AutoSchemaKG: Autonomous Knowledge Graph Construction through Dynamic Schema Induction from Web-Scale Corpora

Jiaxin Bai*, **Wei Fan***, **Qi Hu***, **Qing Zong, Chunyang Li, Hong Ting Tsang**
Hongyu Luo, Yauwai Yim, Haoyu Huang, Xiao Zhou, Feng Qin, Tianshi Zheng
Xi Peng, Xin Yao, Huiwen Yang, Leijie Wu, Yi Ji
Gong Zhang, Renhai Chen, and Yangqiu Song
CSE, Hong Kong University of Science and Technology
Theory Lab, Huawei
{jbai, wfanag, qhuaf, qzong, cliei, httsangaj,
hluoay, ywyimaa, hhuangcp, tzhengad, xzhoucs, fqinac}@cse.ust.hk
{pancy.pengxi, yao.xin1, huiwen.yang, leijie.wu1, jiyi13,
nicholas.zhang, chenrenhai}@huawei.com yqsong@cse.ust.hk

Abstract

We present AutoSchemaKG, a framework for fully autonomous knowledge graph construction that eliminates the need for predefined schemas. Our system leverages large language models to simultaneously extract knowledge triples and induce comprehensive schemas directly from text, modeling both entities and events while employing conceptualization to organize instances into semantic categories. Processing over 50 million documents, we construct ATLAS (Automated Triple Linking And Schema induction), a family of knowledge graphs with 900+ million nodes and 5.9 billion edges. This approach outperforms state-of-the-art baselines on multi-hop QA tasks and enhances LLM factuality. Notably, our schema induction achieves 92% semantic alignment with human-crafted schemas with zero manual intervention, demonstrating that billion-scale knowledge graphs with dynamically induced schemas can effectively complement parametric knowledge in large language models. Code and data are fully available at `https://github.com/HKUST-KnowComp/AutoSchemaKG`.

1 Introduction

In the era of information abundance, transforming vast amounts of unstructured data into structured, machine-readable knowledge remains one of the most significant challenges in artificial intelligence. Knowledge Graphs (KGs) provide the semantic backbone for applications ranging from search engines and question answering (Wu et al., 2024; Chen et al., 2024c; Zong et al., 2024; Sun et al., 2024b) to recommendation systems (Lyu et al., 2024) and complex reasoning tasks (Li et al., 2024b). Yet current KG construction approaches remain hampered by an inherent paradox: they require predefined schemas created by domain experts, fundamentally limiting their scalability and domain coverage.

We present AutoSchemaKG, a framework that enables autonomous knowledge graph construction without predefined schemas (Ye et al., 2023), as shown in Figure 1. Our approach leverages large language models to simultaneously extract knowledge triples and dynamically induce schemas directly from text, eliminating the manual bottleneck in KG development (Zhang and Soh, 2024; Li et al., 2024a; Wang et al., 2025). Unlike entity-only approaches, we model events as first-class semantic units (Zhang et al., 2020, 2022), capturing temporal relationships, causality, and procedural knowledge. This aligns with recent work in proposition-based text decomposition (Hoyle et al., 2023; Jhamtani et al., 2024; Chen et al., 2024b) and discourse relation recognition (Chan et al., 2023, 2024), which demonstrate that event-centric representations preserve richer semantic information. Our experiments confirm this: AutoSchemaKG preserves over 90% of original passage content versus just 70% for entity information alone.

Central to our innovation is conceptualization (Wang et al., 2023c; Bai et al., 2024c; Wang et al., 2023b; He et al., 2024; Wang et al., 2024a,c), which generalizes specific entities, events, and relations into broader categories. This creates semantic bridges between disparate information, enables zero-shot cross-domain inference, and provides hierarchical organization supporting both specific and abstract reasoning (Wang et al., 2024d,b).

By deliberately investing substantial computational resources to process the Dolma 1.7 (Soldaini et al., 2024) pretraining corpus (including English

* Equal contribution.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 20557–20584

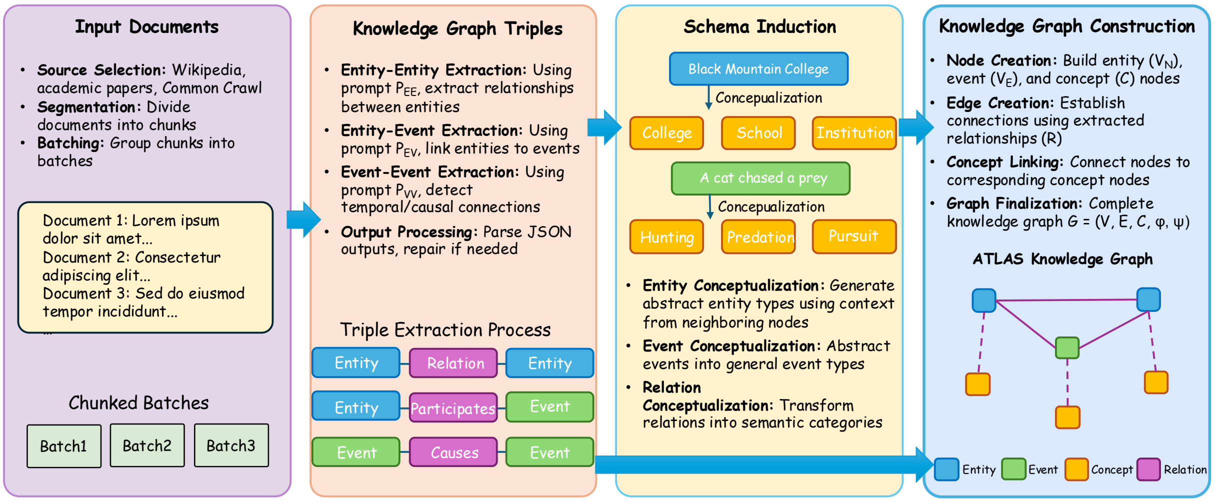

Figure 1: This figure illustrates the AutoSchemaKG pipeline for autonomous knowledge graph construction through four phases: (1) Input Processing: documents are filtered, segmented, and batched; (2) Triple Extraction: relationships between entities and events are extracted using LLM prompts; (3) Schema Induction: elements are conceptualized into abstract categories without predefined schemas; and (4) Knowledge Graph Construction: triples and schema are integrated into the ATLAS knowledge graph with entity nodes (blue), event nodes (green), concept nodes (orange), and relation edges (purple).

Wikipedia, Semantic Scholar abstracts, and 3% of Common Crawl), we construct the ATLAS family of knowledge graphs comprising over 900 million nodes connected by 5.9 billion edges. This scale is crucial: knowledge graphs must reach billions of facts to effectively complement parametric knowledge in contemporary LLMs. Through comprehensive evaluation, we demonstrate that properly structured knowledge representations offer advantages over text-based retrieval, even when drawn from the same sources as model pretraining data. Our schema induction achieves 92% semantic alignment with human-crafted schemas with zero manual intervention.

AutoSchemaKG delivers substantial performance improvements: 12-18% gains on multi-hop question answering (Trivedi et al., 2022; Yang et al., 2018; Ho et al., 2020), up to 9% enhancement in LLM factuality (Chen et al., 2023), and consistent improvements for Llama3.1 7B across diverse reasoning domains including Global Facts, History, Law, Medicine, and Social Sciences.

Our key contributions: **(1)** An entity-event-concept extraction framework creating multi-dimensional knowledge representations; **(2)** Web-scale processing yielding ATLAS—to our knowledge, both the largest automatically constructed KG and largest Graph RAG dataset available; **(3)** A retrieval augmented generation pipeline demonstrating effectiveness across diverse domains without domain-specific customization.

2 Problem Definition

We formally outline the tasks involved in automatically constructing knowledge graphs. We begin by providing a precise definition of a knowledge graph equipped with a conceptual schema.

Definition 1 (Knowledge Graph with Conceptual Schema)**.** Consider a knowledge graph denoted as $G = (V, E, C, \phi, \psi)$, where: $V = V_E \cup V_N$ represents the collection of nodes, with V_E as the set of event nodes, V_N as the set of entity nodes, and $V_E \cap V_N = \emptyset$. $E \subseteq V \times V \times R$ defines the set of edges, where R denotes relation types. Edges may connect entity-entity, entity-event, or event-event nodes. C is the set of conceptual categories. $\phi : V \to \mathcal{P}(C)$ assigns each node a subset of concepts, where $\phi(v) \subseteq C$ for every $v \in V$. $\psi : R \to \mathcal{P}(C)$ links each relation type to a subset of concepts, where $\psi(r) \subseteq C$ for every $r \in R$. $\mathcal{P}(C)$ denotes the power set of C, encompassing all possible subsets. Additional constraints: $\forall v \in V : \phi(v) \neq \emptyset$ and $\forall r \in R : \psi(r) \neq \emptyset$.

3 AutoSchemaKG Framework

Triple Extraction Our approach employs a multi-phase pipeline using Large Language Models to convert unstructured text into knowledge triples from the Dolma corpus (Soldaini et al., 2024). This pipeline extracts Entity-Entity, Entity-Event, and

	Question Answering Corpora			Pre-training Corpora		
	MuSiQue	2WikiQA	HotpotQA	ATLAS-Wiki	ATLAS-Pes2o	ATLAS-CC
# Text Chunks	11,656	6,119	9,221	9.599M	7.918M	35.040M
# Entities	108,582	48,782	95,686	70.104M	75.857M	241.061M
# Events	99,747	50,910	82,833	165.717M	92.636M	696.195M
# Concepts	37,414	19,830	32,410	8.091M	5.895M	31.070M
# Nodes	**245,743**	**119,522**	**210,929**	**243.912M**	**174.387M**	**937.256M**
# Entity-Entity Edges	91,186	40,748	78,467	0.114B	0.076B	0.414B
# Event-Entity Edges	143,254	63,680	123,527	0.265B	0.208B	1.063B
# Event-Event Edges	45,157	21,062	36,602	0.071B	0.044B	0.295B
# Conceptulization Edges	933,330	432,869	789,608	1.041B	0.821B	4.178B
# Edges	**1,212,927**	**558,359**	**1,028,204**	**1.492B**	**1.150B**	**5.958B**

Table 1: Statistics of knowledge graph construction across QA datasets (MuSiQue, 2WikiQA, HotpotQA) and LLM pre-training corpora (En-Wiki, Pes2o-Abstract, Common Crawl) for ATLAS knowledge graphs, showing counts of text chunks, nodes (entities/events), concepts, edges (Entity-Entity, Event-Entity, Event-Event), and conceptualizations. M = million, B = billion.

Event-Event relationships through three sequential stages. We preprocess texts by filtering for English language and segmenting documents that exceed token limits. The segmented texts are grouped into processing batches. Stage 1 extracts Entity-Entity relationships using a system prompt P_{EE} that instructs the LLM to detect entities and their interrelations. The output is parsed into triples (e_1, r, e_2), where $e_1, e_2 \in V_N$ are entity nodes and $r \in R$ is a relation type. Stage 2 identifies Entity-Event relationships with prompt P_{EV}, producing triples (e, r, v) or (v, r, e), where $e \in V_N$, $v \in V_E$, and $r \in R$. Stage 3 targets Event-Event relationships with prompt P_{VV}, generating triples (v_1, r, v_2), where $v_1, v_2 \in V_E$ and $r \in R$. The pipeline supports various LLMs with optimized precision settings and GPU acceleration. Extracted triples with their corresponding texts and metadata are serialized into JSON files.

Schema Induction Following triple extraction, we perform schema induction to abstract specific entities, events, and relations into generalized types. This process uses LLMs to generate conceptual phrases representing types of each graph element, aligning with our formal definition $G = (V, E, C, \phi, \psi)$. For each category (events, entities, and relations), we process elements in batches. The LLM generates at least three phrases per element that encapsulate its type or related concepts at varying abstraction levels. For entities ($e \in V_N$), we enhance abstraction by incorporating contextual information from neighboring nodes. We sample up to N_{ctx} neighbors to construct a context string that provides additional semantic cues. The schema induction pipeline processes the graph serialized from the triple extraction phase. Elements are partitioned into batches, with options for slicing for distributed computation. The generated phrases are recorded in a CSV file, mapping each node $v \in V$ and relation $r \in R$ to a subset of concepts in C via ϕ and ψ. This automated schema enhances the knowledge graph's adaptability across varied domains without requiring manual curation.

4 Construction of ATLAS Families

Corpora As shown in Table 1, the ATLAS-Wiki, ATLAS-Pes2o, and ATLAS-CC are constructed from the subsets from Dolma's subset of Wikipedia & Wikibooks, Semantic Scholar, and Dolma's CC respectively.[1] We use the full Wikipedia & Wikibooks to construct the ATLAS-Wiki, and we use the abstract part of Semantic Scholar to construct ATLAS-Pes2o, and we use the each of 3% from cc-head, cc-middle, and cc-tail to construct ATLAS-CC. According to (Soldaini et al., 2024) the head, middle, tail of CC are used to measure the distribution similarity to Wikipedia text.

