Nanjing, Jiangsu 211189, China*
[3] *School of Artificial Intelligence, Southeast University, Nanjing, Jiangsu 211189, China*
* Correspondence: gqi@seu.edu.cn (Qi G)

Abstract

Large Language Models (LLMs) present unprecedented opportunities for generating scientific questions. However, existing approaches face two key limitations: heavy reliance on costly human annotations and the production of brittle, unverifiable outputs. To address these challenges, we propose **AutoQG**, a fully automated multi-agent framework for evidence-grounded scientific QA generation. AutoQG comprises three complementary agents: (i) **KG Extraction Agent**, which performs ontology-guided knowledge graph construction with section-aware prompts for precise information retrieval; (ii) **KG Evaluation Agent**, a multi-dimensional evaluation module with iterative refinement to ensure accuracy and consistency; and (iii) **QA Generation Agent**, which produces schema-constrained QA pairs grounded in reasoning paths and explicit textual evidence. Applied to over 4,000 scientific papers, AutoQG constructs 243k triples and introduces *AutoQG*^{20k}, a benchmark containing more than 20,000 QA pairs. Each pair is explicitly linked to its reasoning chains and supporting evidence, ensuring transparency and verifiability. We further release *AutoQG*^{7k}, a challenging subset designed with hard questions that strong LLMs struggle to answer. Extensive experiments demonstrate that AutoQG consistently outperforms strong baselines in both human evaluation and LLM-as-a-judge assessments. By transforming LLM output into a controlled and auditable pipeline, AutoQG advances evidence-based AI for the understanding of reliable scientific knowledge.

Keywords: Question generation, Knowledge graph, Evidence traceability, Multi-agent systems, Scientific document understanding

Citation: Shen X, Qi G, Zhang P, Yang D, Yang Y, et al. 2026. AutoQG: automated evidence-traceable QA via ontology-guided KG. *The Knowledge Engineering Review* 41: e016 https://doi.org/10.48130/ker-0026-0017

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in text comprehension and knowledge extraction, presenting transformative opportunities for navigating the expansive and complex domain of scientific literature[1–3]. However, applying these models to scientific texts reveals a critical structural challenge: scientific knowledge is inherently distributed across document sections. A central claim introduced in the abstract is typically elaborated in the methods section and validated by empirical evidence in the results. Existing approaches frequently fail to synthesize this cross-sectional information into coherent reasoning chains, producing question-answering (QA) pairs that lack traceability and robust evidentiary grounding[4,5]. Without a principled framework to constrain their generative behavior, LLM outputs degrade into shallow factual recall or unverifiable claims, falling short of the methodological rigor and auditability required by scientific inquiry[6]. This gap underscores the urgent need for frameworks that transform LLM generative power into a controlled, verifiable, and evidence-grounded process tailored to produce high-fidelity scientific QA pairs with explicit reasoning chains and traceable evidence.

Recent advances in scientific question generation have been shaped by a fundamental trade-off between quality and scalability. Manual annotation by domain experts yields high-quality benchmarks but remains prohibitively costly and difficult to scale[7,8]. Conversely, automated methods achieve scalability at the expense of depth, often producing shallow QA pairs that prioritize factual recall over higher-order reasoning and lack verifiable evidence links[9,10]. Even recent LLM-based frameworks have not fully resolved this tension, as they seldom accommodate the structured, evidence-centric nature of scientific discourse[11,12].

To bridge this gap, we propose that an intermediate structured representation is essential—one that explicitly models reasoning paths and evidentiary links within scientific documents. Knowledge graphs (KGs) serve as an ideal scaffold for this purpose[13]. By transforming unstructured text into a network of entities and their semantic relationships[14], a KG provides an auditable backbone for generating complex, multi-hop reasoning questions[15] that are explicitly traceable to source text[16]. Figure 1 illustrates how a structured knowledge graph serves as the foundation for generating evidence-traceable QA pairs from scientific papers.

We introduce **AutoQG**, an automated multi-agent framework for evidence-traceable question generation via ontology-guided knowledge graph construction. AutoQG comprises three specialized agents: a **KG Extraction Agent** that constructs domain-aligned knowledge graphs using section-aware prompts and ontology constraints; a **KG Evaluation Agent** that performs multi-dimensional evaluation and provides diagnostic feedback in an iterative refinement loop to progressively improve graph accuracy and consistency; and a **QA Generation Agent** that identifies multi-hop reasoning paths and synthesizes them into coherent, evidence-grounded QA pairs through schema-constrained prompting. By enforcing structured intermediate representations and explicit reasoning paths, AutoQG addresses both the scalability challenge of manual annotation and the quality degradation inherent in purely end-to-end generative approaches.

Building upon this structured knowledge representation, we propose a comprehensive solution to automate and evaluate the generation of scientific QA pairs. In summary, our key contributions are as follows:

- We propose AutoQG, an automated multi-agent framework that enables high-fidelity scientific QA generation by integrating ontology-guided and section-aware KG extraction, multi-dimensional evaluation, and schema-constrained QA generation.

© The Author(s)

https://doi.org/10.48130/ker-0026-0017

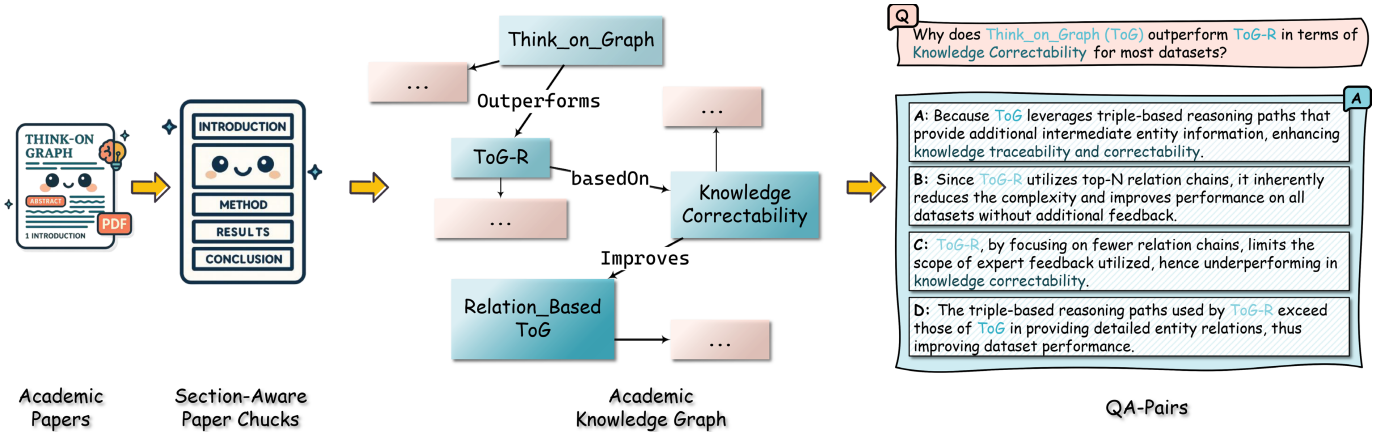


Fig. 1 An illustrative overview of our pipeline. The process begins with an academic paper, which is segmented into section-aware chunks and transformed into a structured knowledge graph capturing the paper's core entities and relations. From this graph, we derive evidence-traceable QA pairs by linking each question and answer directly to its underlying reasoning path and supporting text.

- Our framework generates two large-scale, evidence-traceable QA benchmarks derived from over 4,000 scientific papers: *AutoQG*^{20k}, containing more than 20 k QA pairs explicitly linked to reasoning paths and supporting evidence, and *AutoQG*^{7k}, which focuses on challenging scientific questions that current LLMs struggle to answer.
- Through extensive experiments combining human and LLM-as-a-judge evaluations, we demonstrate that AutoQG consistently outperforms strong baselines, achieves stable cross-domain generalization, supports low-cost open-source deployment, and delivers substantial downstream supervised fine-tuning (SFT) gains. Ablation studies further quantify the significant contribution of each framework component.

2 Related work

2.1 Scientific comprehension benchmarks

Early scientific QA benchmarks such as emrQA and PubMedQA[17,18] compile large-scale corpora from clinical and biomedical documents but fragment papers into isolated passages, neglecting the cross-sectional dependencies critical for deeper scientific reasoning. Subsequent benchmarks including ScienceQA, SciQA, and SciBench[9,10,19] broaden domain coverage yet rely primarily on factoid-style questions that do not require multi-hop reasoning over structured relational knowledge. Domain-specific efforts such as ChemLit-QA[20] and health-oriented QA systems[11] demonstrate that evidence reliability and temporal relevance matter for scientific retrieval, yet remain confined to single-domain, passage-level answering without structured multi-hop reasoning paths. More challenging benchmarks such as Humanity's Last Exam and LongBench v2[21,22] confirm that even frontier LLMs struggle with information-intensive scientific reasoning, and evaluation frameworks such as ETHIC[4] further reveal that high information-density tasks remain a fundamental bottleneck for long-context models. While agentic systems have demonstrated strong knowledge synthesis capabilities[23], no existing benchmark systematically evaluates evidence-grounded, multi-hop QA with explicit reasoning chains—the gap that AutoQG is designed to fill.

2.2 Academic QA generation methods

Automated QA generation methods broadly fall into two paradigms: crowdsourced annotation and LLM-based pipelines. Crowdsourced approaches[7,8] yield high-quality annotations but are prohibitively costly and difficult to scale. Scalable LLM-based methods such as QASA and SciQAG[24,25] achieve broad coverage but generate questions focused on factual recall rather than inferential depth, lacking explicit reasoning paths or traceable evidence. Conversational approaches leveraging ChatGPT for zero-shot information extraction[26] and generative AI tools for literature review synthesis[6] further demonstrate the potential of LLMs in knowledge-intensive workflows, yet do not address the need for structured, evidence-anchored question generation. More recent frameworks, including SciCUEval and AutoPaperBench[27,28] extend automated evaluation across scientific domains but similarly omit evidence anchoring at question construction time. These limitations collectively motivate a structured, evidence-grounded approach to large-scale QA generation.

