Supplementary Materials for "AutoQG: An Automated Framework for Evidence-Traceable Question Generation via Ontology-Guided Knowledge Graph Construction"

Xiaoli Shen[a], Guilin Qi[a,*], Peng Zhang[b], Dingsheng Yang[a], Yang Yang[c], Muyan Yang[a], HaoWen Guo[a] and HuiKang Hu[a]

[a]School of Computer Science and Engineering, Southeast University, Nanjing, Jiangsu, China
[b]School of Life Science and Technology, Southeast University, Nanjing, Jiangsu, China
[c]School of Artificial Intelligence, Southeast University, Nanjing, Jiangsu, China

ABSTRACT

Supplementary File 1: Supplementary material to this study.

A. Supplementary Materials

A.1. AI-Assisted Manuscript Preparation

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.

A.2. KG Annotation Details

To establish the gold standard for our knowledge graph evaluation, we recruited seven master's students with strong backgrounds in natural language processing and machine learning to serve as annotators. Each annotator was assigned five papers and instructed to manually construct knowledge graphs in TTL format, following the predefined ontology detailed in Appendix A.6. The annotation process was overseen by a domain expert to ensure the quality and consistency of the output. The resulting manually annotated graphs served as the gold dataset for evaluating the performance of our KG Extraction Agent against various baselines.

A.3. Gold Set Representativeness

The 30-paper gold set is designed to characterize coverage across scientific NLP-adjacent papers rather than to support a random-sampling statistical claim. Table A.1 summarizes domain, year, venue, paper type, and length coverage.

A.4. Supplementary Diagnostic Tables

This appendix organizes the detailed diagnostic results into three groups: baseline fairness settings, external extraction validation, and controlled ablations.

A.1 Baseline Fairness Settings

Table A.2 documents how each compared method differs in input, target schema, and available components, ensuring all comparisons in the main text are made under controlled conditions.

A.2 External Extraction Validation

Tables A.3 and A.4 report source-grounded diagnostic scores for the adapted agentic baselines and SciLEAD TDMR extraction results respectively. Both tables use the same 30-paper manually annotated gold set.

*Corresponding author: Guilin Qi
gqi@seu.edu.cn (G. Qi)

Table A.1
Representativeness statistics for the 30-paper human-annotated gold set. Counts are descriptive coverage statistics, not random-sampling estimates.

Aspect	Distribution	Coverage Note
Domain	Knowledge Graph 8; LLM Reasoning 8; Retrieval/QA 7; Information Extraction 2; Multimodal 2; Table/Document Understanding 2; Other 1	Covers KG, reasoning, retrieval, IE, multimodal, and document settings.
Year	2024: 18; 2025: 8; 2023: 2; 2021: 1; 2019: 1	Primarily recent papers, with limited older anchor papers.
Venue	arXiv 19; ICLR 4; EMNLP 3; Findings of ACL 2; ACL 1; BioNLP Workshop 1	Mixes preprints, conferences, and workshop papers.
Paper type	Method 22; Dataset/Benchmark 3; System 3; Empirical Analysis 2	Method papers dominate, matching the extraction target.
Length	11–34 pages; median 18.5 pages; extracted text 35,890–106,812 characters	Includes short and long scientific documents.

Table A.2
Baseline fairness settings. All systems use the same source papers within each diagnostic comparison; only the availability of ontology constraints, evidence preservation, refinement, and reasoning-path construction differs.

Method	Input	Target Schema	Ontology	Evidence	Refine	QA Path
Direct prompting	Full/sectioned paper text	Natural-language QA format	No	Post-hoc	No	No
LangExtract	Gold-set paper JSON subset	Gold TTL target	Partial	Yes	No	No
EPT-LLM	Gold-set paper JSON subset	Gold TTL target	Partial	Yes	No	No
AutoQG	Same sectioned papers	Academic KG + QA schema	Yes	Yes, per hop	Yes	Yes

Table A.3
Source-grounded diagnostic scores for adapted agentic extraction baselines on the 30-paper manually annotated gold set. These are controlled diagnostic baselines, not official reproductions of the original systems.

