Diagnosing and Addressing Pitfalls in KG-RAG Datasets: Toward More Reliable Benchmarking

Liangliang Zhang[1], Zhuorui Jiang[2], Hongliang Chi[1], Haoyang Chen[1], Mohammed Elkoumy[1], Fali Wang[3], Qiong Wu[4], Zhengyi Zhou[4], Shirui Pan[5], Suhang Wang[3], Yao Ma[1]

[1]Rensselaer Polytechnic Institute, [2]University of Toronto, [3]Pennsylvania State University, [4]AT&T Chief Data Office, [5]Griffith University

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation: The Diagnosis

Large language models with retrieval-augmented generation (RAG) are increasingly used for knowledge-intensive tasks. For knowledge-graph RAG (KG-RAG), progress depends heavily on benchmark datasets.

- We manually inspected over 1,000 QA pairs sampled from **16 widely used datasets**, which reveals an average factual correctness rate of only **57%**.

Dataset	KG	Year	Sample / Total	Correctness (%)
WebQSP	Freebase	2016	100 / 1639	52.00
ComplexWebQuestions	Freebase	2018	300 / 3531	49.33
ComplexQuestions	Freebase	2016	60 / 800	63.33
GraphQuestions	Freebase	2016	60 / 2607	70.00
QALD	DBpedia	2018	60 / 150	61.67
MetaQA	WikiMovies	2018	60 / 39093	20.00
SimpleDBpediaQA	DBpedia	2018	60 / 8595	43.33
CSQA	Wikidata	2018	60 / 27797	65.00
LC-QuAD	DBpedia/Wikidata	2017/2019	60 / 7046	38.34
FreebaseQA	Freebase	2019	60 / 3996	98.67
CFQ	Freebase	2020	60 / 239357	71.67
GrailQA	Freebase	2020	60 / 13231	30.00
QALD-Plus	DB/Wikidata	2022	60 / 136	63.33
KQA-Pro	FB15k+Wikidata	2022	60 / 11797	66.67
DynamicKGQA	YAGO	2025	60 / 40000	45.00

Pitfalls of Existing KGQA Benchmarks

Inaccurate Ground Truth Answers.

A major issue in existing KGQA benchmarks is **incorrect or unreliable answers**, leading to **misleading evaluation** and **penalizing correct model predictions**.

- **Incorrect annotations**
- **Outdated answers**
- **Incomplete annotations**

Q: "When did Michael Jordan return to the NBA?"
A: 1984
Issue: 1984 is the year of his NBA debut, not his return. The correct answer should be 1995

Low-Quality or Ambiguous Questions.

Many KGQA benchmarks contain **poorly constructed or underspecified questions**, which reduce evaluation reliability and fail to test real reasoning ability.

- **Ambiguous phrasing**
- **Low-complexity questions**
- **Unanswerable, subjective, or ill-formed questions**

Q: "What does George Wilson do for a living?"
A: American football player
Issue: Multiple well-known individuals share the name George Wilson (e.g., *The Great Gatsby* character, *Dennis the Menace*, several athletes).

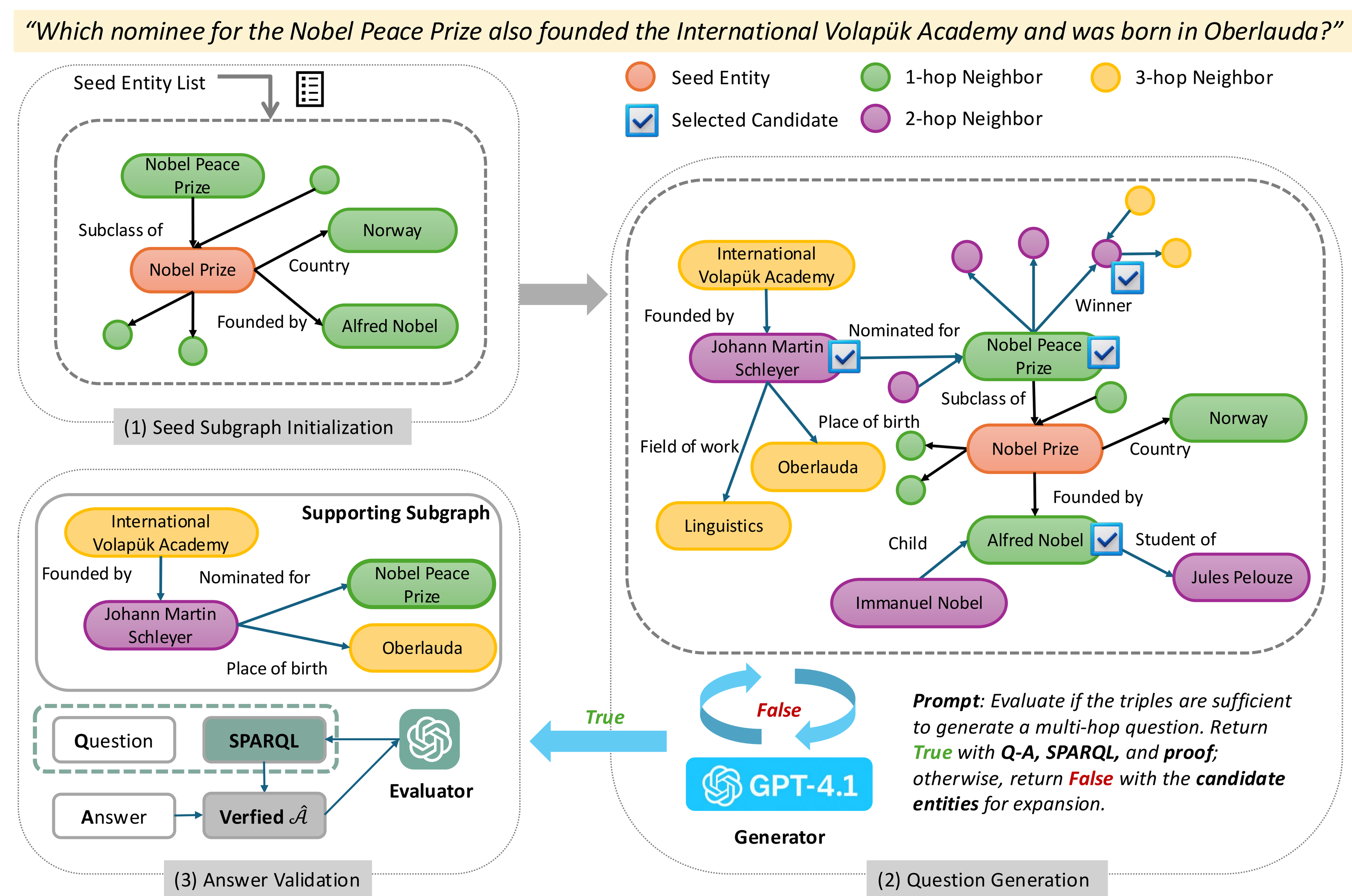
Limitations of Exact-Match Evaluation.