5 Experiment

In the this section, we show that the AutoSchemaKG has accurate triplex extraction, can coherently induce schemas, and has very high information preservations in section 5.1.

5.1 Evaluating AutoSchemaKG

Evaluating Triple Extraction Accuracy We employ a rigorous counting-based evaluation comparing against OpenIE (Kolluru et al., 2020) and Stan-

[1] https://huggingface.co/datasets/allenai/dolma

Knowledge Graph	Triple Type	Precision	Recall	F1
ATLAS-Wiki	OpenIE 6	45.92	81.16	52.65
	Stanford OIE	78.50	93.31	82.00
	Entity-Entity	99.13	90.10	94.09
	Event-Entity	**100.0**	92.59	95.60
	Event-Event	99.60	**93.59**	**96.01**
ATLAS-Pes2o	OpenIE 6	54.08	85.50	59.89
	Stanford OIE	86.23	92.76	87.74
	Entity-Entity	97.66	89.89	93.03
	Event-Entity	**100.0**	**94.29**	**96.83**
	Event-Event	99.54	91.31	94.94
ATLAS-CC	OpenIE 6	45.17	76.73	50.20
	Stanford OIE	79.49	91.76	82.31
	Entity-Entity	95.65	84.64	88.82
	Event-Entity	**99.93**	87.92	92.72
	Event-Event	99.86	**93.20**	**96.16**

Table 2: Triple precision, recall and F1 score across datasets as well as comparisons with baseline methods. Each row displays the performance of a type of extracted triples or the performance of the entity-level triples extracted by a baseline method.

Extraction Model	Precision	Recall	F1
DeepSeek-V3	**99.59**	88.71	93.12
LLaMA-3.2-1B-Instruct	90.68	85.68	86.44
LLaMA-3.2-3B-Instruct	97.41	87.00	90.78
LLaMA-3.1-8B-Instruct	98.88	87.86	92.18
LLaMA-3.3-70B-Instruct	99.50	**89.56**	**93.73**
Qwen-2.5-7B-Instruct	96.19	85.15	89.60
Qwen-2.5-72B-Instruct	98.28	87.94	91.19

Table 3: Triple precision, recall and F1 score across LLM modules with various architectures and parameter sizes on HotpotQA dataset. Each row displays the average performance of three types of extracted triples.

ford OIE (Angeli et al., 2015). Using DeepSeek-V3 (Liu et al., 2024a) as a judge: (1) We present DeepSeek-V3 with both the original text and the triples extracted by `LLaMA-3.1-8B-Instruct`;(2) DeepSeek-V3 identifies triples that are incorrectly extracted (false positives); (3) DeepSeek-V3 lists facts present in the original text but missing from the extracted triples (false negatives); This yields precise precision, recall, and F1 scores.

Table 2 shows our approach achieves superior extraction quality across all datasets, with precision, recall, and F1 scores exceeding 90% in most cases. While traditional methods show strong recall, their precision suffers due to fixed schemas and lack of contextual reasoning. Table 3 reports average HotpotQA performance across LLMs with different architectures (DeepSeek (Liu et al., 2024a),

Dataset	Context	Model	
		LLaMA-3.1-8B	LLaMA-3.3-70B
ATLAS-Wiki	[Lower-Upper]	46.29-99.29	65.69-99.70
	OpenIE 6	87.72	89.64
	Stanford OIE	74.31	81.65
	Entity	65.08	70.96
	Event	92.69	94.82
	Event + Entity	**93.30**	**95.13**
ATLAS-Pes2o	[Lower-Upper]	62.32-98.99	75.05-99.49
	OpenIE 6	87.74	90.55
	Stanford OIE	74.21	82.74
	Entity	80.00	83.33
	Event	96.97	97.78
	Event + Entity	**97.37**	**97.98**
ATLAS-CC	[Lower-Upper]	56.08-97.29	70.25-99.10
	OpenIE 6	87.24	90.45
	Stanford OIE	73.37	82.76
	Entity	76.78	81.01
	Event	94.87	96.78
	Event + Entity	**96.28**	**96.98**

Table 4: KG performance showing bounds (no context to full passage) and results with different knowledge representations and baseline KG extraction methods. Entity, Event, and combined representations preserve most information for MCQs, approaching full-passage performance across all datasets and models.

LLaMA (Touvron et al., 2023), Qwen (Team et al., 2025)) and sizes (1B-70B), demonstrating that larger LLMs improve KG quality while architecture choice has minimal impact. We additionally report multi-judge cross-validation and a baseline-selection rationale in Appendix E.

Measuring Information Preservation We evaluate how well entity-level and event-level triples preserve passage information by testing MCQ performance, comparing against OpenIE 6 (Kolluru et al., 2020) and Stanford OIE (Angeli et al., 2015). Following (Schuhmann et al., 2025), we generate five MCQs per passage using `LLaMA-3.3-70B-Instruct` (prompts in Figure 8), sampling 200 passages and 1,000 MCQs per dataset. We test with no context (lower bound), original passage (upper bound), entity triples (Entity), event triples (Event), and combined (Event + Entity) on En-Wiki, Pes2o-Abstract, and CC.

Table 4 reveals: (1) Information is well preserved: MCQ performance with Entity, Event, or Event + Entity approaches the upper bound, far exceeding the lower bound; (2) Events outperform entities: Event or Event + Entity performance exceeds 95% in most cases, demonstrating events pre-

Context	Extraction Model	Answering Model	
		LLaMA-3.1-8B	**LLaMA-3.3-70B**
[Lower-Upper]	-	56.40-98.90	73.30-99.90
Event + Entity	DeepSeek-V3	97.90	98.80
	LLaMA-3.2-1B-Instruct	97.70	98.60
	LLaMA-3.2-3B-Instruct	94.07	95.88
	LLaMA-3.1-8B-Instruct	96.50	97.70
	LLaMA-3.3-70B-Instruct	92.56	95.18
	Qwen-2.5-7B-Instruct	95.10	96.00
	Qwen-2.5-72B-Instruct	97.99	98.19

Table 5: KG performance across different LLMs on HotpotQA dataset, showing bounds (no context to full passage) and results combining Entity and Event representations via various extraction and answering models. This combination preserves most information for MCQs across seven LLMs of diverse architectures and sizes.

Model	Entity Typing				Event Typing		Relation Typing	
	FB15kET		YAGO43kET		wikiHow		FB15kET	
	BS-R	BS-C	BS-R	BS-C	BS-R	BS-C	BS-R	BS-C
Txt2onto	85.96	75.73	88.06	23.27	93.51	31.94	78.41	13.27
LLaMA-3.1-8B	88.57	86.54	80.67	58.56	**99.18**	99.26	**88.75**	88.41
LLaMA-3.3-70B	**89.49**	**87.30**	**94.26**	**90.56**	98.97	**99.33**	88.58	**88.66**

Table 6: Schema-induction results (%) of Txt2onto, instruct-tuned version of `LLaMA-3.1-8B`, and `LLaMA-3.3-70B` across typing tasks. Bold numbers indicate the best model on the same task-dataset pair.

serve richer information; (3) Superior to baselines: Event + Entity significantly outperforms baseline methods, as traditional approaches extract incomplete triples. Table 5 shows performance across seven LLMs (different architectures and sizes) on HotpotQA maintains >95% accuracy, confirming architecture- and size-agnostic quality.

Measuring Schema Quality To demonstrate our schema induction method's capability, we apply it to entity, event, and relation typing tasks, measuring how many types our method recalls. We also compare our schema induction method with Txt2onto (Hawkins et al., 2022), which classifies unstructured text to terms in an ontology using NLP-ML. Dataset details are in Appendix C.1. Since rule-based evaluation might overlook semantic similarities, we use two semantic-level metrics: **BS-R** and **BS-C**, explained in Appendix C.2. Table 6 shows results using `LLaMA-3.1-8B-Instruct`, `LLaMA-3.3-70B-Instruct`, and an ontology-learning baseline method. Our method achieves over 80% and often 90% recall for entity, event, and relation types in most cases, which also significantly outperforms baseline methods. This is mainly due to the LLMs' powerful zero-shot generalization abilities and their

Induction Model	BS-R	BS-C
DeepSeek-V3	93.12	93.22
LLaMA-3.2-1B-Instruct	93.63	94.67
LLaMA-3.2-3B-Instruct	92.74	92.61
LLaMA-3.1-8B-Instruct	89.29	83.27
LLaMA-3.3-70B-Instruct	93.05	92.36
Qwen-2.5-7B-Instruct	93.20	93.19
Qwen-2.5-72B-Instruct	93.42	93.51

Table 7: Average results of schema induction with various LLMs across three typing tasks on HotpotQA.

freedom from predefined ontologies. Besides, we further evaluate the schema quality across seven LLMs with different architectures and parameter sizes in Table 7. The results suggest that regardless of using any architecture or parameter size of LLMs, our method can still achieve over 80% and even 90% recall and coverage for entity, event, and relation types in most cases. It also demonstrates that our automatic schema induction method can ensure the comprehensiveness of the schemas across diverse LLM architectures and parameter sizes.

5.2 Performance on Multi-hop QA Tasks

This subsection details the experimental setup for open-domain QA, focusing on multi-hop reasoning tasks where our knowledge graph's structure and schema induction are expected to excel.

Datasets We select three benchmark datasets known for their multi-hop reasoning demands, all derived from Wikipedia: MuSiQue (Trivedi et al., 2022), HotpotQA (Yang et al., 2018), and 2WikiMultihopQA (Ho et al., 2020), necessitating complex relational paths across articles. From each dataset, we randomly select one thousand questions following (Gutiérrez et al., 2024).

Baselines and Metrics We compare our knowledge graph-based RAG system against several state-of-the-art approaches. The graph-based baselines include HippoRAG (Gutiérrez et al., 2024), a framework that builds a memory graph from text using entity recognition and relation extraction; HippoRAG2 (Gutiérrez et al., 2025), an advanced iteration with enhanced graph construction; GraphRAG (Edge et al., 2024), Microsoft Research's technique combining text extraction, network analysis, and LLM prompting; LightRAG (Guo et al., 2024), a simpler alternative to GraphRAG focused on efficiency; and MiniRAG (Fan et al., 2025), an extremely simple framework tailored for Small Language Models. For MiniRAG and LightRAG we adjust its prompts so that it outputs brief answers for QA instead of generating long answers with explanations. For text-based RAG comparisons, we evaluate against BM25 + LLM using traditional retrieval with BM25 scoring; Contriever (Izacard et al., 2021), a dense retrieval-augmented system fine-tuned for QA; and RAPTOR (Sarthi et al., 2024), a hierarchical summarization system. These baselines allow us to benchmark our approach against diverse retrieval-augmented methods.

We evaluate our system using standard metrics for open-domain QA. Exact Match (EM) measures binary correctness after normalization: $\text{EM}(a, g) = \mathbf{1}[\text{norm}(a) = \text{norm}(g)]$, where a is the predicted answer, g is the gold answer, and normalization includes lowercasing and removing articles, punctuation, and whitespace. F1 Score measures token overlap between normalized answers: $\text{F1} = \frac{2 \cdot P \cdot R}{P+R}$, where $P = |a \cap g|/|a|$ and $R = |a \cap g|/|g|$ are precision and recall. The implementation details of our experiments and baselines are shown in Appendix D.

Evaluation Results The experimental results in Table 8 and Table 13 demonstrate AutoSchemaKG's effectiveness in multi-hop question answering across three benchmark datasets. With HippoRAG2 integration, the Full-KG configuration (entities, events, and concepts) outperforms traditional retrieval approaches like BM25 and Contriever by 12-18%, highlighting its strength in complex reasoning scenarios. Notably, AutoSchemaKG achieves comparable or better results using `LLaMA-3.1-8B` as graph constructor compared to the original HippoRAG2 implementation that requires `LLaMA-3.3-70B` for both construction and QA reading. We provide an additional apples-to-apples 70B constructor comparison and concept-node ablation clarifications in Appendix E.

Advantages of Events and Concepts Our case studies revealed two key benefits of event and concept nodes: **1) Event nodes provide enriched context.** As shown in Figure 9, they serve as valuable retrieval targets when critical information in triples is ambiguous or missed, helping identify relevant subgraphs containing passage nodes; **2) Concept nodes create alternative pathways.** These nodes establish connections beyond direct entities and events, addressing complex multi-hop question answering limitations. Figure 10 shows how concept nodes link knowledge across disparate subgraphs, enabling systems like HippoRAG to bridge separate subgraph influences via PageRank algorithms.