2.3 Automatic knowledge graph construction

Early work on scientific knowledge graph construction established the value of unified extraction pipelines that jointly identify entities, relations, and coreference from scientific papers[29]. Document-level relation extraction datasets such as DocRED[30] and automated frameworks such as AutoRE[31] further demonstrate that complex relational reasoning across sentences is essential for robust KG population. Recent KG construction methods increasingly leverage LLMs for structured information extraction under few-shot and zero-shot settings. In-context learning approaches including GPT-RE[32] and CodeIE[33] show that carefully engineered prompts and code-style outputs substantially improve LLM extraction fidelity, while instruction-tuning frameworks such as Instruct and Extract[34] enable on-demand extraction across heterogeneous tasks. Unified extraction frameworks such as UIE[35] consolidate NER, relation extraction, and event extraction into a single generation model, while knowledge-integrated approaches like KnowCoder and GoLLIE[36,37] demonstrate that schema design and annotation guidelines are critical for zero-shot generalization. Domain applications such as LLM-powered mental health knowledge graph construction[38] further validate the broader applicability of LLM-based KG pipelines beyond NLP-centric benchmarks. Prompt engineering strategies and empirical benchmarking[39,40] further confirm that task-aligned instruction design is the primary determinant of extraction quality. Despite these advances, most approaches lack explicit joint modeling of reasoning paths and source-text provenance[14,41], limiting their reliability for explainable KG-based question answering. Retrieval-augmented frameworks such as TRACE[12] and graph-structured approaches such

as THE-Tree[5] further establish that explicit, structured reasoning chains are essential for interpretable multi-hop inference—a design principle that directly motivates AutoQG's ontology-guided, evidence-anchored extraction.

3 Task definition

We formalize two interrelated problems that jointly underpin evidence-traceable question generation from academic literature: *academic knowledge graph construction* with source provenance and *evidence-grounded multi-hop QA generation*.

3.1 Academic knowledge graph with source context

An *academic knowledge graph* $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T}, C)$ comprises four components:

- $\mathcal{E}$: a set of entities (e.g., methods, datasets, metrics, contributions) extracted from scholarly documents;
- $\mathcal{R}$: a set of typed relations (e.g., `proposes`, `evaluates_on`, `outperforms`) defined by a domain ontology;
- $\mathcal{T} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$: a collection of triples (h, r, t) representing relational facts;
- $C : \mathcal{T} \to 2^S$: a mapping from each triple to a set of source-text spans S that provide evidential support.

Given a scholarly document D segmented into sections $\{s_1, \ldots, s_n\}$ and a domain ontology $O = (\mathcal{E}_{\text{type}}, \mathcal{R})$ that supplies the admissible entity types and relation vocabulary, the *KG construction task* is to derive the instance components $(\mathcal{E}, \mathcal{T}, C)$ from D—with $\mathcal{R}$ inherited from O—yielding $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T}, C)$ such that every triple $\tau \in \mathcal{T}$ satisfies the following two constraints:

1. **Ontology conformance**: $r \in \mathcal{R}$ must align with a pre-defined schema (Section 4.1), ensuring structural coherence and domain validity.
2. **Textual grounding**: $C(\tau) \neq \emptyset$, i.e., every triple anchors to at least one verbatim passage in the source manuscript, enabling full traceability.

A knowledge graph $\mathcal{G}$ that passes multi-dimensional quality evaluation (Section 4.3) is denoted $\mathcal{G}^\star$ and serves as the validated input to the downstream QA generation task defined below.

3.2 Evidence-traceable multi-hop QA generation

Given a finalized knowledge graph $\mathcal{G}^\star$ and its associated context mapping C, the *evidence-traceable QA generation* task produces a QA pair (q, a) together with an evidence set E such that

- **Reasoning path**: the answer a is derived via a traversal over a sequence of triples $\{\tau_1, \tau_2, \ldots, \tau_k\} \subseteq \mathcal{T}$, where $k \in \{1, 2, 3\}$ encodes question complexity. Here, $k = 1$ corresponds to evidence-grounded factual QA, while $k \in \{2, 3\}$ corresponds to multi-hop reasoning QA;
- **Full evidence provenance**: $E = \bigcup_{i=1}^{k} C(\tau_i)$, guaranteeing that every reasoning step references explicit textual support;
- **Schema-constrained generation**: the relation sequence $(r_1, r_2, \ldots, r_k)$ must form a valid reasoning chain under the ontology $\mathcal{R}$, preventing hallucinated or structurally inconsistent questions.

Formally, we define the generation objective as:

$$(q, a, E) = f(\{\tau_i\}_{i=1}^{k}, C) \quad \text{subject to} \quad \{\tau_i\} \subseteq \mathcal{T},\ E = \bigcup_{i=1}^{k} C(\tau_i),\ |E| > 0, \quad (1)$$

where f denotes the question-generation function implemented by the QA Generation Agent (Section 4.4).

4 Methodology

We propose AutoQG, a multi-agent framework that constructs ontology-guided academic knowledge graphs from scientific papers and generates evidence-traceable, difficulty-controllable QA pairs without manual annotation or model fine-tuning. As illustrated in Fig. 2, the system follows a three-stage *construct–evaluate–generate* paradigm comprising three specialized agents: the KG Extraction Agent, the KG Evaluation Agent, and the QA Generation Agent.

Given a segmented scientific document and a domain-specific ontology, the system operates in three sequential stages. First, the KG Extraction Agent parses the text into a candidate knowledge graph. Then, the KG Evaluation Agent iteratively refines it, passing the validated graph to the QA Generation Agent to synthesize traceable questions. Each agent is implemented as an LLM-based module with tailored prompts and quality-control policies.

4.1 KG Extraction Agent: section-aware and ontology-guided extraction

The KG Extraction Agent transforms each section $s_i \in D$ into structured triples by leveraging section-specific schemas and provenance-tracking constraints. Academic documents exhibit strong section-level coherence—*Methods* describe algorithmic procedures, while *Results* present empirical outcomes. To exploit this regularity, extraction proceeds section-by-section, with the ontology schema O serving as a type constraint on entities and relations.

For each section s_i, the agent queries a large language model f_θ using a schema-constrained prompt $\Phi(O, s_i)$ that enforces entity types T and relation types R defined by O. The output is a set of candidate triples $K_i \subseteq E \times R \times E$. Crucially, for each extracted triple (e_h, r, e_t), the agent records the supporting evidence:

$$\Gamma(e_h, r, e_t) = \{\text{spans in } s_i \text{ that justify } (e_h, r, e_t)\}. \quad (2)$$

This evidence mapping Γ preserves provenance at extraction time, enabling downstream traceability. Aggregating over all sections yields the initial knowledge graph $G = (E, K)$ where $E = \bigcup_i E_i$ and $K = \bigcup_i K_i$.

This design confers three practical advantages: it reduces context length per extraction call, aligns entity type distributions with discourse structure, and enables section-specific relation filtering (e.g., prioritizing `evaluates_on` in *Results* and `proposes` in *Methods*). To illustrate, consider the following sentence drawn from the *Methods* section:

'We trained Model-A on the CIFAR-100 dataset with a learning rate of 0.001'.

A flat baseline may only extract (*Model-A*, `trainedOn`, *CIFAR-100*), whereas our section-specific template, guided by the ontology, yields richer relations:

$$(Model-A, \texttt{evaluatedOn}, CIFAR-100),\ (Model-A, \texttt{hasHyperparameter}, learning\ rate = 0.001). \quad (3)$$

The AutoQG ontology is automatically induced from section-aware prompts applied across 4,435 papers, yielding over 800 open-vocabulary fine-grained relations. These relations exhibit a long-tail distribution that naturally clusters into eight core scientific discourse categories—*Method*, *Usage*, *Evaluation*, *Result*, *Comparison*, *Component*, *Data*, and *Concept*—spanning 23.4% to 4.3% of the total triple frequency. This hierarchical structure preserves extraction granularity while enforcing the discourse-level coherence required for multi-hop reasoning. Figure 3 summarizes the empirical distribution.