Method	Papers	Schema Conf.	Evid. Cov.	Grounding
LangExtract	30/30	92.41	100.00	91.35
EPT-LLM	30/30	93.53	100.00	91.91

A.3 Controlled Component Ablations

Tables A.5 and A.6 isolate the contribution of each AutoQG component by removing it individually. Table A.5 reports KG-level metrics; Table A.6 reports QA-level judge scores for the same settings.

A.5. QA Pair Examples

Three-Hop QA Example

Reasoning Path:

$$\text{KISA} \xrightarrow{\texttt{outperforms}} \text{SelfAttn} \xrightarrow{\texttt{evaluatedOn}} \text{SemEval2010Task8} \xrightarrow{\texttt{measures}} \text{F1}$$

Context:

Table A.4
SciLEAD TDMR extraction on the 30-paper manually annotated gold set. All methods use the same paper inputs, model, and gold annotations; this is not a full leaderboard reproduction.

Method	Papers	ETM-TDMR	ETM-TDM	Task	Dataset	Metric	Result	Evid.
AutoQG	30/30	27.41	63.91	90.00	82.76	76.09	35.56	100.00
LangExtract	30/30	10.32	33.20	76.92	62.89	60.00	15.87	100.00
EPT/SciLead	30/30	23.53	44.53	78.95	72.84	70.33	27.94	100.00

Table A.5
Controlled diagnostic ablation on the same 30-paper manually annotated gold set. The table isolates degradation patterns after removing individual components; the full AutoQG row is reported in the main evaluation tables rather than inserted as a same-run row.

Setting	Valid	Schema	Evid. Cov.	KG Judge	QA Pairs	QA Papers	Path Evid.
w/o ontology guidance	30	58.2%	98.9%	8.66	762	30/30	100.0%
w/o section-aware extraction	30	68.3%	70.8%	8.80	12	3/30	100.0%
w/o iterative refinement	30	76.8%	99.1%	8.47	1,049	30/30	100.0%
w/o evidence/provenance	30	64.3%	0.03%	8.29	1,795	30/30	0.0%

Table A.6
Stratified QA-level judge evaluation for the controlled diagnostic ablations. We score all 12 QA pairs from the section-agnostic variant and 100 sampled QA pairs from each remaining variant, using a two-model judge ensemble (GPT-4o and Qwen2.5-72B).

Setting	n	Rel	Ans	Distr	Flu	Reas	Evid	Total
w/o ontology guidance	100	9.40	9.09	8.44	9.38	7.97	9.02	8.94
w/o section-aware extraction	100	9.52	9.04	8.52	9.42	8.10	8.96	8.99
w/o iterative refinement	100	9.31	9.03	8.40	9.33	7.96	8.84	8.87
w/o evidence/provenance	100	8.61	8.53	8.23	9.37	8.05	7.75	8.48

- *"The integrated models improve performance by at least 0.9 F1 score and achieve new state-of-the-art results."*
- *"Besides TACRED, another dataset called SemEval2010-Task8 ... is used to evaluate the generalization ability."*
- *"We use the official macro-averaged F1 score as evaluation metric."*

Question:
Which model outperformed the Self-Attention model on the SemEval2010 Task 8 by utilizing which specific evaluation metric?

Options:

A. The KISA model outperformed the Self-Attention model using the official macro-averaged F1 score as the evaluation metric.

B. The KISA model surpassed the Self-Attention model using the average precision score on the SemEval2010 Task 8.

C. The Self-Attention model was outperformed by the KISA model, which used the micro-averaged recall score for evaluation.

D. The Self-Attention model was outperformed by the KISA model evaluated by the area under the ROC curve on SemEval2010 Task 8.