Model/Dataset	**MuSiQue**		**2Wiki**		**HotpotQA**	
Metric	**EM**	**F1**	**EM**	**F1**	**EM**	**F1**
Baseline Retrievers						
No Retriever	17.6	26.1	36.5	42.8	37.0	47.3
Contriever	24.0	31.3	38.1	41.9	51.3	62.3
BM25	20.3	28.8	47.9	51.2	52.0	63.4
LLM Embeddings						
GTE-Qwen2-7B	30.6	40.9	55.1	60.0	58.6	71.0
GritLM-7B	33.6	44.8	55.8	60.6	60.7	73.3
NV-Embed-v2 (7B)	34.7	45.7	57.5	61.5	**62.8**	75.3
Existing Graph-based RAG Methods						
RAPTOR	20.7	28.9	47.3	52.1	56.8	69.5
GraphRAG	27.3	38.5	51.4	58.6	55.2	68.6
LightRAG	20.0	29.3	38.6	44.6	33.3	44.9
MiniRAG	9.6	16.8	13.2	21.4	47.1	59.9
HippoRAG	26.2	35.1	65.0	71.8	52.6	63.5
HippoRAG2	**37.2**	**48.6**	65.0	71.0	62.7	75.5
Traditional OpenIE + HippoRAG						
OpenIE6	15.5	25.7	42.7	47.2	39.7	51.1
StanfordIE	15.6	26.3	34.3	38.7	37.3	49.1
Traditional OpenIE + HippoRAG2						
OpenIE6	28.1	42.2	48.3	54.0	51.0	70.0
StanfordIE	30.0	43.8	47.9	53.9	53.6	67.7
AutoSchemaKG (`LLama-3.1-8B-Instruct`) *+ Think-on-Graph*						
Entity-KG	14.8	26.0	36.9	44.0	41.9	55.2
Entity-Event-KG	19.4	32.8	39.0	47.1	47.7	61.2
Full-KG	20.1	31.2	40.0	47.7	48.2	60.5
AutoSchemaKG (`LLama-3.1-8B-Instruct`) *+ HippoRAG*						
Entity-KG	22.5	36.4	57.7	65.8	50.3	65.8
Entity-Event-KG	22.9	36.1	56.4	64.4	48.6	64.6
Full-KG	23.6	36.5	54.8	63.2	50.0	65.3
AutoSchemaKG (`LLama-3.1-8B-Instruct`) *+ HippoRAG2*						
Entity-KG	31.4	47.2	64.2	73.3	60.9	77.5
Entity-Event-KG	31.6	47.3	65.2	73.7	60.0	77.0
Full-KG	31.8	47.3	**65.3**	**73.9**	61.8	**78.3**

Table 8: Performance comparison of AutoSchemaKG integrated with ToG, HippoRAG and HippoRAG2 with bold indicating the highest performance per dataset.

Comparison between AutoSchemaKG and Traditional OpenIE Traditional OpenIE methods, such as OpenIE6 (Kolluru et al., 2020) and Stanford OpenIE (Angeli et al., 2015), rely on fixed schemas and focus solely on triple extraction, often producing incomplete or ambiguous triples due to the lack of context and reasoning ability. For example, given the sentence "After a hiatus of eleven years, the race was revived by the Verizon IndyCar Series in 2016," Stanford OpenIE extracts entities like "hiatus of eleven years" and "revived by Verizon IndyCar Series in 2016" without identifying "the race", resulting in an incoherent graph. Similarly, OpenIE6 identifies "the race" as an entity but fails to link it to "Desert Diamond West Valley

Corpus	Method	Acc	F1
-	-	54.08	26.79
Text Corpora			
Wikipedia	Random	52.77	25.56
	BM25	<u>56.15</u>	<u>30.43</u>
	Dense Retrieval	56.04	30.33
Pes2o-Abstract	Random	53.34	26.00
	BM25	54.60	27.95
	Dense Retrieval	55.43	29.19
Common Crawl	Random	53.31	26.45
	BM25	54.56	28.32
	Dense Retrieval	54.42	28.49
Knowledge Graph			
Freebase	Think on Graph	53.75	24.81
ATLAS-Wiki	HippoRAG2	**56.43**	**30.48**
ATLAS-Pes2o	HippoRAG2	55.30	28.12
ATLAS-CC	HippoRAG2	55.56	29.57

Table 9: Balanced accuracy (%) and F1 score (%) on FELM benchmark of `Llama-3.1-8b-Instruct` with retrieval methods. The best results are in **bold**, and the second best results are <u>underlined</u>.

Phoenix Grand Prix" or extract the revival as an event, restricting subgraph connectivity. In contrast, AutoSchemaKG resolves "the race" to "Desert Diamond West Valley Phoenix Grand Prix," extracts an event node ("the race was revived by the Verizon IndyCar Series in 2016"), and induces a concept node like "IndyCar racing" to connect related entities (e.g., "Verizon IndyCar Series"). This approach enhances graph coherence and retrieval accuracy. In KG-based Retrieval-Augmented Generation, the ambiguous triples from traditional OpenIE methods can skew node weights in algorithms like PageRank, leading to the retrieval of irrelevant passages. Moreover, the absence of concept generation in OpenIE methods limits their ability to connect disparate subgraphs, as demonstrated by the performance gap in multi-hop question answering tasks (see Table 8). Conversely, AutoSchemaKG's dynamic schema induction and inclusion of event and concept nodes address these limitations, producing more coherent and retrievable graphs.

5.3 Enhancing LLM Factuality with KGs

We evaluated our KG using the FELM benchmark (Chen et al., 2023) (847 samples across five domains, 4,425 fine-grained segments). Following FELM's protocol, we applied RAG to world knowledge, science/technology, and writing/recommendation domains, maintaining vanilla settings for math and reasoning. We compared against HippoRAG2, BM25, and dense retrieval using MiniLM (Wang et al., 2021) (implemented via ElasticSearch (Elasticsearch, 2018)). We chose MiniLM over larger embeddings due to computational constraints: dense retrieval with higher-dimensional embeddings (e.g., 4096 dimensions) across one billion nodes would require 16TB storage using 32-bit floating-point representation. All experiments used `LLaMA-3.1-8B-Instruct` with Neo4j integration and zero-shot CoT for fair comparison. Table 9 shows performance via balanced accuracy and F1 score for factual error detection. HippoRAG2 with our KG consistently outperforms baselines on Wikipedia (56.43% accuracy, 30.48% F1) and Common Crawl, with competitive results on Pes2o-Abstract. The superior Wikipedia performance likely stems from FELM samples being partially Wikipedia-sourced. Extended results appear in Appendix I.1.

5.4 General Domain Knowledge Capabilities

To assess AutoSchemaKG's ability to construct knowledge graphs across various domains, we evaluated it on MMLU (Hendrycks et al., 2021), a comprehensive benchmark for LLM reasoning. Previous research on KNN-LMs (Khandelwal et al.) suggests that retrieval-augmented generation can sometimes hinder LLMs' reasoning capabilities (Wang et al., 2023a; Geng et al., 2025). While we do not expect RAG to universally improve LLM performance, our findings demonstrate significant improvements in knowledge-intensive domains, even those covered in LLM training data. Using the same retrieval and generation settings as our FELM experiments, we classified MMLU tasks into subject categories (detailed mapping in Appendix I.2) and focused on knowledge-intensive domains including History, Law, Religion, Philosophy/Ethics, Medicine/Health, Global Facts, Social Sciences, and Logic.

As shown in Table 10, our ATLAS knowledge graphs enhanced performance across these domains on all tested corpora. Notably, each ATLAS variant demonstrated distinct strengths: ATLAS-Pes2o excelled in Religion, Medicine/Health, Global Facts, and Social Sciences, reflecting its academic paper-sourced knowledge; ATLAS-Wiki showed advantages in general knowledge areas like Religion, Philosophy/Ethics, and Global Facts; while ATLAS-CC performed best in Law and History, leveraging its broader web-sourced content. All ATLAS variants consistently outperformed both the no-retrieval

Knowledge Source	History	Law	Religion	Phil/Eth	Med/Hlth	GlbFct	SocSci	Logic	Average
None	76.59	66.86	83.04	63.55	70.38	66.72	79.74	64.35	72.10
Freebase-ToG	**78.42**	69.00	75.44	<u>65.67</u>	<u>72.65</u>	67.27	76.00	66.03	72.34
Random Baseline									
Wikipedia	76.64	66.82	79.53	59.26	70.34	66.46	77.78	59.21	70.62
Common Crawl	74.89	66.52	79.53	61.74	69.82	68.11	77.52	59.30	70.60
Pes2o-Abstract	76.24	64.16	80.70	62.01	70.62	66.59	77.16	62.39	70.82
Text Corpora + BM25									
Wikipedia	76.67	67.35	78.36	63.34	69.35	61.98	76.99	62.30	70.61
Common Crawl	76.15	66.36	80.12	60.43	69.58	64.67	76.71	63.18	70.36
Pes2o-Abstract	78.01	65.89	78.95	63.83	71.01	65.78	77.07	59.34	71.22
Text Corpora + Dense Retrieval									
Wikipedia	73.59	<u>69.60</u>	79.53	63.58	70.82	62.41	76.83	62.21	70.86
Common Crawl	74.47	68.98	79.53	60.46	69.29	64.09	75.21	61.86	69.56
Pes2o-Abstract	75.79	61.82	78.36	65.15	69.72	66.77	76.47	63.05	70.52
ATLAS + HippoRAG2									
ATLAS-Wiki	76.73	67.38	<u>84.21</u>	**66.01**	70.82	<u>68.36</u>	79.16	63.65	72.53
ATLAS-CC	<u>78.16</u>	**70.85**	83.04	65.60	71.28	63.95	78.16	65.42	72.66
ATLAS-Pes2o	77.13	68.41	81.29	65.05	**72.75**	65.67	<u>81.19</u>	62.98	<u>73.25</u>
ATLAS + Think-on-Graph									
ATLAS-Wiki	77.91	66.60	<u>84.21</u>	65.10	70.69	63.85	78.31	<u>67.08</u>	72.18
ATLAS-CC	77.07	68.18	83.63	65.24	72.03	66.87	79.72	66.59	73.07
ATLAS-Pes2o	77.52	66.95	**84.80**	63.44	71.15	**68.92**	**81.59**	**67.87**	**73.28**

Table 10: Performance comparison of `Llama-3.1-8b-Instruct` with our KG-integrated HippoRAG2 and ToG versus baseline methods across Wikipedia, Common Crawl, and Pes2o-Abstract corpora on MMLU benchmarks. Tasks are grouped by subject, with bold and underlined values indicating first and second-highest scores. Phil/Eth, Med/Hlth, GlbFct, and SocSci denote Philosophy/Ethics, Medicine/Health, Global Facts, and Social Sciences.

baseline and Freebase-ToG in these humanities and social science domains. For example, in Law, our approach achieved a 4-point improvement over the baseline, while some other retrieval methods actually decreased performance, as shown in Table 10. The retrieval method also matters for some specific tasks. For example, in Logic, ToG on our ATLAS knowledge graphs performs much better than all the other methods.

The domain-specific performance pattern aligns with intuitive expectations: knowledge graphs excel in retrieving factual relationships critical for humanities and social sciences, while showing limited benefits in mathematical and technical domains where node-relation structures are less effective for capturing procedural knowledge. The complete subject analysis, including technical domains, is available in Appendix I.2.

6 Further Related Work

KG construction transforms unstructured data into machine-readable formats. Traditional approaches using predefined schemas limit cross-domain scalability, while LLMs now enable autonomous construction through improved extraction and schema induction. Recent studies have also advanced this direction through complex query answering and reasoning over standard and eventuality KGs (Bai et al., 2022, 2023c,a,b, 2024b), as well as intention-oriented KG construction in practical domains (Yu et al., 2023; Bai et al., 2024c,a), with emerging discussion on deployable neural graph databases under privacy constraints (Hu et al., 2024). Recent advances include SAC-KG (Chen et al., 2024a), which uses LLMs as domain experts to generate specialized KGs with high precision on million-node graphs; Docs2KG (Sun et al., 2024c) for heterogeneous document processing; and KAG (Liang et al., 2024), which enhances multi-hop reasoning through KG-text mutual indexing. Related conceptualization and abstraction efforts (Wang et al., 2024a,b) also provide useful context for schema and knowledge organization, alongside recent discussions on KG reasoning trends and neural graph database challenges (Liu et al., 2024b; Bai et al., 2025; Zheng et al., 2025; Tsang et al., 2025; Huang et al., 2025). One of the earliest approaches to extracting schemas or concepts from entities, events, or textual descriptions leveraged lexico-syntactic patterns to automatically harvest novel lexico-semantic concepts from documents (Hearst, 1998), which enhanced WordNet (Miller, 1995). Agirre et al. (2000) further enhanced WordNet by extracting topically related concepts from web documents. Txt2onto (Hawkins et al., 2022) proposed a NLP-ML approach to create numerical representations of texts and use these features in a supervised learning classifier that predicts terms and concepts. Recently, due to the strong ability of LLMs to capture complex language patterns in different knowledge domains, LLMs4OL (Babaei Giglou et al., 2023) uses zero-shot prompting method with LLMs for ontology learning tasks. Schema induction automatically derives KG structure without predefined ontologies. Hofer et al. (2024) survey construction pipelines emphasizing ontology learning, while Dash et al. (2021) address canonicalization using variational autoencoders and Dognin et al. (2021) employ reinforcement learning for bidirectional text-to-graph conversion.