https://doi.org/10.48130/ker-0026-0017

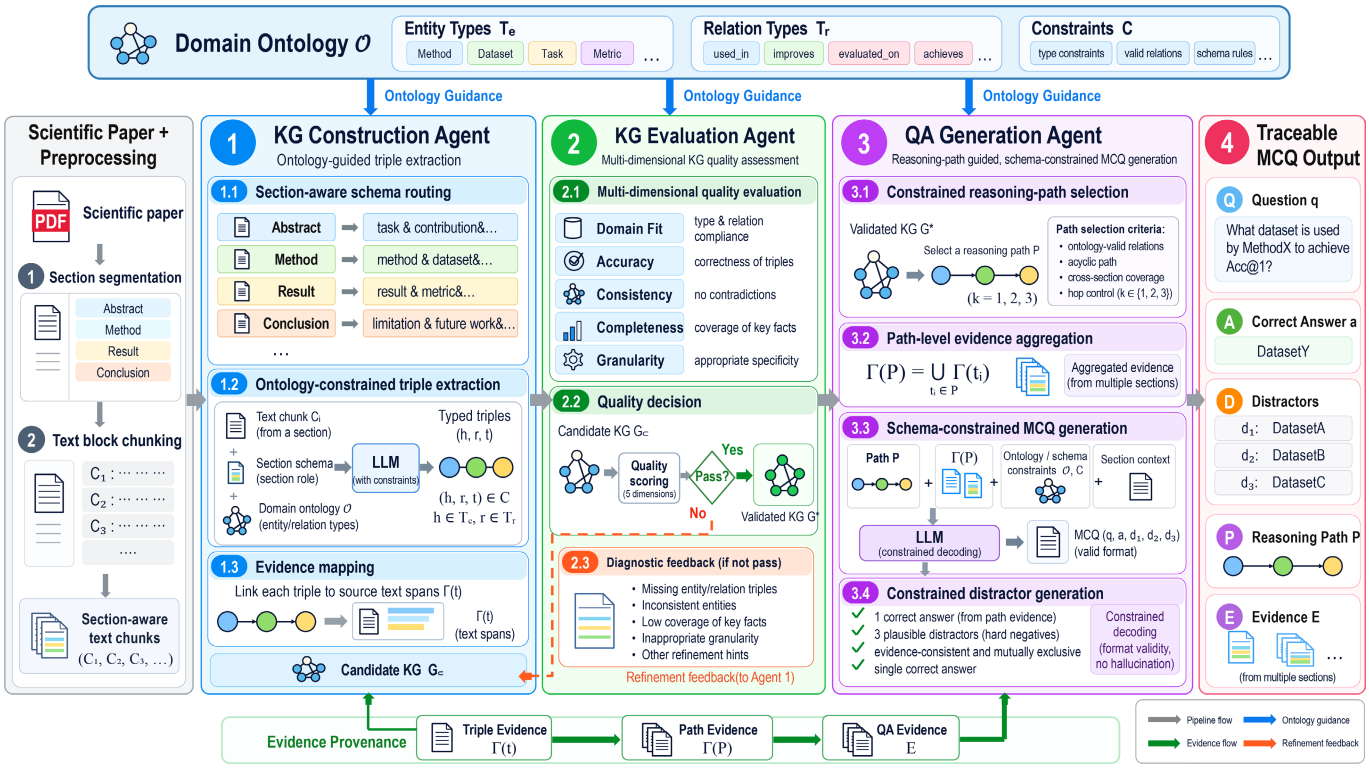


Fig. 2 Overview of the AutoQG framework. AutoQG follows a construct–evaluate–generate pipeline: section-aware scientific papers are transformed into ontology-constrained candidate KGs, audited and refined by a KG Evaluation Agent, and then converted into traceable QA pairs through schema-constrained reasoning paths. Blue arrows indicate ontology constraints, green arrows indicate evidence provenance, and orange dashed arrows indicate diagnostic refinement feedback.

4.2 Ontology construction and validation

Ontology construction proceeds in four stages: (1) initializing a compact academic schema covering recurring discourse roles (models, datasets, metrics, methods, tasks, etc.); (2) allowing the model to propose open-vocabulary predicates only when backed by an explicit evidence span; (3) normalizing predicates by lowercasing, removing punctuation-only variants, and merging near-synonymous surface forms into canonical relation families; and (4) discarding relations that fail type compatibility checks, lack `contextText`, or cannot be assigned to a scientific discourse category.

The resulting ontology is two-level: a coarse level of stable entity and relation families constrains extraction and path traversal, while a fine-grained level preserves scientific specificity (`evaluatedOn`, `achievesScore`, `proposes`, `outperforms`, etc.).

Validation operates at both triple and graph levels. Each triple must pass type-compatibility, evidence-entailment, and source-section checks. The KG Evaluation Agent then audits domain fit, factual accuracy, consistency, completeness, and granularity at the graph level, returning low-scoring graphs to the Extraction Agent for correction. The same schema is enforced during QA generation—reasoning paths must follow schema-valid relation sequences with per-hop evidence mappings—so that extraction and question generation remain mutually consistent.

4.3 KG Evaluation Agent: multi-dimensional assessment and iterative refinement

The KG Evaluation Agent assesses the quality of G across multiple dimensions and orchestrates feedback-driven refinement. For each triple $(e_h, r, e_t) \in K$ with evidence $\Gamma(e_h, r, e_t)$, the agent computes a composite quality score across five interpretable criteria: Domain Fit (s_{dom}), Accuracy (s_{acc}), Consistency (s_{con}), Completeness (s_{com}), and Granularity (s_{gra}). Each dimension is estimated via targeted LLM prompts (see A7 in Supplementary File 1), producing a quality vector

$$\mathbf{q}(G) = (s_{\text{dom}}, s_{\text{acc}}, s_{\text{con}}, s_{\text{com}}, s_{\text{gra}}) \in [0,1]^5. \tag{4}$$

The overall graph quality is defined as a weighted aggregation:

$$Q(G) = \sum_{i=1}^{5} w_i \cdot s_i, \qquad \sum_{i=1}^{5} w_i = 1, \tag{5}$$

where weights $\{w_i\}$ are calibrated via ablation studies (§5), with higher emphasis on accuracy and consistency as they are most predictive of downstream QA performance.

When $Q(G) < \tau$, the evaluator generates actionable diagnostics—for instance, 'low-confidence triples concentrated in *Methods*' or 'entity *Model-A* inconsistently referenced across sections'. These diagnostics are injected into the extractor's next-round prompts:

$$G^{(t+1)} = R\left(G^{(t)}, \mathbf{q}(G^{(t)})\right), \tag{6}$$

where $R(\cdot)$ denotes a refinement operator that integrates evaluator feedback into the extraction process. Iteration continues until $Q(G) \geq \tau$, yielding the finalized graph G^*. This feedback loop transforms extraction from a one-shot process into a robust, self-correcting pipeline.

4.4 QA Generation Agent: schema-constrained and evidence-traceable QA

Given the refined knowledge graph G^* and its evidence mappings Γ, the QA Generation Agent produces QA pairs that satisfy dual constraints: each question is derived from a multi-hop reasoning path in G^*, and each answer is supported by explicit textual evidence from D.

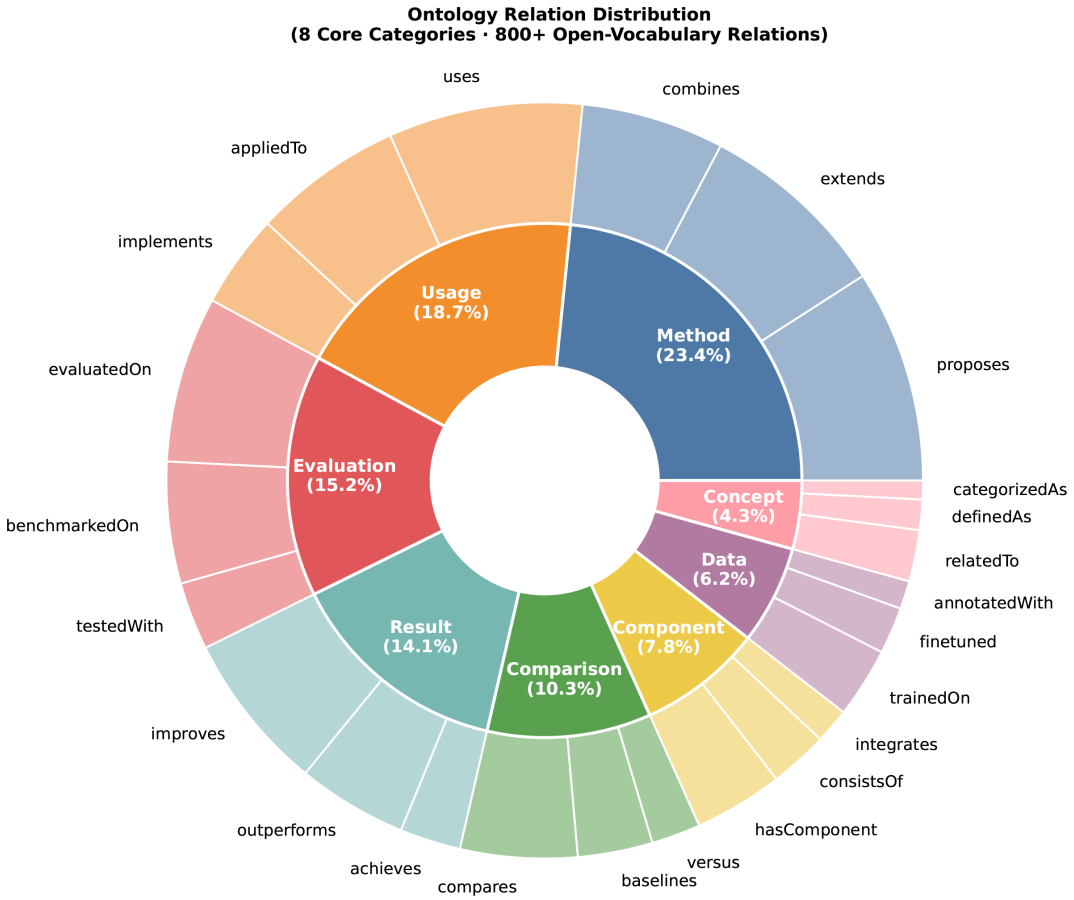


Fig. 3 Sunburst visualization of the AutoQG ontology. Inner ring: eight core discourse categories; outer ring: representative fine-grained relations. Sector size ∝triple frequency across 4,435 papers.

A *reasoning path* is a sequence of entities connected by ontology-aligned relations:

$$P = (e_1 \xrightarrow{r_1} e_2 \xrightarrow{r_2} \dots \xrightarrow{r_k} e_{k+1}), \tag{7}$$

where $e_i \in E$ and $r_i \in R$. The length k determines the reasoning depth required to answer the corresponding question. Each path P is associated with aggregated supporting evidence:

$$\Gamma(P) = \bigcup_{i=1}^{k} \Gamma((e_i, r_i, e_{i+1})). \tag{8}$$

Rather than sampling paths randomly, AutoQG employs a *structure-constrained traversal strategy* that enforces cross-paragraph, acyclic reasoning chains. This strategy prioritizes paths that span multiple discourse sections (e.g., from *Methods* to *Results*), thereby reducing shallow local hops and encouraging deeper, more integrative reasoning. By avoiding cycles and ensuring diversity in section coverage, the traversal process produces paths that are both structurally valid and pedagogically valuable.

For each selected path $P \subseteq G^*$ with evidence $\Gamma(P)$, the agent generates a QA pair via a schema-constrained prompt template $\Psi(\cdot)$ and a constrained decoding policy $\delta(\cdot)$ that enforces type-correct entities and valid answer formatting:

$$(Q, A) = \delta(f_\theta(\Psi(P, \Gamma(P)))). \tag{9}$$

The complete QA corpus is defined as

$$\mathcal{Q} = \{(Q, A, P, \Gamma(P)) \,:\, (Q, A) = \delta(f_\theta(\Psi(P, \Gamma(P)))),\ P \subseteq G^*\}. \tag{10}$$

By varying the hop length k and relation types in R, the generation process provides explicit control over reasoning difficulty while maintaining full traceability from questions to textual evidence.