Correct Answer: A

Explanation:
The KISA model outperformed the Self-Attention model on SemEval2010 Task 8 using the official macro-averaged F1 score, as explicitly indicated by the reasoning path and evidence.

Two-Hop QA Example

Reasoning Path:

CameraSystemForFaceDetection $\xrightarrow{\texttt{uses}}$ RFIDEnergyHarvesting $\xrightarrow{\texttt{type}}$ Technology

Context:

- *"The camera system runs solely on energy harvested from RFID readers."*
- *"The camera system runs solely on energy harvested from RFID readers."*

Question:
What type of technology enables the camera system described in the paper to function solely on energy harvested from RFID readers?

Options:

A. RFID-based energy harvesting technology, which captures and converts energy from RFID systems.

B. Solar panel technology, which converts sunlight into electrical energy to power devices.

C. Piezoelectric energy harvesting technology, which generates power from mechanical stress.

D. Thermoelectric generators, which produce electricity from temperature differences.

Correct Answer: A

Explanation:
The camera system is explicitly described as operating with energy harvested from RFID readers. Therefore, the correct answer is RFID-based energy harvesting technology, consistent with the reasoning path and supporting evidence.

One-Hop QA Example

Reasoning Path:
LuisCeze $\xrightarrow{\texttt{type}}$ Author

Context:

- *"Amrita Mazumdar, Thierry Moreau, Armin Alaghi, Luis Ceze, Mark Oskin, Sung Kim, Meghan Cowan, Visvesh Sathe are mentioned as contributors."*

Question:
Which of the following accurately describes the professional role of Luis Ceze in the context provided in the paper?

Options:

A. Luis Ceze is mentioned as an author among other contributors in the paper.

B. Luis Ceze is recognized as the sole contributor and leading expert in the paper.

C. Luis Ceze is the primary editor responsible for reviewing and editing the paper.

D. Luis Ceze is a guest speaker who provided insights for the paper's development.

Correct Answer: A

Explanation:
Based on the knowledge graph path and context, Luis Ceze is listed along with other contributors as an author. No evidence supports alternative roles such as editor or guest speaker.

A.6. KG Construction Agent Prompts

To ensure transparency and reproducibility, we provide the exact schema and prompt templates used in our framework. The schema defines the set of entity and relation types used to construct academic knowledge graphs, while the prompts demonstrate how section-aware, ontology-guided extraction and evaluation are operationalized in practice. These resources allow others to replicate our pipeline and adapt it to new domains.

Knowledge Graph Schema

Entity Types:

- **Core Academic**: Models (GPT-4, BERT, Llama-2), Datasets (SQuAD, GLUE), Metrics (Accuracy, F1), Methods (Fine-tuning, LoRA), Algorithms (Attention, SGD), Architectures (Transformer).
- **Research Context**: Papers, Authors, Organizations (OpenAI, Google), Tools (PyTorch, TensorFlow), Venues (ACL, NeurIPS).
- **Measurement**: Scores, Parameters (7B, 175B), Configurations (hyperparameters, settings).

Relation Types:

- **Performance/Evaluation**: `:evaluatedOn`, `:achievesScore`, `:measures`, `:outperforms`, `:comparedWith`
- **Development/Creation**: `:basedOn`, `:improves`, `:proposes`, `:developedBy`, `:implementedIn`
- **Training/Application**: `:trainedOn`, `:finetunedOn`, `:uses`, `:appliedTo`, `:optimizedFor`
- **Research/Citation**: `:addresses`, `:publishedIn`, `:citedBy`
- **Structural**: `:partOf`, `:enabledBy`, `:resultIn`, `:relatedTo`

Mandatory Metadata:

```
:contextText "original_text_snippet"
```

Knowledge Extraction Prompt

Role: Expert extractor for academic knowledge. Output ontology-aligned triples for multi-hop reasoning and QA.
Selective Extraction ($\leq$ 200 Triplets):

- Focus on major models, datasets, metrics, methods, and results.
- Extract only significant scores, parameters, comparisons, and settings.
- Skip minor mentions and redundant details.