7 Conclusion

AutoSchemaKG eliminates the need for predefined schemas in knowledge graph construction through LLM-based triple extraction and schema induction. Processing web-scale data, we construct the ATLAS family—over 900 million nodes and 5.9 billion edges—achieving >95% extraction precision and 92% schema alignment with human experts. Our approach delivers 12-18% gains on multi-hop QA and up to 9% improvements in LLM factuality across diverse domains. This work demonstrates that billion-scale knowledge graphs with dynamically induced schemas can effectively complement parametric knowledge in large language models without expert intervention, establishing a scalable pathway for automated knowledge acquisition from unstructured text.

8 Limitations

Despite promising results, our work has several important limitations. Our approach inherits biases and limitations from the underlying LLMs used for triple extraction and schema induction, potentially affecting performance in specialized domains where these models lack expertise. While achieving high semantic alignment with human-crafted schemas, our induction method still struggles with extremely technical domains requiring expert-level conceptual organization. Despite extracting billions of facts, our knowledge graphs may contain inconsistencies, contradictions, or information gaps in sparse knowledge regions.

Acknowledgments

The authors of this paper were supported by the ITSP Platform Research Project (ITS/189/23FP) from ITC of Hong Kong, SAR, China, and the AoE (AoE/E-601/24-N), the RIF (R6021-20) and the GRF (16205322) from RGC of Hong Kong, SAR, China. We thank the computational resource support from HKUST SuperPod for large-scale experiments. We also thank the support from Huawei. We also thank the support of RGC JRFS2526-6S10.

References

Eneko Agirre, Olatz Ansa, Eduard H. Hovy, and David Martínez. 2000. Enriching very large ontologies using the www. *arXiv preprint cs/0010026*.

Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D Manning. 2015. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 344–354.

Hamed Babaei Giglou, Jennifer D'Souza, and Sören Auer. 2023. Llms4ol: Large language models for ontology learning. In *International Semantic Web Conference*, pages 408–427. Springer.

Jiaxin Bai, Xin Liu, Weiqi Wang, Chen Luo, and Yangqiu Song. 2023a. Complex query answering on eventuality knowledge graph with implicit logical constraints. *CoRR*, abs/2305.19068.

Jiaxin Bai, Chen Luo, Zheng Li, Qingyu Yin, and Yangqiu Song. 2024a. Understanding inter-session intentions via complex logical reasoning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Jiaxin Bai, Chen Luo, Zheng Li, Qingyu Yin, Bing Yin, and Yangqiu Song. 2023b. Knowledge graph reasoning over entities and numerical values. *CoRR*, abs/2306.01399.

Jiaxin Bai, Yulin Wang, Tianshi Zheng, Yang Guo, Xin Liu, and Yangqiu Song. 2024b. Advancing abductive reasoning in knowledge graphs through complex logical hypothesis generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Jiaxin Bai, Zehong Wang, Yichao Zhou, Hang Yin, Wenxuan Fei, Qi Hu, Zhenghua Deng, Jiayang Cheng, Tianshi Zheng, and Yangqiu Song. 2025.

Top ten challenges towards agentic neural graph databases. *arXiv preprint arXiv:2501.14224*.

Jiaxin Bai, Zhaobo Wang, Junfei Cheng, Dan Yu, Zerui Huang, Weiqi Wang, Xin Liu, Chen Luo, Qi He, Yanming Zhu, Bo Li, and Yangqiu Song. 2024c. Intention knowledge graph construction for user intention relation modeling. *Preprint*, arXiv:2412.11500.

Jiaxin Bai, Zihao Wang, Hongming Zhang, and Yangqiu Song. 2022. Query2Particles: Knowledge graph reasoning with particle embeddings. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2703–2714, Seattle, United States. Association for Computational Linguistics.

Jiaxin Bai, Tianshi Zheng, and Yangqiu Song. 2023c. Sequential query encoding for complex query answering on knowledge graphs. *Transactions on Machine Learning Research*.

Kurt D. Bollacker, Colin Evans, Praveen K. Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008, Vancouver, BC, Canada, June 10-12, 2008*, pages 1247–1250. ACM.

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2787–2795.

Chunkit Chan, Cheng Jiayang, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2024. Exploring the potential of ChatGPT on sentence level relations: A focus on temporal, causal, and discourse relations. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 684–721, St. Julian's, Malta. Association for Computational Linguistics.

Chunkit Chan, Xin Liu, Jiayang Cheng, Zihan Li, Yangqiu Song, Ginny Wong, and Simon See. 2023. DiscoPrompt: Path prediction prompt tuning for implicit discourse relation recognition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 35–57, Toronto, Canada. Association for Computational Linguistics.

Hanzhu Chen, Xu Shen, Qitan Lv, Jie Wang, Xiaoqi Ni, and Jieping Ye. 2024a. SAC-KG: Exploiting large language models as skilled automatic constructors for domain knowledge graph. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4345–4360, Bangkok, Thailand. Association for Computational Linguistics.

Muhao Chen, Hongming Zhang, Haoyu Wang, and Dan Roth. 2020. " what are you trying to do?" semantic typing of event processes. *arXiv preprint arXiv:2010.06724*.

Shiqi Chen, Yiran Zhao, Jinghan Zhang, I-Chun Chern, Siyang Gao, Pengfei Liu, and Junxian He. 2023. Felm: benchmarking factuality evaluation of large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Hongming Zhang, and Dong Yu. 2024b. Dense X retrieval: What retrieval granularity should we use? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15159–15177, Miami, Florida, USA. Association for Computational Linguistics.

Zhongwu Chen, Long Bai, Zixuan Li, Zhen Huang, Xiaolong Jin, and Yong Dou. 2024c. A new pipeline for knowledge graph reasoning enhanced by large language models without fine-tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1366–1381, Miami, Florida, USA. Association for Computational Linguistics.

Sarthak Dash, Gaetano Rossiello, Nandana Mihindukulasooriya, Sugato Bagchi, and Alfio Gliozzo. 2021. Open knowledge graphs canonicalization using variational autoencoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10379–10394, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pierre Dognin, Inkit Padhi, Igor Melnyk, and Payel Das. 2021. ReGen: Reinforcement learning for text and knowledge base generation using pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1084–1099, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kuicai Dong, Aixin Sun, Jung-jae Kim, and Xiaoli Li. 2023. Open information extraction via chunks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15390–15404, Singapore. Association for Computational Linguistics.

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

BV Elasticsearch. 2018. Elasticsearch. *software], version*, 6(1).

Tianyu Fan, Jingyuan Wang, Xubin Ren, and Chao Huang. 2025. Minirag: Towards extremely simple retrieval-augmented generation. *arXiv preprint arXiv:2501.06713*.

Shangyi Geng, Wenting Zhao, and Alexander M Rush. 2025. Great memory, shallow reasoning: Limits of kNN-LMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 471–482, Albuquerque, New Mexico. Association for Computational Linguistics.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2024. Lightrag: Simple and fast retrieval-augmented generation.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. 2025. From rag to memory: Non-parametric continual learning for large language models. *arXiv preprint arXiv:2502.14802*.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. Hipporag: Neurobiologically inspired long-term memory for large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Nathaniel T Hawkins, Marc Maldaver, Anna Yannakopoulos, Lindsay A Guare, and Arjun Krishnan. 2022. Systematic tissue annotations of genomics samples by modeling unstructured metadata. *Nature Communications*, 13(1):6736.

Mutian He, Tianqing Fang, Weiqi Wang, and Yangqiu Song. 2024. Acquiring and modeling abstract commonsense knowledge via conceptualization. *Artificial Intelligence*, 333:104149.

Marti A Hearst. 1998. Automated discovery of wordnet relations. *WordNet: an electronic lexical database*, 2.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Marvin Hofer, Daniel Obraczka, Alieh Saeedi, Hanna Köpcke, and Erhard Rahm. 2024. Construction of knowledge graphs: Current state and challenges. *Information*, 15(8):509.

Alexander Hoyle, Rupak Sarkar, Pranav Goel, and Philip Resnik. 2023. Natural language decompositions of implicit content enable better text representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13188–13214, Singapore. Association for Computational Linguistics.

Qi Hu, Haoran Li, Jiaxin Bai, Zehong Wang, and Yangqiu Song. 2024. Privacy-preserved neural graph databases. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Haoran Huang, Ho Ting Tsang, Jiaxin Bai, Xiaotian Peng, Guangxiang Zhang, and Yangqiu Song. 2025. Atlaskv: Augmenting llms with billion-scale knowledge graphs in 20gb vram. *arXiv preprint arXiv:2510.17934*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning.

Harsh Jhamtani, Hao Fang, Patrick Xia, Eran Levy, Jacob Andreas, and Benjamin Van Durme. 2024. Natural language decomposition and interpretation of complex utterances. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 6306–6314.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*.

Keshav Kolluru, Vaibhav Adlakha, Samarth Aggarwal, Mausam, and Soumen Chakrabarti. 2020. Openie6: Iterative grid labeling and coordination analysis for open information extraction. *arXiv (Cornell University)*.

Mahnaz Koupaee and William Yang Wang. 2018. Wikihow: A large scale text summarization dataset. *arXiv preprint arXiv:1810.09305*.

Muzhi Li, Minda Hu, Irwin King, and Ho-fung Leung. 2024a. The integration of semantic and structural knowledge in knowledge graph entity typing. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6625–6638, Mexico City, Mexico. Association for Computational Linguistics.

Yinghao Li, Haorui Wang, and Chao Zhang. 2024b. Assessing logical puzzle solving in large language models: Insights from a minesweeper case study. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 59–81, Mexico City, Mexico. Association for Computational Linguistics.

Lei Liang, Mengshu Sun, Zhengke Gui, Zhongshu Zhu, Zhouyu Jiang, Ling Zhong, Yuan Qu, Peilong Zhao, Zhongpu Bo, Jin Yang, Huaidong Xiong, Lin Yuan, Jun Xu, Zaoyang Wang, Zhiqiang Zhang, Wen Zhang, Huajun Chen, Wenguang Chen, and Jun Zhou. 2024. Kag: Boosting llms in professional domains via knowledge augmented generation. *Preprint*, arXiv:2409.13731.

Yan Liao, Jialin Hua, Liangqing Luo, Weiying Ping, Xuewen Lu, and Yuansheng Zhong. 2024. Aprcoie: An open information extraction system for chinese. *SoftwareX*, 26:101649.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Ling Liu, Zehong Wang, Jiaxin Bai, Yangqiu Song, and Hanghang Tong. 2024b. New frontiers of knowledge graph reasoning: Recent advances and future trends. In *Companion Proceedings of the ACM Web Conference 2024*, pages 1294–1297.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Hanjia Lyu, Song Jiang, Hanqing Zeng, Yinglong Xia, Qifan Wang, Si Zhang, Ren Chen, Chris Leung, Jiajie Tang, and Jiebo Luo. 2024. LLM-rec: Personalized recommendation via prompting large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 583–612, Mexico City, Mexico. Association for Computational Linguistics.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Changsung Moon, Paul Jones, and Nagiza F Samatova. 2017a. Learning entity type embeddings for knowledge graph completion. In *Proceedings of the 2017 ACM on conference on information and knowledge management*, pages 2215–2218.

Changsung Moon, Paul Jones, and Nagiza F. Samatova. 2017b. Learning entity type embeddings for knowledge graph completion. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, page 2215–2218, New York, NY, USA. Association for Computing Machinery.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. 2024. Raptor: Recursive abstractive processing for tree-organized retrieval. In *International Conference on Learning Representations (ICLR)*.

Christoph Schuhmann, Gollam Rabby, Ameya Prabhu, Tawsif Ahmed, Andreas Hochlehnert, Huu Nguyen, Nick Akinci Heidrich, Ludwig Schmidt, Robert Kaczmarczyk, Sören Auer, et al. 2025. Project alexandria: Towards freeing scientific knowledge from copyright burdens via llms. *arXiv preprint arXiv:2502.19413*.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Pete Walsh, Luke Zettlemoyer, Noah A. Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research. *CoRR*, abs/2402.00159.

Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706.

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel Ni, Heung-Yeung Shum, and Jian Guo. 2024a. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. In *The Twelfth International Conference on Learning Representations*.

Kai Sun, Yifan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. 2024b. Head-to-tail: How knowledgeable are large language models (LLMs)? A.K.A. will LLMs replace knowledge graphs? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 311–325, Mexico City, Mexico. Association for Computational Linguistics.

Qiang Sun, Yuanyi Luo, Wenxiao Zhang, Sirui Li, Jichunyang Li, Kai Niu, Xiangrui Kong, and Wei Liu. 2024c. Docs2kg: Unified knowledge graph construction from heterogeneous documents assisted by large language models. *Preprint*, arXiv:2406.02962.

Qwen Team, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report. *arXiv (Cornell University)*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*.

Ho Ting Tsang, Jiaxin Bai, Haoran Huang, Qiyuan Xiao, Tianshi Zheng, Baixuan Xu, Shuo Liu, and Yangqiu Song. 2025. Autograph-r1: End-to-end reinforcement learning for knowledge graph construction. *arXiv preprint arXiv:2510.15339*.