5 Experimental results

We conduct comprehensive experiments to evaluate AutoQG across multiple dimensions: knowledge graph extraction quality, QA generation performance, comparison with state-of-the-art frontier models, ablation studies on reasoning-path strategies and input modalities, iterative refinement dynamics, and downstream supervised fine-tuning utility.

5.1 Experimental setup

Datasets and materials. We construct AutoQG-7k, a benchmark comprising 7,921 QA pairs derived from 4,435 academic papers spanning computer science, biology, and medicine. The corpus is sourced from arXiv (https://arxiv.org) and ACL Anthology (https://aclanthology.org), covering diverse topics including transformer architectures, reinforcement learning, and multimodal understanding. Each paper undergoes section-aware parsing, ontology-guided extraction, and iterative refinement to produce validated knowledge graphs and evidence-traceable questions. The framework generates 243,214 knowledge graph triples and constructs reasoning paths of varying hop lengths ($k \in \{1, 2, 3\}$) with schema-constrained decoding to ensure type-correct answers and evidence traceability. Dataset statistics are provided in Table 1.

The multi-agent framework is implemented using LangGraph (https://www.langchain.com/langgraph), with GPT-4o as the backbone LLM for all agents. Each agent operates within a structured reasoning loop: the KG Extraction Agent constructs initial graphs from parsed document chunks, the KG Evaluation Agent performs

https://doi.org/10.48130/ker-0026-0017

Table 1. Dataset statistics of *AutoQG*^{20k} and *AutoQG*^{7k}.

Metric	*AutoQG*^{20k}	*AutoQG*^{7k}
Papers	4,435	4,435
Triples	243,214	243,214
QA pairs	20,424	7,921

schema-constrained verification and iterative refinement, and the QA Generation Agent synthesizes evidence-traceable QA pairs from validated reasoning paths.

Baselines and frontier models. For knowledge graph extraction, we compare against zero-shot methods (Self-Prompting, QA4RE, ChatIE) and LLM-based approaches (Gemini-1.5-Pro, DeepSeek-V2.5, Claude-3.5 Opus, o3, o1). For QA generation, we benchmark AutoQG against leading large language models, including GPT-5.1, Claude Sonnet 4, Baichuan4, Qwen3-32B, GLM-4, Gemini-1.5 Pro, and DeepSeek-V2.5. All baselines receive identical input documents and are prompted to generate questions with distractors and evidence citations.

To clarify baseline fairness, we keep the source papers fixed within each diagnostic comparison and vary only the availability of ontology constraints, evidence preservation, refinement, and reasoning-path construction. The full baseline-setting matrix is reported in Supplementary File 1: A2.

Evaluation metrics. For knowledge graph quality, we employ both manual annotation and automated LLM-as-a-judge protocols. Human evaluation reports precision, recall, and F1 on a gold dataset of 30 manually annotated papers, with inter-annotator agreement Cohen's $\kappa = 0.82$. Automated evaluation assesses five dimensions: *Dom (domain fit)*, *Acc (accuracy)*, *Con (consistency)*, *Com (completeness)*, and *Gra (granularity)*.

For QA generation, we employ six dimensions: *Rel (relevance)*, *Ans (answer correctness)*, *Distr (distractor quality)*, *Flu (fluency)*, *Reas (reasoning depth)*, and *Evid (evidence traceability)*. Heterogeneous judge panels—comprising GPT-5, Claude-3.5-Sonnet, and Qwen2.5-72B—score each dimension on a 10-point scale. Replacing GPT-5 with DeepSeek-V2.5 alters scores by at most 0.2, and the resulting judgments correlate strongly with human experts (Pearson $r = 0.973$, MAE = 0.15).

Iterative refinement protocol. We conduct up to four rounds of graph refinement. In each round, the validation agent scores triples and flags low-quality instances for regeneration. We track average KG and QA scores per round and the proportion of graphs exceeding the quality threshold (Qual ≥ 8.0).

5.2 Knowledge graph quality evaluation

Benchmark comparison. Table 2 compares AutoQG against zero-shot, direct LLM-based, and adapted agentic extraction methods. AutoQG achieves F1 = 72.3 on human-annotated gold triples, substantially outperforming the best direct baseline (F1 = 37.0). Among zero-shot and direct LLM extraction baselines, AutoQG also achieves the strongest LLM-as-a-judge final score (8.3), reflecting strong domain fit, factual accuracy, cross-sectional consistency, relational completeness, and fine-grained granularity. The adapted LangExtract (https://github.com/google/langextract) and EPT-LLM baselines receive competitive source-grounded KG-judge scores, but strict ontology-aligned exact matching is highly sensitive to surface-form normalization. Therefore, we report both strict gold-terse RDF triple language (TTL) exact-match scores and diagnostic schema/evidence/grounding metrics, rather than relying on exact-match F1 alone. This separation indicates that source-grounded extraction alone does not recover the normalized ontology-aligned triples required by AutoQG. These gains stem from three synergistic design choices: (i) **section-aware extraction** exploits discourse structure to reduce noise and prioritize domain-relevant triples; (ii) **ontology-guided constraints** suppress spurious relations and enforce structural validity; and (iii) **iterative refinement** ensures cross-sectional coherence and completeness through evaluator feedback loops.

External adapted extraction validation. We also conduct a SciLEAD-adapted TDMR extraction evaluation to test whether AutoQG's ontology-guided and evidence-grounded extraction transfers beyond our own KG schema. Because SciLEAD evaluates leaderboard task-dataset-metric-result (TDMR) extraction rather than QA generation, we evaluate only the extraction component under a task-specific TDMR schema and do not claim full SciLEAD leaderboard reproduction. AutoQG achieves the best ETM-TDMR F1 and exact tuple matching on task–dataset–metric (ETM-TDM) F1 on the same 30-paper manually annotated gold set, while preserving complete evidence coverage. Detailed SciLEAD diagnostic results are reported in Supplementary File 1: A4.

Iterative refinement analysis. As shown in Fig. 4 and Table 3, the KG Evaluation Agent's diagnostic feedback drives consistent quality gains across the first three rounds. Rounds 1 and 2 yield modest improvements ($Q(G) = 7.12$ for both; QA rises from 8.86 to 8.90), as the evaluator identifies and corrects surface-level extraction errors. Round 3 achieves the optimal configuration: $Q(G) = 8.30$, QA = 9.30, with 80% of graphs exceeding the quality threshold τ, reflecting the full benefit of structured feedback on multi-hop reasoning paths and evidence grounding.

Table 2. Evaluation of knowledge graph extraction quality on the manually annotated gold set.

Category	Method	Gold-set evaluation			LLM-as-a-Judge evaluation					
		Prec.	Rec.	F1	Dom	Acc	Con	Com	Gra	Final
Zero-shot extraction methods	Self-prompting	35.7	35.7	35.7	4.9	5.0	5.2	5.1	5.6	5.2
	QA4RE	32.3	30.4	31.3	5.8	5.1	5.0	5.5	5.3	5.3
	ChatIE	16.5	34.1	21.4	5.9	5.3	5.2	5.6	5.3	5.3
LLM-based extraction methods	Gemini-1.5-pro	23.4	17.5	20.0	8.9	6.9	7.1	6.2	7.1	7.3
	DeepSeek-V2.5	37.0	34.0	37.0	8.1	7.3	7.7	6.5	7.4	7.4
	Claude-3.5 Opus	30.0	28.6	29.3	9.1	7.0	7.4	6.7	7.2	7.5
	O3	35.0	26.4	30.1	8.1	7.3	7.5	7.3	7.4	7.5
	O1	53.8	26.4	35.4	9.3	8.1	8.8	6.8	7.7	8.1
	Ours	**73.6**	**71.2**	**72.3**	9.3	8.5	8.9	7.5	7.5	8.3
Agentic baselines	LangExtract	16.7	26.5	30.2	8.8	8.7	8.5	7.4	7.3	8.2
	EPT-LLM	30.5	29.1	30.1	8.7	8.9	8.4	7.5	7.4	8.3

The table reports gold-set Precision/Recall/F1, and LLM-as-a-judge scores. Bold numbers denote the best results.

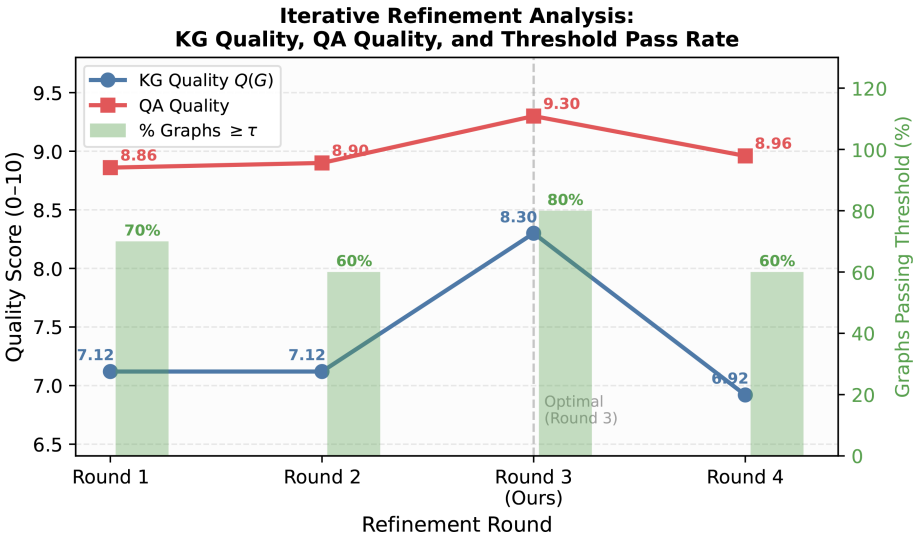


Fig. 4 Dual-axis trend chart of iterative refinement. Blue line (left axis): KG quality $Q(G)$; red line (left axis): overall QA quality; green bars (right axis): percentage of graphs passing threshold $\tau \geq 8.0$. Round 3 achieves peak performance before over-refinement degrades quality in Round 4.

Table 3. Iterative refinement results across KG and QA quality dimensions.