Triple Format:

```
:Entity1 :relation :Entity2 ;
    :contextText "Original snippet..." .
```

Examples:

```
:GPT-4 :evaluatedOn :WinoGrande ;
    :contextText "GPT-4 was evaluated on WinoGrande dataset" .
:GPT-4 :achievesScore "94.2"^^xsd:float ;
    :contextText "GPT-4 achieved 94.2
:ThisPaper :proposes :LogitLens ;
    :contextText "This paper proposes LogitLens" .
```

Success Criteria: 150–200 high-quality triples, schema-aligned, evidence-traceable, and relevant for QA.

A.7. KG Evaluation Agent Prompt

To complement the extraction schema, we also release the evaluation prompt used for graph-level quality assessment. The evaluator is designed to score an entire sectioned knowledge graph rather than individual triples, providing structured diagnostics and a weighted final score across five dimensions (domain fit, accuracy, consistency, completeness, and granularity). This ensures that graph evaluation is transparent, reproducible, and aligned with downstream QA requirements.

Sectioned KG Evaluator Prompt

Role. You are an evaluator for a sectioned knowledge graph extracted from an academic paper. Score the *entire graph* (not individual triples) using **only** the given question, ontology, and sectioned evidence. Return **JSON only** (no extra text).

Inputs

- Task/Question: `{QUESTION}`
- Ontology & constraints (optional): `{ONTOLOGY_OR_RULES}`
- Sectioned KG snippet (RDF/Turtle-like): `{SECTIONED_TURTLE_BLOCK}`

Notes: Entities/relations may contain `:sourceSection`, `:sourceChunk`, and `:contextText`.

What you must do

- Parse the snippet into a graph (nodes/edges) with per-edge provenance `{section, chunk, evidence}`.
- Section-aware evidence: prefer evidence from the same `sourceSection`; if multiple, choose the strongest/clearest `contextText`.
- Compute graph-level diagnostics:
  - `triple_count`, `entity_count`, `section_coverage` (present / missing),
  - `conflict_count`, `missing_slot_rate`, `redundancy_rate`,
  - `granularity_notes`, `evidence_coverage` (portion of edges with `contextText`).

Scoring dimensions (0–10) & section rules

- **Domain Fit** – Alignment to task/question and paper domain across sections. Abstract/Conclusion weight high-level relevance; Methods/Results weight task-level relevance.
- **Accuracy** – Proportion of edges explicitly entailed by evidence; penalize speculation unless typical for the section (e.g., Discussion).
- **Consistency** – Cross-section coherence: contradictions, unit/time/version mismatches, ontology violations.
- **Completeness** – Coverage of section-appropriate slots/metadata; use `missing_slot_rate`.
- **Granularity** – Appropriate detail (normalized terms, alias resolution, versions, metric names/values); penalize systematically coarse/fine patterns per section norms.

$$\begin{aligned}\texttt{final_score} = {} & 0.20 \times \texttt{domain_fit} + 0.20 \times \texttt{accuracy} \\ & + 0.20 \times \texttt{consistency} + 0.20 \times \texttt{completeness} \\ & + 0.20 \times \texttt{granularity}\end{aligned}$$

(round to one decimal).