Shufan Wang, Yixiao Song, Andrew Drozdov, Aparna Garimella, Varun Manjunatha, and Mohit Iyyer. 2023a. kNN-LM does not improve open-ended text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15023–15037, Singapore. Association for Computational Linguistics.

Weiqi Wang, Tianqing Fang, Wenxuan Ding, Baixuan Xu, Xin Liu, Yangqiu Song, and Antoine Bosselut. 2023b. CAR: Conceptualization-augmented reasoner for zero-shot commonsense question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13520–13545, Singapore. Association for Computational Linguistics.

Weiqi Wang, Tianqing Fang, Chunyang Li, Haochen Shi, Wenxuan Ding, Baixuan Xu, Zhaowei Wang, Jiaxin Bai, Xin Liu, Cheng Jiayang, Chunkit Chan, and Yangqiu Song. 2024a. CANDLE: Iterative conceptualization and instantiation distillation from large language models for commonsense reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2351–2374, Bangkok, Thailand. Association for Computational Linguistics.

Weiqi Wang, Tianqing Fang, Haochen Shi, Baixuan Xu, Wenxuan Ding, Liyu Zhang, Wei Fan, Jiaxin Bai, Haoran Li, Xin Liu, and Yangqiu Song. 2024b. On the role of entity and event level conceptualization in generalizable reasoning: A survey of tasks, methods, applications, and future directions. *Preprint*, arXiv:2406.10885.

Weiqi Wang, Tianqing Fang, Baixuan Xu, Chun Yi Louis Bo, Yangqiu Song, and Lei Chen. 2023c. CAT: A contextualized conceptualization and instantiation framework for commonsense reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13111–13140, Toronto, Canada. Association for Computational Linguistics.

Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. 2021. MiniLMv2: Multi-head self-attention relation distillation for compressing pretrained transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2140–2151, Online. Association for Computational Linguistics.

Zehong Wang, Sidney Liu, Zheyuan Zhang, Tianyi Ma, Chuxu Zhang, and Yanfang Ye. 2025. Can LLMs convert graphs to text-attributed graphs? In *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)*, pages 1412–1432, Albuquerque, New Mexico. Association for Computational Linguistics.

Zhaowei Wang, Wei Fan, Qing Zong, Hongming Zhang, Sehyun Choi, Tianqing Fang, Xin Liu, Yangqiu Song, Ginny Wong, and Simon See. 2024c. AbsInstruct: Eliciting abstraction ability from LLMs through explanation tuning with plausibility estimation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 973–994, Bangkok, Thailand. Association for Computational Linguistics.

Zhaowei Wang, Haochen Shi, Weiqi Wang, Tianqing Fang, Hongming Zhang, Sehyun Choi, Xin Liu, and Yangqiu Song. 2024d. AbsPyramid: Benchmarking the abstraction ability of language models with a unified entailment graph. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3991–4010, Mexico City, Mexico. Association for Computational Linguistics.

Yike Wu, Yi Huang, Nan Hu, Yuncheng Hua, Guilin Qi, Jiaoyan Chen, and Jeff Z. Pan. 2024. CoTKR: Chain-of-thought enhanced knowledge rewriting for complex knowledge graph question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3501–3520, Miami, Florida, USA. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Hongbin Ye, Honghao Gui, Xin Xu, Xi Chen, Huajun Chen, and Ningyu Zhang. 2023. Schema-adaptable knowledge graph construction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6408–6431, Singapore. Association for Computational Linguistics.

Changlong Yu, Weiqi Wang, Xin Liu, Jiaxin Bai, Yangqiu Song, Zheng Li, Yifan Gao, Tianyu Cao, and Bing Yin. 2023. Folkscope: Intention knowledge graph construction for e-commerce commonsense discovery. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1173–1191. Association for Computational Linguistics.

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

Hongming Zhang, Xin Liu, Haojie Pan, Haowen Ke, Jiefu Ou, Tianqing Fang, and Yangqiu Song. 2022. ASER: towards large-scale commonsense knowledge acquisition via higher-order selectional preference over eventualities. *Artificial Intelligence*, 309:103740.

Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. 2020. Aser: A large-scale eventuality knowledge graph. In *Proceedings of the web conference 2020*, pages 201–211.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Tianshi Zheng, Zhenghua Deng, Ho Ting Tsang, Weiqi Wang, Jiaxin Bai, Zehong Wang, and Yangqiu Song. 2025. From automation to autonomy: A survey on large language models in scientific discovery. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

Chang Zong, Yuchen Yan, Weiming Lu, Jian Shao, Yongfeng Huang, Heng Chang, and Yueting Zhuang. 2024. Triad: A framework leveraging a multi-role LLM-based agent to solve knowledge base question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1698–1710, Miami, Florida, USA. Association for Computational Linguistics.

A Prompts for Triple Extractions

The prompts used for extracting entity-entity, entity-event, and event-event triples are given in Figure 2, Figure 3, and Figure 4 respectively.

B Implementation Details of Knowledge Graph Construction Framework

In this section, we elaborate on the process of fully automating knowledge graph construction. Given a collection of n documents $D = \{D_1, D_2, D_3, \ldots, D_n\}$, our method systematically builds the graph.

B.1 Triple Extraction

Our approach to triple extraction employs a multi-phase pipeline that utilizes the generative power of Large Language Models (LLMs) to convert unstructured text into structured knowledge triples, drawing from the Dolma corpus (Soldaini et al., 2024). This pipeline systematically extracts three categories of relationships—Entity-Entity, Entity-Event, and Event-Event—forming the foundation of a comprehensive knowledge graph. Designed for scalability and resilience, the method incorporates batch processing, text segmentation, and robust output parsing to efficiently process large-scale datasets.

Entity Relationship Extraction

Given a passage, summarize all the important entities and the relations between them in a concise manner. Relations should briefly capture the connections between entities, without repeating information from the head and tail entities. The entities should be as specific as possible. Exclude pronouns from being considered as entities. The output should strictly adhere to the following JSON format:

```
[
  {
    "Head": "{a noun}",
    "Relation": "{a verb}",
    "Tail": "{a noun}",
  }
  ...
]
```

Here is the passage:

Figure 2: The figure demonstrates the prompts we use to generate the text triples describing relations between entities.

The extraction unfolds across three sequential stages, each tailored to a specific relationship type, leveraging a single LLM guided by distinct prompts to produce structured outputs in JSON format. To manage the constraints of LLM input capacity, denoted as L_{max} tokens, we preprocess the text corpus to ensure compatibility, segmenting documents as needed and organizing them into batches for efficient processing. This section outlines the preprocessing strategy, the staged extraction process, and key implementation details.

B.1.1 Text Preprocessing and Data Organization

Given a corpus $D = \{D_1, D_2, \ldots, D_n\}$ of n documents, we first filter the dataset to include only English-language texts, identified through metadata or assumed if unspecified, to match the linguistic capabilities of our LLMs. To adhere to the token limit L_{max}, we account for an instructional prompt length, L_{inst} derived from empirical observations. The maximum token length per text segment, C_{max}, is calculated as: $C_{max} = (L_{max} - L_{inst})$

Event and Entity Triple Extraction

Please analyze and summarize the participation relations between the events and entities in the given paragraph. Each event is a single independent sentence. Additionally, identify all the entities that participated in the events. Do not use ellipses. Please strictly output in the following JSON format:

```
[
  {
    "Event": "{a simple sentence describing an event}",
    "Entity": ["{entity 1}", "{entity 2}", "..."]
  }
  ...
]
```

Figure 3: The figure demonstrates the prompts we use to generate the text triples describing relations between entities and events.

Event Relationship Extraction

Please analyze and summarize the relationships between the events in the paragraph. Each event is a single independent sentence. Identify temporal and causal relationships between the events using the following types: before, after, at the same time, because, and as a result. Each extracted triple should be specific, meaningful, and able to stand alone. Do not use ellipses. The output should strictly adhere to the following JSON format:

```
[
  {
    "Head": "{a simple sentence describing the event 1}",
    "Relation": "{temporal or causality relation between the events}",
    "Tail": "{a simple sentence describing the event 2}"
  }
  ...
]
```

Figure 4: The figure demonstrates the prompts we use to generate the text triples describing relations between events.

Documents exceeding C_{max} are divided into smaller chunks, each tagged with a unique identifier and metadata to maintain traceability. This segmentation ensures that inputs remain within the LLM's token capacity, preserving contextual integrity without truncation.

The preprocessed text is then grouped into batches of size B, utilizing a custom data management framework that integrates with standard dataset loading tools. Tokenization is applied to each batch, adjusting for padding and truncation to produce consistent input representations suitable for LLM processing.

B.1.2 Stage 1: Extraction of Entity-Entity Relationships

In the initial stage, as shown in Figure 2, we extract Entity-Entity relationships, identifying connections between named entities such as individuals, organizations, or locations. For each batch, we prepend a system prompt, P_{EE}, which instructs the LLM to detect entities and their interrelations, followed by the segmented text. The LLM generates a response in JSON format, which is subsequently decoded and parsed. The parsing process isolates the structured content by locating the model's answer start token, T_{start}, repairs any malformed JSON, and extracts a list of dictionaries. Each dictionary represents a triple (e_1, r, e_2), where $e_1, e_2 \in V_N$ are entity nodes and $r \in R$ is a relation type. If parsing encounters errors, an empty list is returned to ensure pipeline continuity .

B.1.3 Stage 2: Extraction of Entity-Event Relationships

The second stage focuses on Entity-Event relationships, as shown in Figure 3, linking entities to specific occurrences or events. Using the original text segments from Stage 1, we apply a new prompt, P_{EV}, directing the LLM to identify events and their associated entities. The generation and parsing steps mirror those of Stage 1, producing triples of the form (e, r, v) or (v, r, e), where $e \in V_N$, $v \in V_E$ (event nodes), and $r \in R$. This bidirectional extraction captures both entities involved in events and events tied to entities, enhancing the graph's relational depth.

B.1.4 Stage 3: Extraction of Event-Event Relationships

The third stage targets Event-Event relationships, as shown in Figure 4,, detecting causal, temporal, or logical connections between events. A specialized prompt, P_{VV}, is applied to the text segments, prompting the LLM to generate triples of the form (v_1, r, v_2), where $v_1, v_2 \in V_E$ and $r \in R$. The parsing process follows the same methodology, repairing JSON outputs as needed. To accommodate potentially intricate event descriptions, we extend the generation limit to $L_{ext} = \alpha \cdot L_{max}$, where $\alpha > 1$ is a scaling factor, ensuring comprehensive capture of event interactions.

B.1.5 Implementation Considerations

The pipeline supports a variety of LLMs, including models from Google, Meta, Mistral, Microsoft, and others, configured with optimized precision settings (e.g., bfloat16 or float16) and enhanced with acceleration techniques where applicable. Deployment occurs on GPUs, with input-output formatting governed by model-specific chat templates, T_{chat}, to ensure compatibility. The extracted triples, along with their corresponding texts and metadata, are serialized into JSON files per batch, enabling subsequent schema induction and evaluation.

This multi-stage pipeline achieves thorough triple extraction by addressing each relationship type systematically, harnessing the LLM's generative capabilities within a scalable and fault-tolerant framework. The use of variables such as L_{max}, C_{max}, and B ensures flexibility across different models and datasets, reinforcing the methodology's adaptability and generalizability.

B.2 Schema Induction

Following the extraction of knowledge triples, our methodology advances to schema induction, a critical step that abstracts specific entities, events, and relations into generalized types to form a coherent and adaptable schema for the knowledge graph. This process leverages the contextual understanding of Large Language Models (LLMs) to generate conceptual phrases that represent the types or related concepts of each graph element, enabling the graph to scale across diverse domains without manual schema design. The induced schema aligns with the formal definition of a knowledge graph $G = (V, E, C, \phi, \psi)$, where C denotes the set of concepts, and ϕ and ψ map nodes and relations to subsets of C, respectively.

Our schema induction pipeline processes the triples extracted from the Dolma corpus (Soldaini et al., 2024), organizing them into batches and employing a generative approach to derive abstract representations. The process targets three components—events (V_E), entities (V_N), and relations (R)—producing a set of conceptual phrases for each, which collectively form the concept set C. This section outlines the abstraction methodology, the role of context in entity conceptualization, and key implementation details.

B.2.1 Abstraction Methodology

The schema induction begins by categorizing the nodes and edges of the knowledge graph G into events, entities, and relations. For each category, we process the elements in batches of size B_s to optimize computational efficiency and scalability. The LLM is prompted with tailored instructions to generate a list of phrases, each containing one to two words, that abstractly represent the input element. These phrases must satisfy several criteria: they should encapsulate the element's type or related concepts, vary in abstraction level, and avoid repetition or inclusion of the original input term. For each element, a minimum of three phrases is targeted, though more may be generated depending on the LLM's output.