Round	KG dimensions					$Q(G)$	$\Delta Q(G)$	% $\geq \tau$	QA dimensions						QA	ΔQA
	Dom	Acc	Con	Com	Gra				Rel	Ans	Distr	Flu	Reas	Evid		
1	8.8	6.9	6.7	6.2	7.0	7.12	—	70%	9.1	9.0	8.4	9.3	8.3	9.1	8.86	—
2	8.8	7.2	6.5	6.3	6.8	7.12	+0.00	60%	9.2	8.9	8.5	9.4	8.4	9.0	8.90	+0.04
3 (Ours)	9.3	8.5	8.9	7.5	7.5	**8.30**	**+1.18**	**80%**	9.6	9.6	9.0	9.5	8.5	9.6	**9.30**	**+0.40**
4	8.9	6.7	6.4	6.0	6.6	6.92	−1.38	60%	9.4	9.1	8.5	9.4	8.3	9.1	8.96	−0.34

The table reports both per-dimension scores and composite scores for each refinement round. Round 3 is selected as the default configuration because it achieves the best balance between KG quality, QA quality, and threshold pass rate before over-refinement begins. Round 3 represents the optimal configuration with the highest overall scores and threshold pass rate. Bold numbers denote the best results.

Table 4. Human evaluation and LLM-as-a-judge evaluation results.

Method	Human evaluation (0–10)						LLM-as-a-judge (0–10)					
	Rel	Ans	Distr	Flu	Reas	Evid	Rel	Ans	Distr	Flu	Reas	Evid
GPT-5.1	8.9	9.4	8.8	9.2	8.3	8.6	9.4	9.0	**9.2**	9.4	8.6	8.4
Claude Sonnet 4	8.4	8.4	8.3	9.2	8.2	8.2	8.4	7.9	8.7	9.3	7.6	7.6
Baichuan4	8.9	8.4	8.6	9.2	8.1	8.3	8.9	8.6	<u>9.0</u>	9.4	8.0	7.9
Qwen3-32B	8.9	8.2	8.3	9.1	8.1	8.1	8.2	8.0	8.0	8.9	7.4	7.5
GLM-4	7.6	7.4	7.3	9.2	8.1	7.9	7.6	7.4	7.5	8.6	7.0	7.0
Gemini-1.5 Pro	8.6	8.4	8.3	9.2	8.1	8.0	8.1	8.9	8.6	9.3	7.6	8.0
DeepSeek-V2.5	8.8	8.5	8.8	9.0	7.9	8.2	8.9	9.0	8.3	**9.5**	8.0	8.0
Ours	**9.8**	**9.8**	**8.9**	**9.4**	**8.4**	**9.7**	**9.6**	**9.6**	<u>9.0</u>	**9.5**	**8.5**	**9.6**

All LLM-as-a-judge scores are averaged across three independent judges. Scores are averaged across QA pairs and reported on a 0–10 scale. Bold numbers denote the best results and underlined numbers denote the second best results.

Round 4, however, reveals a clear over-refinement effect: $Q(G)$ drops by $\Delta = -1.38$ to 6.92, and the threshold pass rate falls back to 60%, despite a marginal QA score of 8.96. This degradation arises because excessive correction disrupts well-formed triples and introduces inconsistencies, confirming that diminishing returns set in after the third iteration. We therefore adopt three refinement rounds as the optimal stopping criterion for all experiments.

5.3 QA generation performance

Table 4 presents the multi-dimensional evaluation of AutoQG and baseline models. AutoQG achieves the highest reasoning depth (Reas = 8.4) in human evaluation and (Reas = 8.5) in LLM-as-a-judge evaluation, while maintaining superior evidence traceability (Evid = 9.7 and 9.6, respectively). Across all six metrics, AutoQG consistently ranks in the top tier, demonstrating that the integration of validated knowledge graphs and schema-constrained generation enhances both reasoning complexity and evidential precision. Figure 5 visualizes the performance comparison between AutoQG and two representative baseline models (Gemini-1.5 Pro and DeepSeek-V2.5).

Figure 5a illustrates AutoQG's higher scores and uniformly expanded radar shape compared to representative baselines (Gemini-1.5 Pro and DeepSeek-V2.5), indicating balanced improvements across all metrics. Figure 5b reveals a near-perfect correlation between human and LLM-as-a-judge scores ($r = 0.982$ for AutoQG), confirming the reliability of human-verifiable outputs. While all baseline models exhibit competitive fluency, AutoQG demonstrates superior performance in reasoning depth and evidence grounding.

Benchmark evaluation on Large Language Models. To further validate the quality and difficulty of the generated benchmark, we evaluate a diverse set of large language models on AutoQG-20k and AutoQG-7k. Table 5 presents the results across difficulty-controlled

https://doi.org/10.48130/ker-0026-0017

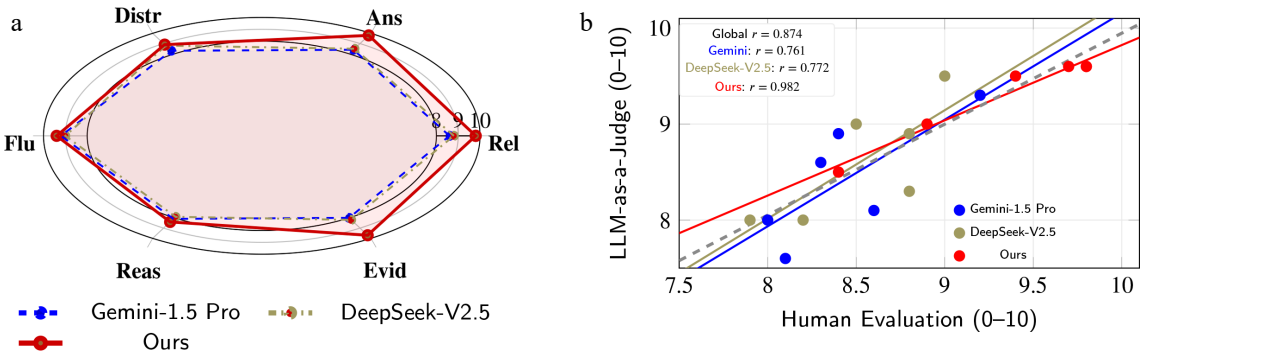


Fig. 5 Comparison of human and LLM-as-a-judge evaluations. (a) Radar chart across six dimensions. (b) Scatter plot with individual regression lines for each model and an overall (dashed) global regression.

Table 5. Extended evaluation of the generated QA benchmark under two settings: AutoQG-20k (left) and AutoQG-7k (right).

Category	Model	AutoQG-20k				AutoQG-7k
		1-hop	2-hop	3-hop	Acc (w/Ev.)	QA-Pair
Base models	GPT-4o	95.2	91.0	72.5	96.1	43.3
	Mistral Large 2	94.2	91.0	76.4	95.2	55.3
	WizardLM-2 70B	89.6	72.4	64.6	92.6	35.9
	Qwen-2.5 14B	86.0	81.3	67.3	99.7	30.6
	DeepSeek-V2.5	86.9	82.7	64.7	98.6	33.7
	Gemini-1.5-flash	87.5	77.3	65.8	98.7	29.0
	GLM-4	83.1	75.0	65.6	98.7	29.0
	Claude-haiku	82.9	79.7	60.2	99.7	31.4
	Mistral-7B (Base)	85.7	82.0	62.2	99.2	28.0
	DBRX	83.4	79.2	56.2	90.8	30.7
Reasoning models	Claude-3.5 Sonnet	86.0	84.0	75.8	95.9	52.9
	Grok-2	94.0	88.0	69.6	92.9	41.9
	Mistral-7B	76.5	69.2	60.7	90.9	27.6

Results are strict accuracy (%) on QA pairs, difficulty-controlled hops, and accuracy with/without evidence.

reasoning hops (1-hop, 2-hop, 3-hop) and evidence-grounded accuracy. Base models such as GPT-4o achieve 95.2% on 1-hop questions but drop to 72.5% on 3-hop questions, demonstrating the increasing difficulty with reasoning depth. On the challenging AutoQG-7k subset, even advanced models struggle, with GPT-4o achieving only 43.3% accuracy. Reasoning-oriented models like Claude-3.5 Sonnet show improved performance on complex questions (75.8% on 3-hop), validating that AutoQG effectively captures multi-hop reasoning requirements. The high accuracy with evidence (Acc w/Ev.) across all models confirms that the generated questions are well-grounded and traceable to the source text.

5.4 Ablation studies

To validate the contribution of each design choice, we conduct ablation experiments along three dimensions: reasoning-path strategies, input modalities, and individual agent components.

Reasoning-path strategies. To isolate the contribution of structure-constrained traversal, we compare three path-selection policies: **P0** (structure-constrained, cross-paragraph, acyclic traversal; ours), **P1** (random path sampling), and **P2** (length-restricted random sampling). P0 achieves Reas = 8.5, significantly outperforming P1 (Reas = 7.9) and P2 (Reas = 8.2). Moreover, P0 leads in relevance and evidence traceability, confirming that enforcing multi-paragraph connectivity and avoiding trivial local hops strengthens reasoning quality and evidential rigor.

Table 6. Ablation on input modalities.

Strategy	Rel	Ans	Distr	Flu	Reas	Evid
S1 (KG + sent) – ours	**9.6**	**9.6**	**9.0**	**9.5**	**8.5**	**9.6**
S2 (sent-only)	9.4	9.5	8.5	9.4	8.0	9.1
S3 (KG-only)	9.4	9.5	8.3	9.45	8.3	9.4
S4 (random-sent)	8.0	8.2	7.8	9.4	7.5	8.5

Bold numbers denote the best results.

Input modalities. Table 6 presents the ablation results on input modalities. We investigate the interplay between knowledge graphs and raw sentence inputs via four configurations: **S1** (KG + Sentence; AutoQG), **S2** (Sentence-only), **S3** (KG-only), and **S4** (Random-Sentence baseline). S1 achieves the best overall performance (Reas = 8.5, Evid = 9.6). Removing the KG (S2) degrades reasoning depth and structural coherence. Eliminating source text (S3) diminishes evidence traceability and distractor plausibility. Random sentence selection (S4) severely underperforms across all metrics. These results demonstrate that the dual-stream architecture—combining symbolic graph structure and natural-language evidence—is essential for high-quality, traceable QA generation.