Output (JSON only)

Return a single JSON object. Keep reasons concise (each reason $\leq$ 30 chars; `summary_advice` $\leq$ 120 chars; each `top_fix` $\leq$ 60 chars).

```
{
  "meta": {
    "triple_count": <int>,
    "entity_count": <int>,
    "section_coverage": {
      "present": ["Abstract","Methods","..."],
      "missing": ["Results","..."]
    },
    "evidence_coverage": <0-1>,
    "conflict_count": <int>,
    "redundancy_rate": <0-1>,
    "missing_slot_rate": <0-1>,
    "granularity_notes": ["...","..."]
  },
  "scores": {
    "domain_fit":   {"score": <0-10>, "reason": "$\leq$30 chars"},
    "accuracy":     {"score": <0-10>, "reason": "$\leq$30 chars"},
    "consistency":  {"score": <0-10>, "reason": "$\leq$30 chars"},
    "completeness": {"score": <0-10>, "reason": "$\leq$30 chars"},
    "granularity":  {"score": <0-10>, "reason": "$\leq$30 chars"}
  },
  "final_score": <0-10 one-decimal>,
  "summary_advice": "$\leq$ 120 chars, prioritized",
  "top_fixes": [
    "Fix 1 ($\leq$ 60 chars)",
    "Fix 2 ($\leq$ 60 chars)",
    "Fix 3 ($\leq$ 60 chars)"
  ]
}
```

A.8. QA Generation Agent Prompt

QA Generation Prompt (System)

Role: Expert generator of educational multiple-choice questions grounded in academic paper knowledge graphs.
Task: Produce *one* high-quality MCQ with four options (A–D) from a KG reasoning path anchored to a specific paper section. This prompt instantiates the 3-hop setting used in our prompt library; the same template is adapted to 1-hop and 2-hop generation by replacing the hop-count constraint and the supplied path length.
Constraints:

- Require multi-hop reasoning that follows the provided path end-to-end.
- Create exactly four options with *only one* correct answer.
- Keep options plausible yet clearly distinguishable; include mildly misleading overlaps.
- Use an academic, precise, objective style in **English**.
- Ground the question in the given section context; ensure it is answerable from the information provided.
- Prefer phenomena that deviate from common knowledge and are surprising.
- You may use a term a domain expert would understand but that is likely ambiguous for an LLM; this term must *not* appear elsewhere in the text.
- Each option must contain **more than 10 words**.
- Do not introduce facts beyond the provided path and context; all answer-critical information must be derivable from the supplied evidence.
- Make the item challenging but resolvable without external resources beyond the provided inputs.

Output Format (JSON only):

```
{
  "question": "Clear, specific question text",
  "options": {
    "A": "First option (must be longer than 10 words)",
    "B": "Second option (must be longer than 10 words)",
    "C": "Third option (must be longer than 10 words)",
    "D": "Fourth option (must be longer than 10 words)"
  },
  "correct_answer": "A/B/C/D",
  "explanation": "Brief explanation of why the answer is correct"
}
```

Success Criteria: One valid-JSON output; reasoning explicitly implied by the supplied path; options long, plausible, and non-ambiguous; single correct answer with a concise justification.

QA Generation Prompt (User)

Inputs: Section name, knowledge path, and supporting context from the paper.
Instruction: Based on the following knowledge graph path and the accompanying section-level context, generate *one* multiple-choice question that requires reasoning through the entire supplied path. The item should test understanding of entities and relations involved. For 1-hop and 2-hop settings, the same instruction is used with the corresponding path length.
Template (model receives the following fields):

```
KNOWLEDGE PATH:
{path_description}=
CONTEXT FROM PAPER:
{context_text}
SECTION: {section}
```

Return: Respond with *valid JSON only* following the schema in the System prompt.

A.9. More evaluation Results on LLMs

To provide a more comprehensive view of our benchmark, we extend the main evaluation by reporting additional results on a broader set of large language models in the Appendix. Consistent with the main paper, our analysis emphasizes three complementary dimensions: (i) **Difficulty Control**, measured by model performance across 1–3 hop reasoning paths; (ii) **Evidence Traceability**, quantified by accuracy with and without providing supporting paths; and (iii) **Model Robustness**, evaluated through comparisons between reasoning-oriented and base LLMs, as shown in Table A.7.