For events ($v \in V_E$), the prompt directs the LLM to identify abstract event types or related notions. For example, an event such as "Sam playing with his dog" might yield phrases like "playing," "bonding," and "relaxing event," reflecting different levels of generality. For entities ($e \in V_N$), the prompt similarly elicits abstract entity types, augmented by contextual information derived from the graph structure, as detailed below. Relations ($r \in R$) are abstracted into phrases that capture their semantic essence, such as transforming "participated in" into "engage in," "attend," and "involve in." The

Abstract Event Phrase Generation

I will give you an EVENT. You need to give several phrases containing 1-2 words for the ABSTRACT EVENT of this EVENT. You must return your answer in the following format: phrases1, phrases2, phrases3,... You can't return anything other than answers. These abstract event words should fulfill the following requirements:
1. The ABSTRACT EVENT phrases can well represent the EVENT, and it could be the type of the EVENT or the related concepts of the EVENT.
2. Strictly follow the provided format, do not add extra characters or words.
3. Write at least 3 or more phrases at different abstract level if possible.
4. Do not repeat the same word and the input in the answer.
5. Stop immediately if you can't think of any more phrases, and no explanation is needed.

Examples:
EVENT: A man retreats to mountains and forests Your answer: retreat, relaxation, escape, nature, solitude
EVENT: A cat chased a prey into its shelter Your answer: hunting, escape, predation, hidding, stalking
EVENT: Sam playing with his dog Your answer: relaxing event, petting, playing, bonding, friendship
EVENT: [EVENT] Your answer:

Figure 5: This figure shows the prompt used for generating the concepts for an event.

LLM generates these outputs in a structured format, which we parse into lists of phrases, forming the mappings $\phi(v)$, $\phi(e)$, and $\psi(r)$ to the concept set C.

The abstraction process operates in batches, processing B_s elements simultaneously. The input prompts are tokenized to a maximum length L_{tok}, and the LLM generates responses under controlled parameters (e.g., temperature τ and top-p sampling with probability p) to balance creativity and coherence. The resulting phrases are stored alongside their corresponding elements, ensuring traceability and enabling subsequent analysis.

B.2.2 Contextual Enhancement for Entities

As shown in Figure 6, to enhance the accuracy of entity abstraction, we incorporate contextual information extracted from the knowledge graph. For each entity $e \in V_N$, we examine its neighboring nodes—predecessors and successors—along with their associated relations. A subset of these neighbors, limited to N_{ctx} (e.g., one predecessor and one successor), is randomly sampled to construct a context string. This string concatenates the neighbor's identity and relation (e.g., "neighbor1 relation1, relation2 neighbor2"), providing the LLM with additional semantic cues. For instance, an entity "Black Mountain College" with context "started by John Andrew Rice" might yield phrases like "college," "school," and "liberal arts college." This contextual enrichment ensures that the abstracted types are grounded in the entity's role within the graph, improving the schema's relevance and specificity.

Events and relations, as shown in Figure 5 and Figure 7, by contrast, rely solely on their textual descriptions without additional context, as their abstraction focuses on inherent semantics rather than graph connectivity. This distinction reflects the differing roles of nodes and edges in the knowledge graph structure.

B.2.3 Implementation Details

The schema induction pipeline processes a graph G serialized from the triple extraction phase, typically stored in a binary format and loaded into memory. The elements are partitioned into batches, with the option to apply slicing (dividing the workload into S_{total} slices and processing the S_{slice}-th portion) for distributed computation. If a sample size N_{sample} is specified, a random subset of batches is selected to reduce processing time during experimentation.

Abstract Entity Phrase Generation

I will give you an ENTITY. You need to give several phrases containing 1-2 words for the ABSTRACT ENTITY of this ENTITY. You must return your answer in the following format: phrases1, phrases2, phrases3,... You can't return anything other than answers. These abstract intention words should fulfill the following requirements:
1. The ABSTRACT ENTITY phrases can well represent the ENTITY, and it could be the type of the ENTITY or the related concepts of the ENTITY.
2. Strictly follow the provided format, do not add extra characters or words.
3. Write at least 3 or more phrases at different abstract level if possible.
4. Do not repeat the same word and the input in the answer.
5. Stop immediately if you can't think of any more phrases, and no explanation is needed.

Examples:
ENTITY: Soul CONTEXT: premiered BFI London Film Festival, became highest-grossing Pixar release Your answer: movie, film
ENTITY: Thinkpad X60 CONTEXT: Richard Stallman announced he is using Trisquel on a Thinkpad X60 Your answer: Thinkpad, laptop, machine, device, hardware, computer, brand
ENTITY: Harry Callahan CONTEXT: bluffs another robber, tortures Scorpio Your answer: person, American, character, police officer, detective
ENTITY: Black Mountain College CONTEXT: was started by John Andrew Rice, attracted faculty Your answer: college, university, school, liberal arts college
ENTITY: 1st April CONTEXT: Utkal Dibas celebrates Your answer: date, day, time, festival
ENTITY: [ENTITY] CONTEXT: [CONTEXT] Your answer:

Figure 6: This figure shows the conceptualization prompts for entities enhanced with context.

The LLM, configured with a precision setting (e.g., float16) and optimized with acceleration techniques, operates on a GPU to handle the batched inference efficiently. Prompts are formatted using a model-specific chat template, T_{chat}, ensuring compatibility with the LLM's input-output conventions. The generated phrases are written to a CSV file, with each row recording the original element, its abstracted phrases, and its type (event, entity, or relation). Post-processing aggregates these phrases to compute the unique concepts in C, providing statistics on the schema's coverage, such as the number of distinct event types, entity types, and relation types.

This approach yields a flexible and automated schema, mapping each node $v \in V$ and relation $r \in R$ to a subset of concepts in C via ϕ and ψ. By abstracting specific instances into general types, the

Abstract Relation Phrase Generation

I will give you a RELATION. You need to give several phrases containing 1-2 words for the ABSTRACT RELATION of this RELATION. You must return your answer in the following format: phrases1, phrases2, phrases3,... You can't return anything other than answers. These abstract intention words should fulfill the following requirements:
1. The ABSTRACT RELATION phrases can well represent the RELATION, and it could be the type of the RELATION or the simplest concepts of the RELATION.
2. Strictly follow the provided format, do not add extra characters or words.
3. Write at least 3 or more phrases at different abstract level if possible.
4. Do not repeat the same word and the input in the answer.
5. Stop immediately if you can't think of any more phrases, and no explanation is needed.
Examples:
RELATION: participated in Your answer: become part of, attend, take part in, engage in, involve in
RELATION: be included in Your answer: join, be a part of, be a member of, be a component of
RELATION: [RELATION] Your answer:

Figure 7: This figure shows the conceptualization prompts for relations enhanced with context.

induced schema enhances the knowledge graph's adaptability, supporting downstream applications across varied domains without requiring manual curation.

C Experiment Settings of Schema Accuracy

C.1 Datasets

Entity Typing. We conduct experiments on the typed entities of two real-world knowledge graphs, FB15kET (Bordes et al., 2013) and YAGO43kET (Moon et al., 2017a), which are the subsets of Freebase (Bollacker et al., 2008) and YAGO (Suchanek et al., 2007), respectively. The types of entities are collected from (Moon et al., 2017b). There are 3,584 and 45,182 entity types in FB15kET and YAGO43kET, respectively. We utilize the entities in the testing sets of these two datasets with their types as ground truths to validate the entity induction performance of our schema induction method.

Event Typing. We conduct experiments on the typed events of wikiHow (Koupaee and Wang, 2018), which is an online community contains a collection of professionally edited how-to guideline articles. The types of events are collected by P2GT (Chen et al., 2020). There are 625 event types among 12,795 events. We utilize the events in the testing set of wikiHow with their types as ground truths to validate the event induction performance of our schema induction method.

Relation Typing. We make use of the domain segments separated by "/" in the relations of FB15kET (Bordes et al., 2013) to extract the relation types. These domain segments serve as ground truth types, with the last domain component functioning as the relation itself. There are 607 relation types among 1,345 relations in FB15kET. We utilize the relations in the testing set of FB15kET with their types as ground truths to validate the relation induction performance of our schema induction method.

C.2 Metrics

We employ **BertScore-Recall** and **BertScore-Coverage** as the evaluation metrics, which are denoted as **BS-R** and **BS-C** respectively. They are used to calculate how many types in each instance or entire testing set are recalled by our schema induction method. The BertScore (Zhang et al., 2019), which is denoted as **BS**, between each pair of type and induced schema are calculated as follows:

$$\text{BertRecall} = \frac{1}{|t|}\sum_{t_i \in t} \max_{\hat{t}_j \in \hat{t}} \mathbf{x}_{t_i}^{\top}\mathbf{x}_{\hat{t}_j}, \quad (1)$$

$$\text{BertPrec} = \frac{1}{|\hat{t}|}\sum_{\hat{t}_i \in \hat{t}} \max_{t_j \in t} \mathbf{x}_{t_j}^{\top}\mathbf{x}_{\hat{t}_i}, \quad (2)$$

$$\text{BS}(t,\hat{t}) = 2\frac{\text{BertRecall}\cdot\text{BertPrec}}{\text{BertRecall}+\text{BertPrec}}, \quad (3)$$

where t and $\hat{t}$ represents the tokens of a ground truth type and induced schema, respectively. The embedding vector of each token t_i or $\hat{t}_j$ of a type t or induced schema $\hat{t}$ is denoted as $\mathbf{x}_{t_i}$ and $\mathbf{x}_{\hat{t}_j}$, which are obtained with RoBERTa (Liu et al., 2019). Then the BS-R and BS-C can be calculated as follows:

$$\text{BS-R}(\mathcal{T},\hat{\mathcal{T}}) = \frac{1}{|\hat{\mathcal{T}}|}\sum_{\hat{t}\in\hat{\mathcal{T}}} \max_{t\in\mathcal{T}} \text{BS}(t,\hat{t}), \quad (4)$$

$$\text{BS-C}(\mathcal{S}_t,\mathcal{S}_{\hat{t}}) = \frac{1}{|\mathcal{S}_{\hat{t}}|}\sum_{\hat{t}\in\mathcal{S}_{\hat{t}}} \max_{t\in\mathcal{S}_t} \text{BS}(t,\hat{t}), \quad (5)$$

where $\mathcal{T}$ and $\hat{\mathcal{T}}$ represent a set of ground truth types and induced schemas in each testing instance, respectively. Similarly, $\mathcal{S}_t$ and $\mathcal{S}_{\hat{t}}$ denote the set of ground truth types and induced schemas across the entire testing set, respectively.

D Experiment Details

Think on Graph Settings We implement Think on Graph (ToG) (Sun et al., 2024a) using a knowledge graph derived from our corpus. Nodes represent extracted entities and concepts, with edges showing semantic relations. We use `multi-qa-MiniLM-L6-dot-v1` to compute embeddings for all graph elements, indexed with FAISS. `LLama-3.3-70B-Instruct` performs entity recognition, path scoring, reasoning, and answer generation. The workflow extracts query entities, retrieves starting nodes, explores graph paths through depth-limited search, prunes irrelevant paths, and generates answers based on retrieved paths. Appendix Algorithm 1 provides details.

HippoRAG 1&2 Settings In our implementation of HippoRAG (Gutiérrez et al., 2024), we extend the original framework to operate on a customized graph. Initially presented in the foundational paper, we employ Named Entity Recognition (NER) to build a personalized dictionary for PageRank execution. Regarding HippoRAG2 (Gutiérrez et al., 2025), we select the top 30 edges (musique dataset 50 edges) for LLM filtering, incorporating a weight adjustment factor of 0.9. Considering the capability of our graph to effectively locating subgraphs, combined with various graph configurations (entity, event, concept) resulting in graphs of differing densities, we set the damping factor to 0.9 to concentrate on propagation within the local subgraph. For further implementation details, please refer to Algorithm 2.

Implementation Details The knowledge graph is constructed from the corresponding context corpora for each dataset folowing (Gutiérrez et al., 2024) using the framework of AutoSchemaKG with $L_{max} = 1024$ and $B = 16$, and the schema induction pipeline (Section 3) with $B_s = 5$. We employ Meta's `LLaMA-3.1-8B-Instruct` to construct the graphs, optimized with bfloat16 precision and Flash Attention 2. The graph is stored in NetworkX for retrieval, with subgraphs fed into `LLaMA-3.3-70B-Instruct` for answer generation.

Computational Cost We constructed our knowledge graphs using 80GB GPUs with 1,513 TFLOPS of FP16 compute, running `Llama-3-8B-instruct` with Flash Attention. The computational demands were substantial: 14,300 GPU hours for En-Wiki (243.9M nodes, 1.49B edges), 11,800 GPU hours for Pes2o-Abstract (174.4M nodes, 1.15B edges), and 52,300 GPU hours for Common Crawl (937.3M nodes, 5.96B edges). Processing 1024-token chunks in batches, we invested approximately 78,400 GPU hours total to extract billions of semantic relationships.