Component-wise ablation on agents. To quantify the indispensability of each design component, we conduct a leave-one-out ablation by disabling one module at a time: *Section-aware Prompting (SP)* and *Ontology Constraints (OC)* in the KG Extraction Agent; *Iterative Feedback Refinement (FB)* in the KG Evaluation Agent; and *Question-Format Constraints (QF)* and *Evidence Grounding (E)* in the QA Generation Agent. Figure 6 visualizes the resulting score drops in Reasoning Depth (Reas) and Evidence Traceability (Evid) relative to the full AutoQG baseline (Reas = 8.5, Evid = 9.6).

The results reveal a clear hierarchy of component importance. The KG Extraction Agent's ontology constraints (**w/o OC**) prove most critical: their removal causes the steepest Reas drop (8.5 → 5.0, Δ=−3.5), as unconstrained extraction produces noisy triples that undermine multi-hop reasoning path quality. Disabling section-aware prompting (**w/o SP**) incurs a similarly severe Reas penalty (Δ=−3.0), confirming that cross-section discourse structure is essential for capturing complete reasoning chains. In the KG Evaluation Agent, removing iterative feedback refinement (**w/o FB**) degrades Reas by −1.0 and Evid by −0.5, demonstrating that one-shot extraction without self-correction allows inaccurate triples to propagate to QA generation. Finally, ablating evidence grounding (**w/o E**) in the QA Generation Agent produces the largest Evid drop (Δ=−1.4), while removing question-format constraints (**w/o QF**) substantially weakens reasoning depth (Δ=−2.4). Collectively, these results confirm that every component plays a non-redundant role, and the full AutoQG pipeline

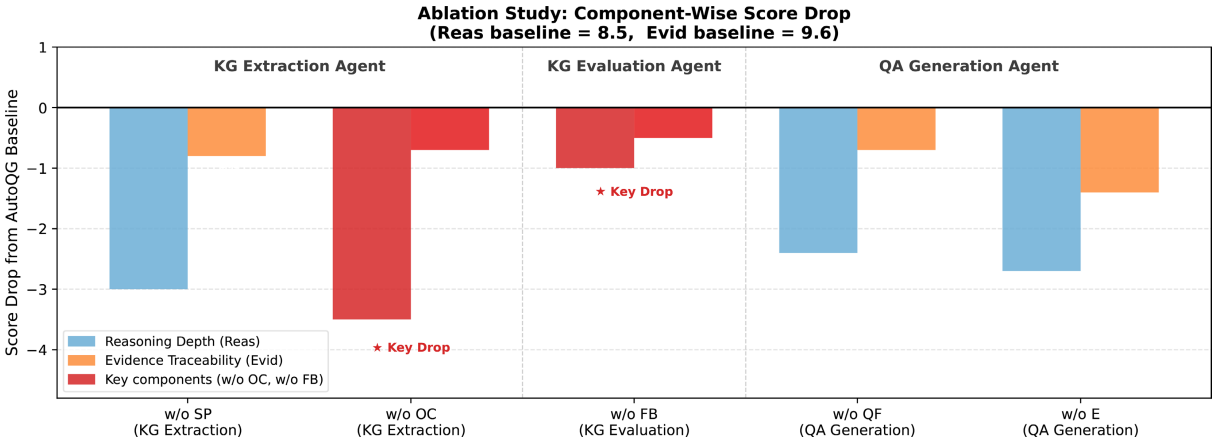


Fig. 6 Component-wise score drop from the full AutoQG baseline (Reas = 8.5, Evid = 9.6). Each bar shows how much Reasoning Depth (**Reas**, blue) and Evidence Traceability (**Evid**, orange) decrease when a single component is removed from the pipeline. Bars highlighted in red correspond to the two most critical ablations: removing *Ontology Constraints* (**w/o OC**) from the KG Extraction Agent causes the largest Reas drop (-3.5), while removing *Iterative Feedback Refinement* (**w/o FB**) from the KG Evaluation Agent degrades both metrics by up to -1.0. Dashed vertical lines separate the three agent modules.

achieves its performance through their tight coordination. Detailed component-level diagnostic ablation tables on the 30-paper manually annotated gold set are reported in Supplementary File 1: A4.

5.5 Downstream supervised fine-tuning utility

To assess the practical utility of AutoQG-generated data, we fine-tune Qwen2.5-32B on AutoQG-7k. Zero-shot performance on a held-out reasoning benchmark stands at 44.9%. Fine-tuning on GPT-5-generated data yields 46.9%, a modest 2.0% gain. In contrast, fine-tuning on AutoQG-7k elevates accuracy to 79.3%, an absolute improvement of 34.4% over zero-shot and 32.4% over GPT-5 data. This substantial gain underscores that AutoQG's evidence-traceable, schema-constrained questions provide superior instructional signals for reasoning-focused model adaptation, validating the framework's contribution to downstream task performance and model alignment.

Paper-level split and leakage check. Because downstream utility claims can be confounded by source-paper overlap, we add a paper-level leakage audit for the AutoQG-7k SFT split. Using fixed seed 20260627, we sample 200 unique full-paper fingerprints for evaluation and use the remaining 1,794 unique fingerprints for training. This split contains 6,303 training QA items and 697 evaluation QA items. Table 7 reports zero overlap for sample IDs, source-paper IDs, full-paper text fingerprints, exact questions, exact QA pairs, and exact reasoning paths. One exact evidence-span overlap is observed, but it is not used as primary leakage evidence because short evidence spans may recur across unrelated papers.

Table 7. Paper-level SFT split and leakage diagnostics for AutoQG-7k.

Item/check	Value
Train unique paper fingerprints	1,794
Eval. unique paper fingerprints	200
Train QA items	6,303
Eval QA items	697
Sample ID overlap	0
Source-paper ID overlap	0
Full paper SHA1 overlap	0
Exact question overlap	0
Exact QA-pair overlap	0
Exact reasoning-path overlap	0

6 System efficiency and practical analysis

While Section 5 establishes the quality and generalizability of AutoQG, this section examines its practical viability through three complementary lenses: computational efficiency, open-source reproducibility, and cross-domain robustness. We demonstrate that the multi-agent decomposition not only enhances quality but also reduces long-context overhead and associated costs.

6.1 Cost control and open-source deployment

We compare the computational cost of AutoQG against a baseline end-to-end direct QA generation approach to quantify the efficiency gains of task decomposition.

Baseline: direct QA generation. The direct approach feeds the entire paper into a long-context prompting workflow and requests the LLM to produce QA pairs without explicit graph construction or agent-level decomposition. As detailed in Supplementary File 1: A15, this end-to-end baseline averages 13,421.61 calls per paper and 13,421.61 tokens per call. Despite this high token throughput, the overall QA quality is limited (7.31), with particularly weak reasoning depth (Reas = 2.4) and evidence traceability (Evid = 4.5). This confirms that monolithic prompting struggles with complex reasoning and fine-grained evidence grounding in long documents.

AutoQG multi-agent pipeline. AutoQG decomposes the task into modular stages (KG extraction, KG evaluation, and QA generation), each operating on shorter, focused contexts. Although the average number of calls increases to 23,592.53 per paper, the average tokens per call drop sharply to 4,831.52. This modularization therefore controls the growth of long-context computation while yielding a substantial quality gain: the overall QA quality rises to 8.45, with Reas = 8.4 and Evid = 9.6. In particular, AutoQG improves overall QA quality by +1.14 and reasoning depth by +6.0 over the end-to-end baseline, showing that multi-agent decomposition is not merely a design choice but a cost-efficiency mechanism for high-fidelity, evidence-grounded QA generation. Figure 7 summarizes the cost-quality trade-off between the two approaches.

Open-source deployment. To ensure accessibility and reproducibility, we validate AutoQG using open-source LLMs deployable on local infrastructure without proprietary API dependencies. We employ Qwen-7B-Instruct[42] for the KG Generation Agent and DeepSeek-V2.5[43] for both the KG Evaluation and QA Generation

https://doi.org/10.48130/ker-0026-0017

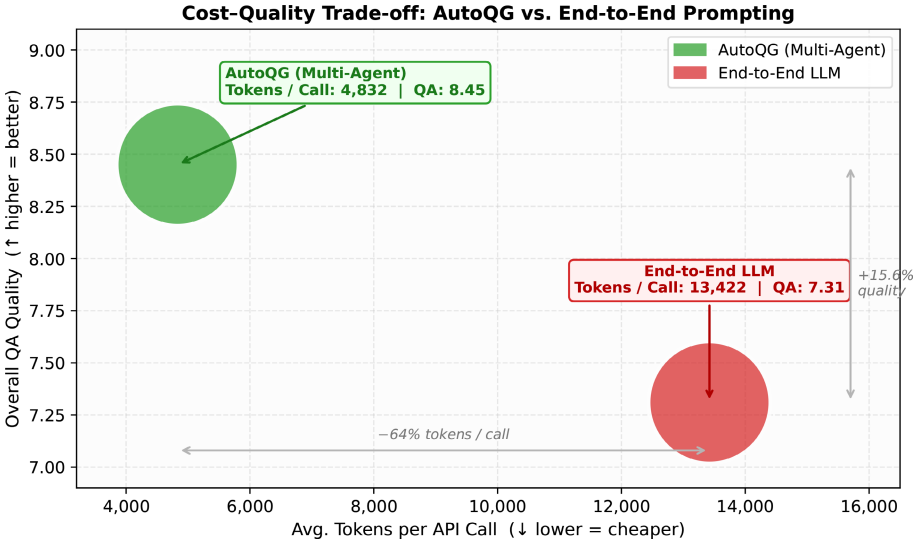


Fig. 7 Cost–quality scatter: AutoQG achieves substantially lower average tokens per call and higher QA quality than end-to-end prompting.

Table 8. Performance reproducibility using open-source models within the AutoQG framework for Knowledge Graph (KG) evaluation and QA generation.