Table A.7
Extended evaluation of the generated QA benchmark under two settings: *AutoQG*^{20k} (left block) and *AutoQG*^{7k} (right, QA-Pair only). Results are Strict Accuracy (%) on QA pairs, difficulty-controlled hops, and accuracy with/without evidence.

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

A.10. Case Study on LLM Failure Cases

Example QA Pair

Question:
In the study described in the *researchcontextandstudydesign* section, how was the approach for balancing functional requirements and non-functional requirements (NFRs) presented to the experts?
Options:

A. The approach was embedded within the closed questions of the questionnaire on a 4-point Likert scale.

B. Experts were invited to a workshop where the approach was discussed in detail before starting the questionnaire.

C. It was detailed at the beginning of the study to introduce experts to the main focus of research.

D. It was presented in the third part of a questionnaire that included both open and closed questions.

Correct Answer: D
Reasoning Path:
Questionnaire → Open and Closed Questions → Three Parts → Approach for Balancing Functional and NFRs
Context:
"The questionnaire comprised 20 open questions and 4 closed questions... Finally, in the third part we presented the experts our approach for balancing functional and NFRs."

Case Study on Incorrect Answers by LLMs. We analyze a representative failure case in which most LLMs struggle to answer correctly. This example requires a three-hop reasoning chain that links the questionnaire to its division into three parts and finally to the presentation of the proposed approach, yet many models fail to follow this path and instead latch onto superficial cues. Distractor options are particularly challenging, as references to Likert scales or common research practices (e.g., introducing the method at the beginning) appear plausible and mislead models into incorrect selections. Success thus hinges on capturing the precise detail that the approach was presented in the *third part* of the questionnaire, which most models overlook. This case was manually checked to ensure that the explanation, evidence, and correct option are aligned. This case highlights that our benchmark effectively probes weaknesses in multi-hop reasoning, evidence grounding, and robustness against distractors.

A.11. QA Human Evaluation Details

We conducted a double-blind human evaluation to rigorously assess the quality of the generated QA pairs. The evaluation was performed by a separate group of annotators who were blind to the source of each QA pair. We collected QA pairs from multiple sources, including our framework and various baselines, and presented them to the annotators in a randomized, interleaved fashion. This approach prevented annotators from inferring the source and ensured an unbiased assessment. The annotators were provided with a detailed scoring rubric to evaluate each QA pair across six key dimensions: Relevance, Answerability, Distractor Quality, Fluency, Reasoning Depth, and Evidence Traceability. The scoring criteria for each dimension are summarized in the table A.9.

Table A.8 summarizes the evaluation transparency controls added for the revision. Human annotators and LLM judges score the same conceptual dimensions, but the LLM judges additionally receive structured inputs that make the evidence and reasoning path explicit. Scores are parsed per dimension and averaged across judges; disagreement is inspected when a dimension differs sharply across judges.

Table A.8
Evaluation transparency summary for human evaluation and LLM-as-a-Judge.

Evaluator	Sample Scope	Blind?	Inputs Seen	Metrics	Aggregation
Human annotators	Interleaved QA pairs from AutoQG and baselines	Yes	Question, answer/options, evidence/context, rubric	Rel, Ans, Distr, Flu, Reas, Evid	Mean score; agreement checked
LLM-as-a-Judge	Same evaluation dimensions across systems	System labels hidden where applicable	Question, answer/options, reasoning path, evidence spans, rubric	Rel, Ans, Distr, Flu, Reas, Evid; KG Dom/Acc/Con/Com/Gra	Mean across heterogeneous judges
KG judge	Sectioned KG snippets and ontology	N/A	Ontology, sectioned triples, `contextText`, source sections	Domain fit, accuracy, consistency, completeness, granularity	Dimension mean and final score