Model	MuSiQue		2Wiki		HotpotQA	
	EM	F1	EM	F1	EM	F1
HippoRAG-2 (baseline)	37.2	48.6	65.0	71.0	62.7	75.5
HippoRAG-2 (AutoSchemaKG, 70B)	36.3	**50.5**	62.4	70.6	61.4	**78.7**

Table 11: Apples-to-apples comparison using `LLaMA-3.3-70B` as graph constructor. AutoSchemaKG improves F1 on MuSiQue and HotpotQA, while baseline remains slightly better on 2Wiki.

E Additional Technical Analyses

Multi-judge validation for triple extraction quality To reduce potential bias from using a single LLM judge, we additionally cross-validated extraction quality with Kimi-K2-Instruct and GPT-4o. Across ATLAS-Wiki, ATLAS-Pes2o, and ATLAS-CC, the relative ranking is consistent with our primary evaluation: AutoSchemaKG remains better than OpenIE 6 and Stanford OIE in overall extraction quality when considering entity-entity, event-entity, and event-event triples jointly.

Why we keep classic OpenIE baselines We include OpenIE 6 (Kolluru et al., 2020) and Stanford OIE (Angeli et al., 2015) as representative systems of the traditional SPO-triple paradigm that our work aims to move beyond. While newer systems improve extraction quality (Dong et al., 2023; Liao et al., 2024), they still mainly operate within the same disconnected triple-centric structure. Our comparison therefore focuses on this structural difference (entity-event-concept graph with induced schema vs. traditional triple-only graph), rather than a narrow incremental gain on one extraction module.

Apples-to-apples 70B constructor comparison We also re-ran AutoSchemaKG graph construction with `LLaMA-3.3-70B` (instead of 8B) and evaluated with the same QA setup as HippoRAG-2.

Clarifying the contribution of concept nodes To make the concept-node contribution explicit, Table 12 reports the ablation with the same reader (`HippoRAG2`) and constructor (`LLaMA-3.1-8B`). The progression from Entity-KG to Entity-Event-KG to Full-KG isolates the effect of adding concept nodes.

Graph	MuSiQue		2Wiki		HotpotQA	
	EM	F1	EM	F1	EM	F1
Entity-KG	31.4	47.2	64.2	73.3	60.9	77.5
Entity-Event-KG	31.6	47.3	65.2	73.7	60.0	77.0
Full-KG (with Concepts)	**31.8**	**47.3**	**65.3**	**73.9**	**61.8**	**78.3**

Table 12: Concept-node ablation under AutoSchemaKG + HippoRAG2. Full-KG (entities + events + concepts) yields the strongest overall multi-hop QA results.

Schema conformation beyond component typing For a schema-free setting, we view holistic schema conformation at the triple level through downstream functional validity: if extracted triples are not semantically coherent as a whole, they cannot support multi-hop reasoning. Therefore, we interpret the strong QA and retrieval results as complementary evidence to the component-level typing metrics (Tables 6 and 7).

Clarification on LLM dependency and deployment cost AutoSchemaKG intentionally uses strong LLMs as a construction engine, while the core contribution is the graph-construction framework (entity-event-concept extraction plus dynamic schema induction) that can improve as base models improve. We explicitly acknowledge the compute and storage overhead as a current trade-off for web-scale construction, and view compression/distillation and domain-specific adaptation (fine-tuning or retrieval grounding) as practical paths to improve deployment in specialized or resource-constrained scenarios.

F Case Study Examples

Figures 9 and 10 demonstrate specific cases where events and concepts are crucial for effective knowledge graph utilization in retrieval-augmented generation. Figure 9 illustrates how event nodes provide essential contextual information that entity-only representations miss, while Figure 10 showcases how concept nodes establish semantic bridges across otherwise disconnected subgraphs, enabling more comprehensive reasoning for complex questions.

G Algorithm for RAG

We include all the algorithms used in our RAG evaluation on various graphs constructed by AutoSchemaKG. Algorithm 1 presents the Think-on-Graph reasoning method that leverages our knowledge graphs for multi-hop question answering. For adapting our entity-event-concept graphs at different scales, we implemented two variants of HippoRAG2: Algorithm 2 for smaller, more focused graph traversal, and Algorithm 3 for large-scale graph exploration with optimized memory management. These adaptations enable efficient navigation of the rich semantic structures in our ATLAS knowledge graphs.

H The Recall Metrics in Opendomain QA Tasks

We also use Retrieval Quality metrics at $k \in \{2, 5\}$: PR@$k = |D_k \cap S|/|S|$ where PR@k (Partial Recall) measures the fraction supporting document is in top-k, D_k is the set of top-k retrieved documents, and S is the set of supporting documents. For multi-hop QA datasets (HotpotQA, 2WikiMultihopQA, MuSiQue), these retrieval metrics are crucial as they measure how effectively our system retrieves the evidence needed for multi-step reasoning.

Questions in datasets like 2WikiMultihopQA (Ho et al., 2020) and HotpotQA (Yang et al., 2018) tend to be more entity-centric, with relationships and entities more explicitly represented, which aids retrievers in easily locating relevant subgraphs. In contrast, MuSiQue (Trivedi et al., 2022), due to its questions' increased complexity in both description and multi-hop nature, poses greater challenges for retrieval. Additionally, differences in graph construction cause the retrievers to perform differently across datasets.

I Details and Full Results on General Benchmarks

I.1 Implementation and Evaluation Details on FELM

For the evaluation metrics, we follow the original paper (Chen et al., 2023) and use balanced accuracy and F1 score to evaluate the factuality checking capability. For the classification of segments in an instance, we ask the model to generate the ID of false segments, and then get the true positive (TP), false positive (FP), true negative (TN) and false negative (FN) results. The balanced accuracy is calculated as:

$$\text{Balanced Accuracy} = \frac{TP}{TP + FN} + \frac{TN}{TN + FP} \quad (6)$$

Since we use F1 score to evaluate the factual error detection capability, we calculate the F1 score as:

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (7)$$

Algorithm 1 Think on Graph (ToG) (Sun et al., 2024a) for Question Answering

Require: Knowledge Graph G, Query q, Top-N parameter, Maximum depth D_{max}
Ensure: Answer to query q
Extract entities from query q using NER
Retrieve top-k initial nodes from G based on entity similarity
Let $P \leftarrow$ set of paths, each containing a single initial node
$D \leftarrow 0$ ▷ Current search depth
while $D \leq D_{max}$ **do**
  $P \leftarrow \texttt{Search}(q, P, G)$ ▷ Expand paths by one hop
  $P \leftarrow \texttt{Prune}(q, P, N)$ ▷ Keep top-N most relevant paths
  if $\texttt{Reasoning}(q, P)$ determines paths sufficient **then**
    return $\texttt{Generate}(q, P)$ ▷ Generate answer using paths
  end if
  $D \leftarrow D + 1$
end while
return $\texttt{Generate}(q, P)$ ▷ Generate answer using best available paths
procedure SEARCH(q, P, G)
  $P_{new} \leftarrow \emptyset$
  for each path $p \in P$ **do**
    $e_{tail} \leftarrow$ last entity in path p
    $S \leftarrow$ successors of e_{tail} in G not already in p
    $R \leftarrow$ predecessors of e_{tail} in G not already in p
    if $S = \emptyset$ and $R = \emptyset$ **then**
      $P_{new} \leftarrow P_{new} \cup \{p\}$ ▷ Keep dead-end paths
    else
      for each node $n \in S$ **do**
        $r \leftarrow$ relation from e_{tail} to n in G
        $P_{new} \leftarrow P_{new} \cup \{p + [r, n]\}$ ▷ Extend path forward
      end for
      for each node $n \in R$ **do**
        $r \leftarrow$ relation from n to e_{tail} in G
        $P_{new} \leftarrow P_{new} \cup \{p + [r, n]\}$ ▷ Extend path backward
      end for
    end if
  end for
  return P_{new}
end procedure
procedure PRUNE(q, P, N)
  Score each path in P using LLM relevance assessment (1-5 scale)
  Sort paths by decreasing score
  return top-N highest scoring paths
end procedure
procedure REASONING(q, P)
  Extract triples from paths in P
  Ask LLM if triples are sufficient to answer q (Yes/No)
  return True if answer is "Yes", False otherwise
end procedure
procedure GENERATE(q, P)
  Extract triples from paths in P
  Prompt LLM with triples and query q to generate answer
  return generated answer
end procedure

Algorithm 2 HippoRAG2 (Gutiérrez et al., 2025)

General algorithm follows the original implementation, while we modify the initialization of graph and embeddings.

```
function INIT(graph_type)
    if graph_type is entity then
        graph, embedding ← Graph(entity), Embeddings(entity)
    else if graph_type is entity+event then
        graph, embedding ← Graph(entity, event), Embeddings(entity, event)
    else if graph_type is entity+event+concept then
        graph, embedding ← Graph(entity, event, concept),
            Embeddings(entity, event, concept)
    end if
end function
function QUERY2EDGE(query, topN)
    Q_emb ← Retriever(query)
    S = Q · W_e                    ▷ Calculate similarity scores with precomputed edge embeddings
    E = argsort_i(S)[: N]          ▷ Select topN edges based on scores
    filtered_edges ← LLM_filter(E)          ▷ Filter edges using Large Language Model
    mapped_edges ← Map_edges(filtered_edges)          ▷ Map filtered edges to original edges
    return_node_scores ← Calculate_node_scores(mapped_edges)
    return return_node_scores
end function
function QUERY2PASSAGE(query, weight_adjust)
    Q_pass ← Encode(query)          ▷ Encode query into passage representation
    S_text ← Similarity_Scores(Q_pass, text_embeddings)
    return Scores_Dictionary(S_text)
end function
function RETRIEVE_PERSONALIZATION_DICT(query, topN)
    node_dict ← query2edge(query, topN)
    text_dict ← query2passage(query, weight_adjust)
    return node_dict, text_dict
end function
function RETRIEVE_PASSAGES(query, topN)
    node_dict, text_dict ← retrieve_personalization_dict(query, topN)
    if node_dict is empty then
        return TopN_Text_Passages(text_dict)
    else
        personalization_dict ← {node_dict, text_dict}
        page_rank_scores ← PageRank(personalization_dict)
        return TopN_Passages(page_rank_scores)
    end if
end function
```

Algorithm 3 LargeKGRetriever

A variant of HippoRAG2 (Gutiérrez et al., 2025), optimized with dynamic graph sampling and common word filtering

```
function INIT(graph_type)
    keyword ← Default_graph_keyword                ▷ Keyword can be cc, pes2o, wiki
    Initialize_resources(keyword)
    Load_node_and_edge_indexes()
end function
function RETRIEVE_TOPK_NODES(query, top_k_nodes)
    entities ← LLM_NER(query)
    KG_entities ← Encode_and_Search(entities, FAISS_index)
    filtered_keywords ← LLM_filter(KG_entities)
    return filtered_keywords
end function
function RETRIEVE_PERSONALIZATION_DICT(query, number_of_source_nodes)
    topk_nodes ← retrieve_topk_nodes(query, number_of_source_nodes)
    if topk_nodes == {} then
        return {}
    end if
    Update personalization dictionary with topk_nodes
    return Personalization dictionary
end function
function PAGERANK(personalization_dict, topN, sampling_area)
    G_Sample ← Random Walk with Restart Sampling
    Scores = PageRank(G_Sample, personalization_dict)
    topN_nodes = argsort_i(Scores)[: N]
    for node in topN nodes do
        Connected_Passage += node.score
    end for
    return TopN_Ranked_Passages
end function
function RETRIEVE_PASSAGES(query, topN, number_of_source_nodes, sampling_area)
    personalization_dict ← retrieve_personalization_dict(query, number_of_source_nodes)
    if personalization_dict is empty then
        return {}, [0]
    end if
    topN_passages ← pagerank(personalization_dict, topN, sampling_area)
    return topN_passages
end function
```

Model/Dataset	MuSiQue		2Wiki		HotpotQA	
Metric	**Recall@2**	**Recall@5**	**Recall@2**	**Recall@5**	**Recall@2**	**Recall@5**
Baseline Retrievers						
Contriever	34.8	46.6	46.6	57.5	58.4	75.3
BM25	32.4	43.5	55.3	65.3	57.3	74.8
LLM Embeddings						
GTE-Qwen2-7B-Instruct	48.1	63.6	66.7	74.8	75.8	89.1
GritLM-7B	49.7	65.9	67.3	76.0	79.2	92.4
NV-Embed-v2 (7B)	52.7	69.7	67.1	76.5	84.1	94.5
Existing Graph-based RAG Methods						
RAPTOR (Llama-3.3-70B-Instruct)	47.0	57.8	58.3	66.2	76.8	86.9
HippoRAG (Llama-3.3-70B-Instruct)	41.2	53.2	71.9	90.4	60.4	77.3
HippoRAG2 (Llama-3.3-70B-Instruct)	56.1	74.7	76.2	90.4	83.5	96.3
AutoSchemaKG (`LLama-3.1-8B-Instruct`) + HippoRAG1						
Entity-KG (Llama-3-8B-Instruct)	41.37	51.08	61.72	75.45	51.89	65.95
Entity-Event-KG (Llama-3-8B-Instruct)	41.28	51.12	61.37	74.56	51.31	65.93
Full-KG (Llama-3-8B-Instruct)	40.78	50.36	61.08	71.9	52.8	65.4
AutoSchemaKG (`LLama-3.1-8B-Instruct`) + HippoRAG2						
Entity-KG (Llama-3-8B-Instruct)	48.33	72.58	67.34	84.25	77.59	92.16
Entity-Event-KG (Llama-3-8B-Instruct)	48.83	72.7	68.59	85.85	81.26	92.66
Full-KG (Llama-3-8B-Instruct)	49.12	72.48	68.46	84.6	84.17	93.04

Table 13: Recall performance in the knowledge graph created by Llama-3-8B-Instruct shows strong performance that is comparable with the knowledge graph created with 70B model.

where Precision = $\frac{TN}{TN+FN}$ and Recall = $\frac{TN}{TN+FP}$.