Model	KG Evaluation Performance						QA Generation Performance					
	Dom	Acc	Con	Com	Gra	Overall KG	Rel	Ans	Distr	Flu	Reas	Evid
AutoQG (Qwen-7B-Instruct)	7.0	7.7	7.0	6.0	6.3	7.0	8.6	8.3	7.7	9.0	7.4	8.3
AutoQG (DeepSeek-V2.5)	9.0	8.3	8.3	7.0	7.3	8.0	9.4	9.1	8.5	9.5	8.4	9.2

Agents. Table 8 presents the performance of these open-source configurations. Under the DeepSeek-V2.5 configuration, the overall KG quality reaches 8.0, and QA quality metrics are Rel = 9.4, Ans = 9.1, Distr = 8.5, Flu = 9.5, Reas = 8.4, and Evid = 9.2. These results closely approximate the performance of proprietary models, confirming that AutoQG's design is robust to model choice and does not rely exclusively on expensive closed-source APIs. The average end-to-end cost per paper is approximately 0.06 USD, making the framework highly scalable for large-scale educational and research applications.

6.2 Cross-domain generalization

To assess the framework's adaptability beyond computer science, we conduct experiments on papers from the PubMed (biomedical) and Physics arXiv domains.

Evaluation setup. We apply the same AutoQG pipeline to 20 papers from each domain and compare it against direct GPT-based prompting using the same LLM-judge protocol.

Results. Table 9 presents the cross-domain evaluation results. AutoQG maintains stable performance across domains, with quality variations within ±0.3 points. On PubMed papers, evidence traceability remains high (Evid = 9.0), and on Physics papers, reasoning depth reaches Reas = 8.8. In contrast, direct GPT-based prompting suffers notable degradation, dropping to Reas = 6.8 and 7.0, and Evid = 7.0 and 6.9 on PubMed and Physics, respectively. These results indicate that AutoQG's ontology-guided KG construction and structured agent coordination generalize effectively across diverse scientific domains, whereas monolithic prompting struggles with domain-specific terminologies and reasoning patterns.

Table 9. Cross-domain generalization results.

Domain (method)	Rel	Ans	Distr	Flu	Reas	Evid
PubMed (AutoQG)	9.2	8.9	8.4	9.5	8.5	9.0
PubMed (GPT-5)	7.6	7.5	7.2	8.5	6.8	7.0
Physics (AutoQG)	9.4	9.1	8.5	9.6	8.8	9.3
Physics (GPT-5)	7.6	7.5	6.2	8.6	7.0	6.9
CS (AutoQG - ours)	9.6	9.6	9.0	9.5	8.5	9.6

AutoQG maintains stable performance across PubMed (biomedical) and Physics domains, while direct GPT-based prompting shows significant degradation. CS results are included for reference.

7 Limitations and error analysis

Manual auditing of 42 erroneous triples from the generated knowledge graphs reveals two dominant failure modes. The majority of errors (57.1%) stem from *surface verb drift*, where relation classification relies on shallow verb-pattern heuristics that conflate semantically distinct predicates—for example, mislabeling 'our method uses pretrained embeddings' as a `basedOn` relation rather than the correct `uses` relation. Passive voice and complex syntactic structures exacerbate this confusion. The second major category (38.1%) involves *granularity mismatch*, where the extractor produces overlapping or redundant triples at excessive detail levels, creating knowledge graph edges that complicate downstream reasoning without adding informational value. A smaller fraction (4.8%) arises from missed long-distance cross-paragraph dependencies due to section-level chunking.

These limitations suggest two concrete improvement directions. First, incorporating *contrastive prompting* that explicitly differentiates semantically similar relations through in-context examples could substantially reduce verb-drift errors, and a two-stage extraction-then-classification pipeline may further enhance disambiguation. Second, adopting longer-context models or retrieval-augmented generation with sliding-window approaches would preserve inter-paragraph coherence during extraction, enabling the framework to capture implicit cross-section relationships that the current chunking strategy misses.

Two evaluation-scope limitations should also be noted. First, the 30-paper manually annotated gold set is intended as a descriptive coverage set rather than a statistically random sample of all scientific papers; Supplementary File 1: A3 reports its domain, year, venue, paper-type, and document-length distributions. Second, the SciLEAD experiment is an adapted TDMR extraction evaluation on the same 30-paper gold set designed to test external transfer of the extraction component; it is not a full reproduction of the SciLEAD leaderboard-generation pipeline and does not evaluate AutoQG's QA generation stage.

8 Conclusions

This paper introduced **AutoQG**, a fully automated multi-agent framework that transforms scientific papers into ontology-guided knowledge graphs and generates evidence-traceable, multi-hop QA pairs without any manual annotation or model fine-tuning. Unlike prior approaches that treat QA generation as a flat prompting task, AutoQG introduces a structured intermediate representation—the academic knowledge graph—as an auditable backbone that explicitly captures entities, typed relations, and their source-text provenance. Through a three-stage *construct–evaluate–generate* paradigm, the framework coordinates three specialized agents: the KG Extraction Agent applies section-aware, ontology-constrained prompting to produce provenance-anchored triples; the KG Evaluation Agent performs multi-dimensional scoring and iterative feedback-driven refinement to ensure graph accuracy and consistency; and the QA Generation Agent traverses validated reasoning paths to synthesize schema-constrained, evidence-grounded QA pairs. This design directly addresses the persistent tension between scalability and quality in academic QA generation—a challenge that has long limited both crowdsourced and LLM-based approaches. Applied to over 4,000 papers, AutoQG constructed 243k structured triples and produced two large-scale benchmarks—**AutoQG^{20k}** and **AutoQG^{7k}**—achieving leading evidence traceability (Evid = 9.6) and strong reasoning depth (Reas = 8.5). Furthermore, supervised fine-tuning on AutoQG-7k yielded a 34.4% absolute accuracy gain over zero-shot baselines, demonstrating the downstream utility of evidence-grounded QA data at scale.

Beyond effectiveness, AutoQG demonstrates robust cross-domain generalization with quality variations within ±0.3 across PubMed and Physics domains, and supports low-cost open-source deployment at approximately 0.06 USD using Qwen-7B and DeepSeek-V2.5. These results establish that structured decomposition, ontology guidance, and evidence-grounded evaluation can unlock the potential of LLMs for large-scale academic knowledge extraction, providing a foundation for next-generation scientific AI systems that are factually grounded, citation-aware, and capable of transparent reasoning.

Several promising directions remain for future exploration. First, extending AutoQG to multi-paper joint processing, generating knowledge graphs and QA pairs simultaneously from multiple documents, presents a non-trivial challenge. The current framework deliberately processes single papers, as their self-contained referential and logical coherence ensures extracted triples are grounded in co-located evidence and multi-hop reasoning remains verifiable within a unified context. As our benchmarks demonstrate, this boundary-respecting design already yields high-quality, evidence-traceable QA pairs at scale. Generalising to multiple papers introduces difficulties like cross-paper coreference resolution, heterogeneous ontology alignment across corpora, and constructing reasoning chains with broadly distributed, hard-to-verify evidence. Designing a robust pipeline preserving traceability and coherence under these conditions remains future work. Second, scientific communication increasingly relies on non-textual modalities, since figures, tables, equations, and diagrams convey information frequently absent from running text. Incorporating multi-modal understanding to extract structured knowledge from charts and schematics and generate questions requiring cross-modal reasoning represents another important avenue. Together, multi-paper and multi-modal extensions will substantially broaden AutoQG toward a comprehensive scientific knowledge mining system.

Ethical statements

During the preparation of this work, the authors used OpenAI Codex (GPT-5; accessed 28 August 2026) for manuscript formatting and language refinement. The authors reviewed and edited all content produced with the assistance of this tool, verified its accuracy, and take full responsibility for the integrity and originality of the final manuscript. This work represents the authors' own intellectual contribution, and no AI tool is credited as an author.

Author contributions

The authors confirm their contributions to this study as follows: conceptualization, writing − review and editing: Shen X, Qi G; methodology, software, writing − original draft preparation: Shen X; validation: Zhang P, Shen X, Qi G; formal analysis: Zhang P, Yang D; investigation: Guo H, Hu H, Yang D, Yang Y; data curation: Zhang P, Yang D, Yang Y, Yang M; visualization: Guo H, Hu H; supervision, funding acquisition: Qi G. All authors reviewed the results and approved the final version of the manuscript.

Data availability

The derived metadata, hashes, prompts, reasoning paths, evidence mappings, split manifests, and evaluation scripts supporting this study are available from the corresponding author upon reasonable request and are planned for public release upon publication. Full source-paper text is not redistributed where copyright or licensing terms do not permit redistribution.

Acknowledgments

This work was supported by the National Social Science Foundation Key Program of China (No. 23&ZD222) and the National Natural Science Foundation of China (No. 62476058). We thank the Big Data Computing Center of Southeast University for providing facility support for the numerical calculations in this paper.

Conflict of interest

The authors have no competing interests to declare that are relevant to the content of this article.

Supplementary information accompanies this paper online at: https://doi.org/10.48130/ker-0026-0017.