Dimension	Definition	Scoring Guideline (0–10, with Example)
Relevance	Alignment of QA with context, path, and evidence.	0: Irrelevant; 10: Fully relevant. High score if QA targets core methods/results, not background.
Answerability	Answer can be derived from provided context/path.	0: Unanswerable; 10: Clearly answerable. Penalize if external knowledge is needed.
Distractor Quality	Plausibility and misleadingness of distractors.	0: Trivial/repetitive; 10: Plausible and non-trivial.
Fluency	Clarity and naturalness of language.	0: Unclear; 10: Fluent and academic. High score if concise and grammatically correct.
Reasoning Depth	Need for multi-hop reasoning/evidence integration.	0: Fact recall; 10: Multi-hop reasoning. Reflects difficulty control.
Evidence Traceability	Explicit grounding in source text and path.	0: No link; 10: Explicitly linked. High score if tied to specific findings/experiments.

Table A.9
Evaluation dimensions for QA pairs with scoring rubric (0–10 scale).

A.12. The prompt of KG Construction of Representative LLMs for KG evaluation

Knowledge Graph Schema

Entity Types:

- **Core Academic**: Models (GPT-4, BERT, Llama-2, Transformer), Datasets (WinoGrande, SQuAD, GLUE), Metrics (Accuracy, F1, BLEU, ROUGE, Perplexity), Methods (Fine-tuning, LoRA, Pre-training), Algorithms (Attention, Backpropagation, SGD), Architectures (Encoder-Decoder, Attention-only).
- **Research Context**: Papers, Authors, Organizations (OpenAI, Google, Meta), Tools (PyTorch, TensorFlow, Hugging Face), Venues (ACL, NeurIPS, journals).
- **Measurement**: Scores, Parameters (7B, 175B), Configurations (hyperparameters, experimental settings).

Relation Types:

- **Performance & Evaluation**: `:evaluatedOn`, `:achievesScore`, `:measures`, `:outperforms`, `:comparedWith`
- **Development & Creation**: `:basedOn`, `:improves`, `:proposes`, `:developedBy`, `:implementedIn`
- **Training & Application**: `:trainedOn`, `:finetunedOn`, `:uses`, `:appliedTo`, `:optimizedFor`
- **Research & Citation**: `:addresses`, `:publishedIn`, `:citedBy`
- **Structural**: `:partOf`, `:enabledBy`, `:resultIn`, `:relatedTo`

Mandatory Metadata:

```
:contextText "original_text_snippet"
```

A.15. Cost Analysis of the AutoQG Multi-Agent Pipeline

We further analyze the computational cost of the AutoQG multi-agent pipeline against an end-to-end LLM QA baseline. The purpose of this comparison is not to claim that AutoQG uses fewer calls, but to show that modularization substantially reduces the average context length per call while improving QA quality. Table A.10 summarizes the comparison.

Table A.10
Cost characteristics and QA quality: end-to-end prompting vs. AutoQG.

Setting	Avg. Calls / Paper	Avg. Tokens / Call	QA Quality	Rel	Ans	Distr	Flu	Reas	Evid
End-to-end LLM QA	13,421.61	13,421.61	7.31	8.0	8.9	7.4	8.6	2.4	4.5
AutoQG (multi-agent)	**23,592.53**	**4,831.52**	**8.45**	**9.6**	**9.6**	**9.1**	**8.6**	**8.4**	**9.6**

Although AutoQG incurs more agent calls per paper than the end-to-end baseline, it drastically reduces the average tokens per call (4,831.52 vs. 13,421.61). This reflects the role of multi-agent decomposition as a cost-efficiency mechanism: the pipeline distributes reasoning across specialized agents, avoids repeatedly forcing all operations into very long contexts, and maintains predictable token usage as task complexity grows.

The quality gains are substantial. AutoQG improves the overall QA quality score by +1.14 and raises reasoning depth by +6.0, which directly addresses the primary failure mode of end-to-end generation. In practice, this means that AutoQG produces higher-fidelity, evidence-grounded QA pairs while keeping token-level costs stable and manageable, rather than allowing cost to grow with increasingly complex long-context reasoning.