We use the Retrieval-Augmented Generation method with different knowledge bases on 3 domains (world knowledge, science and technology, and writing/recommendation) of FELM benchmark. For the math domain and reasoning domain, we use the vanilla setting and their results are the same across different knowledge bases. The detailed results of the 3 domains are shown in Table 14.

I.2 Implementation and Evaluation Details on MMLU

Table 15 presents our classification for organizing MMLU tasks into distinct subject categories, providing a structured framework for domain-specific performance analysis. Table 16 displays comprehensive results across all MMLU subject areas, revealing an important insight: while retrieval-augmented generation enhances performance in knowledge-intensive domains, it can negatively impact performance on reasoning-focused tasks such as mathematics and logical reasoning. This finding aligns with previous research suggesting that RAG may sometimes interfere with LLMs' inherent reasoning capabilities.

Finally, this paper uses LLMs to fix grammar and typos for better writing quality.

Corpus	Method	World Knowledge				Science and Technology				Writing/Recommendation			
		P	R	F1	Acc	P	R	F1	Acc	P	R	F1	Acc
-	-	36.67	29.93	32.96	55.10	15.43	24.51	18.94	50.49	25.95	12.73	17.09	52.69
Text Corpora													
Wikipedia	Random	31.78	27.89	29.71	52.52	7.95	11.76	9.49	43.94	21.95	26.97	24.20	53.78
	BM25	26.82	32.65	29.45	49.31	15.23	38.24	21.79	50.48	29.21	48.69	36.52	62.40
	Dense Retrieval	33.93	38.78	36.19	54.97	16.92	43.14	24.31	53.01	25.22	43.45	31.91	58.68
Pes2o-Abstract	Random	27.36	19.73	22.92	49.86	10.13	15.69	12.31	45.64	26.92	28.84	27.85	56.50
	BM25	32.43	32.65	32.54	53.34	11.18	17.65	13.69	46.54	26.75	32.96	29.53	57.34
	Dense Retrieval	31.43	29.93	30.66	52.50	20.51	47.06	28.57	57.55	25.37	32.21	28.38	56.51
Common Crawl	Random	30.14	29.93	30.03	51.72	11.17	21.57	14.72	45.75	23.73	28.09	25.73	54.91
	BM25	27.21	27.21	27.21	49.71	12.21	25.49	16.51	46.68	27.32	40.82	32.73	59.42
	Dense Retrieval	25.50	34.69	29.39	48.00	15.48	36.27	21.70	50.78	24.41	38.95	30.01	57.27
Knowledge Graph													
Freebase	Think on Graph	26.00	8.84	13.20	49.62	23.76	23.53	23.65	55.15	42.86	12.36	19.19	54.51
ATLAS-Wiki	HippoRAG2	33.33	42.18	37.24	54.98	16.45	50.00	24.76	52.75	32.82	32.21	32.51	59.43
ATLAS-Pes2o		39.17	31.97	35.21	56.51	21.35	40.20	27.89	57.13	30.37	15.36	20.40	54.11
ATLAT-CC		33.80	48.98	40.00	56.18	18.06	52.94	26.93	55.42	24.20	25.47	24.82	54.66

Table 14: Factuality results (%) on different domains of FELM benchmark with different Text Corporas and retrieval methods. P, R, F1, and Acc denote Precision, Recall, F1 score, and Balanced Accuracy, respectively.

Subject	Task
History	high school european history, high school us history, high school world history, prehistory
Formal Logic	formal logic, logical fallacies
Law	international law, jurisprudence, professional law
Philosophy and Ethics	philosophy, moral disputes, moral scenarios, business ethics
Religion	world religions
Medicine and Health	clinical knowledge, college medicine, medical genetics, professional medicine, virology, human aging, nutrition, anatomy
Social Sciences	high school geography, high school government and politics, high school psychology, professional psychology, sociology, human sexuality, us foreign policy, security studies
Economics	high school macroeconomics, high school microeconomics, econometrics
Business and Management	management, marketing, professional accounting, public relations
Math	abstract algebra, college mathematics, elementary mathematics, high school mathematics, high school statistics
Natural Sciences	astronomy, college biology, college chemistry, college physics, conceptual physics, high school biology, high school chemistry, high school physics
Computer Science and Engineering	college computer science, high school computer science, computer security, electrical engineering, machine learning
Global Facts	global facts, miscellaneous

Table 15: The correspondence between subjects and tasks.

Corpus	MMLU													
	overall	History	Law	Religion	PaE	MaH	GF	BaM	SS	Logic	Econ	Math	NS	CSaE
None	69.18	76.59	66.86	83.04	63.55	70.38	66.72	72.20	79.74	64.35	68.35	57.31	65.27	66.70
Freebase-ToG	**70.36**	**78.42**	69.00	75.44	65.67	72.65	67.27	**73.67**	76.00	66.03	67.34	60.56	69.39	**68.23**
Random Baseline														
Wikipedia	68.06	76.64	66.82	79.53	59.26	70.34	66.46	67.34	77.78	59.21	65.35	60.91	67.52	61.87
Common Crawl	67.93	74.89	66.52	79.53	61.74	69.82	68.11	67.67	77.52	59.30	64.20	59.80	67.42	62.22
Pes2o-Abstract	68.07	76.24	64.16	80.70	62.01	70.62	66.59	69.27	77.16	62.39	64.70	60.07	66.41	62.18
Text Corpora + DBM25														
Wikipedia	68.99	76.67	67.35	78.36	63.34	69.35	61.98	71.39	76.99	62.30	65.56	61.67	69.31	65.60
Common Crawl	68.33	76.15	66.36	80.12	60.43	69.58	64.67	69.47	76.71	63.18	68.22	62.26	65.55	65.04
Pes2o-Abstract	69.04	78.01	65.89	78.95	63.83	71.01	65.78	68.29	77.07	59.34	**68.87**	61.20	67.73	65.74
Text Corpora + Dense Retrieval														
Wikipedia	69.37	73.59	69.60	79.53	63.58	70.82	62.41	72.57	76.83	62.21	67.35	61.79	**69.81**	65.39
Common Crawl	67.03	74.47	68.98	79.53	60.46	69.29	64.09	68.88	75.21	61.86	62.27	57.13	64.54	64.47
Pes2o-Abstract	69.07	75.79	61.82	78.36	65.15	69.72	66.77	69.02	76.47	63.05	63.07	**63.92**	69.53	67.86
ATLAS + HippoRAG2														
ATLAS-Wiki	68.22	76.73	67.38	84.21	**66.01**	70.82	68.36	72.35	79.16	63.65	64.53	50.09	65.45	62.10
ATLAS-CC	68.26	78.16	**70.85**	83.04	65.60	71.28	63.95	68.84	78.16	65.42	67.51	52.87	63.32	63.40
ATLAS-Pes2o	69.19	77.13	68.41	81.29	65.05	**72.75**	65.67	72.32	81.19	62.98	65.29	54.24	65.41	64.04
ATLAS + ToG														
ATLAS-Wiki	68.29	77.91	66.60	84.21	65.10	70.69	63.85	70.49	78.31	67.08	66.24	54.41	63.65	64.18
ATLAS-CC	68.40	77.07	68.18	83.63	65.24	72.03	66.87	71.21	79.72	66.59	66.42	48.74	63.97	64.22
ATLAS-Pes2o	68.97	77.52	66.95	**84.80**	63.44	71.15	**68.92**	69.98	**81.59**	**67.87**	66.17	55.46	63.35	64.75

Table 16: Performance comparison of our knowledge graph (KG) integrated with HippoRAG2 and ToG against baseline retrieval methods (Random, BM25, Dense Retrieval) across Wikipedia, Common Crawl, and Pes2o-Abstract corpora on MMLU benchmarks. Tasks are classified according to subjects, with bold and underline indicating the highest and the second highest performance. PaE, MaH, GF, BaM, SS, Econ, NS and CSaE represent Philosophy_and_Ethics, Medicine_and_Health, Global_Facts, Business_and_Management, Social_Sciences, Economics, Natural_Sciences and Computer_Science_and_Engineering respectively.

Multiple-Choice Question Generation and Answering

MCQ Generation Prompt:

You are an expert in generating multiple-choice questions (MCQs) from scientific texts. Your task is to generate 5 multiple-choice questions based on the following passage.

Each question should:

- Focus on factual claims, numerical data, definitions, or relational knowledge from the passage.
- Have 4 options (one correct answer and three plausible distractors).
- Clearly indicate the correct answer.

The output should be in JSON format, with each question as a dictionary containing:

- "question": The MCQ question.
- "options": A list of 4 options (e.g., ["A: ..", "B: ..", "C: ..", "D: .."]).
- "answer": The correct answer (e.g., "A").

Output Example:

```
[
  {
    "question": "What is the primary role of a catalyst in a chemical reaction?",
    "options": [
      "A: To make a thermodynamically unfavorable reaction proceed",
      "B: To provide a lower energy pathway between reactants and products",
      "C: To decrease the rate of a chemical reaction",
      "D: To change the overall reaction itself"
    ],
    "answer": "B"
  }
]
```

Passage: {passage}

MCQ Answering Prompt:

Given the contexts or evidences: {contexts}

Here is a multiple-choice question:

Question: {question}

Options: A. {options_0} B. {options_1} C. {options_2} D. {options_3}

Please select the correct answer by choosing A, B, C, or D. Respond with only the letter of your choice.

Figure 8: The prompts for generating and answering MCQ questions for evaluating knowledge retention in knowledge graph.

Question: When did the country the top-ranking Warsaw Pact operatives came from, despite it being headquartered in the country where A Generation is set, agree to a unified Germany inside NATO?

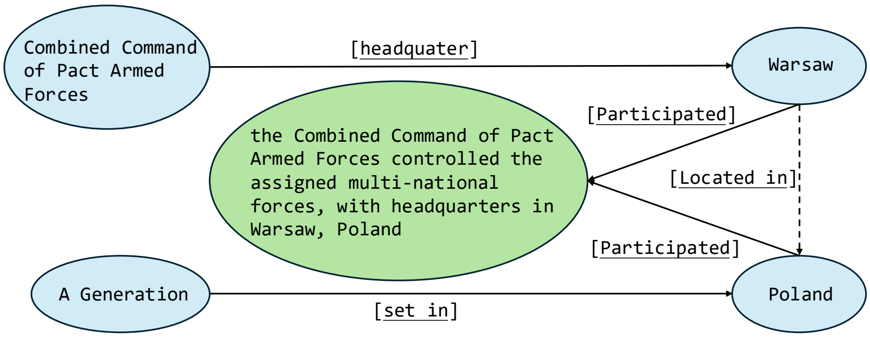

Figure 9: Event Node (green) offers enriched context over triplets (blue); dotted line indicates missing edge.

Question: Who won the Indy Car Race in the largest populated city of the state where the performer of Mingus Three is from?

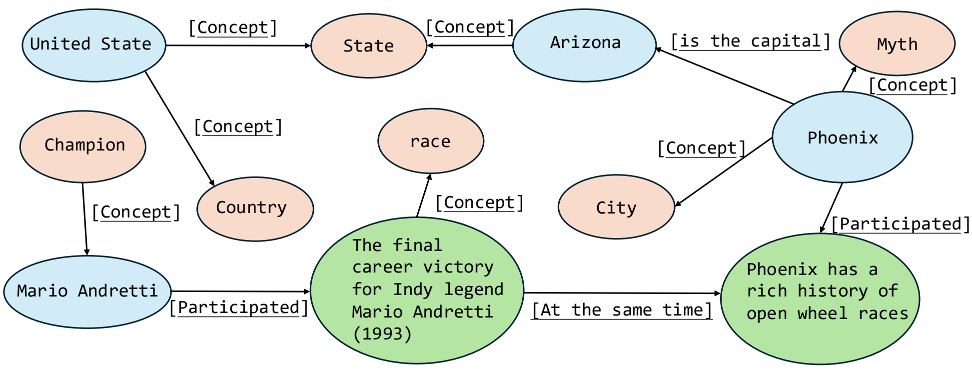

Figure 10: Concept nodes (orange) provide alternate pathways to access information beyond entities and events.