Dates

Received 8 May 2026; Revised 1 July 2026; Accepted 23 August 2026; Published online 9 October 2026

https://doi.org/10.48130/ker-0026-0017

References

[1] OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, et al. 2023. GPT-4 technical report. *arXiv* Preprint:2303.08774

[2] Grattafiori A, Dubey A, Jauhri A, Pandey A, Kadian A, et al. 2024. The llama 3 herd of models. *arXiv* Preprint:2407.21783

[3] Georgiev P, Lei VI, Burnell R, Bai L, Gulati A, et al. 2024. Gemini 1.5: unlocking multimodal understanding across millions of tokens of context. *arXiv* Preprint:2403.05530

[4] Lee T, Yoon C, Jang K, Lee D, Song M, et al. 2025. ETHIC: evaluating large language models on long-context tasks with high information coverage. *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). April 2025. Albuquerque, New Mexico*. USA: Association for Computational Linguistics. pp. 5497–5512 doi: 10.18653/v1/2025.naacl-long.283

[5] Wang X, Liu J, Xiao Y, Ning J, Liu L, et al. 2025. THE-Tree: can tracing historical evolution enhance scientific verification and reasoning? *arXiv* Preprint:2506.21763

[6] Schryen G, Marrone M, Yang J. 2025. Exploring the scope of generative AI in literature review development. *Electronic Markets* 35(1):13

[7] Lála J, O'Donoghue O, Shtedritski A, Cox S, Rodriques SG, et al. 2023. PaperQA: retrieval-augmented generative agent for scientific research. *arXiv* Preprint:2312.07559

[8] Asai A, He J, Shao R, Shi W, Singh A, et al. 2024. OpenScholar: synthesizing scientific literature with retrieval-augmented lms. *arXiv* Preprint:2411.14199

[9] Auer S, Barone DAC, Bartz C, Cortes EG, Jaradeh MY, et al. 2023. The SciQA scientific question answering benchmark for scholarly knowledge. *Scientific Reports* 13(1):7240

[10] Wang X, Hu Z, Lu P, Zhu Y, Zhang J, et al. 2024. SciBench: evaluating college-level scientific problem-solving abilities of large language models. *arXiv* Preprint:2307.10635

[11] Vladika J, Matthes F. 2024. Improving health question answering with reliable and time-aware evidence retrieval. *Findings of the Association for Computational Linguistics: NAACL 2024. June 2024. Mexico City, Mexico*. USA: Association for Computational Linguistics. pp. 4752–4763 doi: 10.18653/v1/2024.findings-naacl.295

[12] Fang J, Meng Z, MacDonald C. 2024. TRACE the evidence: constructing knowledge-grounded reasoning chains for retrieval-augmented generation. *Findings of the Association for Computational Linguistics: EMNLP 2024. November 2024. Miami, Florida, USA*. USA: Association for Computational Linguistics. pp. 8472–8494 doi: 10.18653/v1/2024.findings-emnlp.496

[13] Lan Y, He G, Jiang J, Jiang J, Zhao WX, et al. 2021. A survey on complex knowledge base question answering: methods, challenges and solutions. *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, Survey Track. 19–26 August 2021. Montreal, Canada*. USA: International Joint Conferences on Artificial Intelligence Organizationpp. pp. 4483–4491 doi: 10.24963/ijcai.2021/611

[14] Zhang N, Gui T, Nan G. 2022. Efficient and robust knowledge graph construction. *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Tutorial Abstracts. November 2022. Taipei*. USA: Association for Computational Linguistics. pp. 1–7 doi: 10.18653/v1/2022.aacl-tutorials.1

[15] Dong J, Zhang Q, Huang X, Duan K, Tan Q, et al. 2023. Hierarchy-aware multi-hop question answering over knowledge graphs. *Proceedings of the ACM Web Conference 2023. 30 April–4 May 2023, Austin, TX, USA*. USA: Association for Computing Machinery. pp. 2519–2527 doi: 10.1145/3543507.3583376

[16] Yih WT, Chang MW, He X, Gao J. 2015. Semantic parsing via staged query graph generation: question answering with knowledge base. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. 26–31 July 2015. Beijing, China*. USA: Association for Computational Linguistics. pp. 1321–1331 https://aclanthology.org/P15-1128.pdf

[17] Pampari A, Raghavan P, Liang J, Peng J. 2018. emrQA: a large corpus for question answering on electronic medical records. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. October−November 2018. Brussels, Belgium*. USA: Association for Computational Linguistics. pp. 2357–2368 doi: 10.18653/v1/D18-1258

[18] Jin Q, Dhingra B, Liu Z, Cohen W, Lu X. 2019. PubMedQA: a dataset for biomedical research question answering. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). November 2019. Hong Kong, China*. USA: Association for Computational Linguistics. pp. 2567–2577 doi: 10.18653/V1/D19-1259

[19] Saikh T, Ghosal T, Mittal A, Ekbal A, Bhattacharyya P. 2022. *ScienceQA*: a novel resource for question answering on scholarly articles. *International Journal on Digital Libraries* 23(3):289−301

[20] Wellawatte GP, Guo H, Lederbauer M, Borisova A, Hart M, et al. 2025. ChemLit-QA: a human evaluated dataset for chemistry RAG tasks. *Machine Learning: Science and Technology* 6(2):020601

[21] Phan L, Gatti A, Han Z, Li N, Hu J, et al. 2025. Humanity's last exam. *arXiv* Preprint:2501.14249

[22] Bai Y, Tu S, Zhang J, Peng H, Wang X, et al. 2025. LongBench v2: towards deeper understanding and reasoning on realistic long-context multitasks. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). July 2025. Vienna, Austria*. USA: Association for Computational Linguistics. pp. 3639–3664 doi: 10.18653/v1/2025.acl-long.183

[23] Skarlinski MD, Cox S, Laurent JM, Braza JD, Hinks M, et al. 2024. Language agents achieve superhuman synthesis of scientific knowledge. *arXiv* Preprint:2409.13740

[24] Lee Y, Lee K, Park S, Hwang D, Kim J, et al. 2023. QASA: advanced question answering on scientific articles. *Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research (PMLR) 202. 23–29 July 2023. Honolulu, Hawaii, USA*. USA: PMLR. pp. 19036–19052 https://openreview.net/pdf?id=5ud0h8OXwD

[25] Wan Y, Liu Y, Ajith A, Grazian C, Hoex B, et al. 2024. SciQAG: a framework for auto-generated science question answering dataset with fine-grained evaluation. *arXiv* Preprint:2405.09939

[26] Wei X, Cui X, Cheng N, Wang X, Zhang X, et al. 2023. ChatIE: zero-shot information extraction via chatting with ChatGPT. *arXiv* Preprint:2302.1020

[27] Yu J, Tang Y, Feng K, Liang L, Zhang Q, et al. 2026. SciCUEval: a comprehensive dataset for evaluating scientific context understanding in large language models. *Scientific Data* 13:530

[28] Kim MW, Park HB, Ahn HJ, Park WR, Jeon JW, et al. 2025. AutoPaperBench: an MLLM-based framework for automatic generation of paper understanding evaluation benchmarks. *Electronics* 14(6):1175

[29] Luan Y, He L, Ostendorf M, Hajishirzi H. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. October–November 2018. Brussels, Belgium*. USA: Association for Computational Linguistics. pp. 3219–3232 doi: 10.18653/v1/D18-1360

[30] Yao Y, Ye D, Li P, Han X, Lin Y, et al. 2019. DocRED: a large-scale document-level relation extraction dataset. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. July 2019. Florence, Italy*. USA: Association for Computational Linguistics. pp. 764–777 doi: 10.18653/v1/p19-1074

[31] Xue L, Zhang D, Dong Y, Tang J. 2024. AutoRE: document-level relation extraction with large language models. *arXiv* Preprint:2403.14888

[32] Wan Z, Cheng F, Mao Z, Liu Q, Song H, et al. 2023. GPT-RE: in-context learning for relation extraction using large language models. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. December 2023. Singapore. USA: Association for Computational Linguistics. pp. 3534–3547 doi: 10.18653/V1/2023.EMNLP-MAIN.214.

[33] Li P, Sun T, Tang Q, Yan H, Wu Y, et al. 2023. CodeIE: large code generation models are better few-shot information extractors. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). July 2023. Toronto, Canada*. USA: Association for Computational Linguistics. pp. 15339–15353 doi: 10.18653/V1/2023.ACL-LONG.855.

[34] Jiao Y, Zhong M, Li S, Zhao R, Ouyang S, et al. 2023. Instruct and extract: instruction tuning for on-demand information extraction. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. December 2023. Singapore*. USA: Association for Computational Linguistics. pp. 10030–10051 doi: 10.18653/V1/2023.EMNLP-MAIN.620.

[35] Lu Y, Liu Q, Dai D, Xiao X, Lin H, et al. 2022. Unified structure generation for universal information extraction. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). May 2022. Dublin, Ireland*. USA: Association for Computational Linguistics. pp. 5755–5772 doi: 10.18653/v1/2022.acl-long.395

[36] Li Z, Zeng Y, Zuo Y, Ren W, Liu W, et al. 2024. KnowCoder: coding structured knowledge into llms for universal information extraction. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). August 2024. Bangkok, Thailand*. USA: Association for Computational Linguistics. pp. 8758–8779 doi: 10.18653/V1/2024.ACL-LONG.475.

[37] Sainz O, García-Ferrero I, Agerri R, de Lacalle OL, Rigau G, et al. 2024. GoLLIE: annotation guidelines improve zero-shot information-extraction. *arXiv* Preprint:2310.03668

[38] Gao S, Yu K, Yang Y, Yu S, Shi C, et al. 2025. Large language model powered knowledge graph construction for mental health exploration. *Nature Communications* 16(1):7526

[39] Han R, Yang C, Peng T, Tiwari P, Wan X, et al. 2023. An empirical study on information extraction using large language models. *arXiv* Preprint:2305.14450

[40] Liu S, Li Y, Li J, Yang S, Lan Y. 2024. Unleashing the power of large language models in zero-shot relation extraction via self-prompting. *Findings of the Association for Computational Linguistics: EMNLP 2024. November 2024. Miami, Florida, USA*. USA: Association for Computational Linguistics. pp. 13147–13161 doi: 10.18653/v1/2024.findings-emnlp.769

[41] Bai J, Fan W, Hu Q, Zong Q, Li C, et al. 2026. AutoSchemaKG: autonomous knowledge graph construction through dynamic schema induction from web-scale corpora. *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). July 2026. San Diego, California, USA*. USA: Association for Computational Linguistics. pp. 20557–20584 doi: 10.18653/v1/2026.acl-long.942

[42] Bai J, Bai S, Chu Y, Cui Z, Dang K, et al. 2023. Qwen technical report. *arXiv* Preprint:2309.16609

[43] DeepSeek-AI. 2024. DeepSeek-V2: a strong, economical, and efficient mixture-of-experts language model. *arXiv* Preprint:2405.04434

Copyright: © 2026 by the author(s). Published by Maximum Academic Press, Fayetteville, GA. This article is an open access article distributed under Creative Commons Attribution License (CC BY 4.0), visit https://creativecommons.org/licenses/by/4.0/.