Rely on rigid exact-match criteria that fail to account for **semantically correct answers expressed in different surface forms**.

Q: "How big is the Earth's diameter?"
A: 1.2742e+07
Issue: "12,742 km," "12,742,000 meters," "about 7,918 miles," and "approximately 12,700 kilometers." Variations in units, notation, or approximation are all reasonable.

KGQAGen Framework

Reliable, scalable, and challenging: Built from real Wikidata subgraphs and **SPARQL-verified** for correctness and reproducibility. **LLM-guided KG exploration** enables large-scale, diverse **multi-hop** question generation, providing a benchmark that tests **real reasoning**, not template matching or trivial 1-hop lookup..



Stage 1: Seed Subgraph Construction

- Start from a seed entity and extract a local 1-hop subgraph to define the knowledge scope and candidate reasoning space.

Stage 2: LLM-Guided Subgraph Expansion & Question Generation

- Iteratively expand the subgraph if **insufficient** for meaningful reasoning.
- LLM decides whether more structure is needed and then generates a question, candidate answer(s), and supporting subgraph.

Stage 3: Symbolic Validation & Refinement

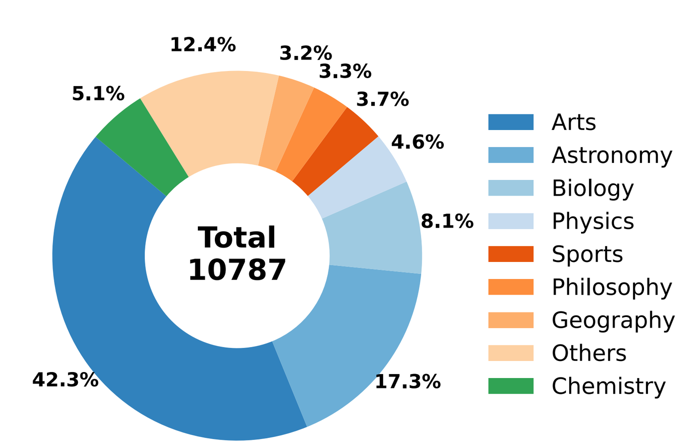
- Execute the SPARQL query to verify answer correctness.
- If a mismatch occurs, apply refinement loops; discard if unresolved.

Output: question, answer, supporting subgraph, and executable SPARQL query.

KGQAGen-10k Dataset

A **10,787-instance**, **SPARQL-verified**, **multi-hop** KGQA dataset generated by KGQAGen.

- **High quality: 96.3%** manually verified correctness (289/300).
- **Non-trivial reasoning: 98% require 2–5 hops**, with **5–30 entities** and **4–28 relations** per supporting graph.
- **Linguistic diversity: 61.1% questions** are 16–30 words; only **7.5%** are short factoids.



Topic Coverage

Experiment Results

Benchmarked KG-RAG frameworks and LLM baselines on **KGQAGen-10k** using a standardized **8,629 / 1,079 / 1,079** split, evaluated with **exact-match** and **LM-Assisted Semantic Match (LASM)** scoring.

Baselines

- **RoG**, **ToG**, **PoG**, **GCR**
- **LLMs:** GPT-4, GPT-4o, GPT-4.1, GPT-4o-mini, DeepSeek-Chat, LLaMA-3.1-8B-Instruct, LLaMA2-7B, Mistral-7B-Instruct-v0.2

Type	Model	Accuracy		Hit@1		F1		Precision		Recall	
		EM	LASM	EM	LASM	EM	LASM	EM	LASM	EM	LASM
Pure LLM	LLaMA-3.1-8B-Instruct	8.87	11.91	9.27	12.42	8.97	11.98	9.27	12.42	8.87	11.81
	LLaMA2-7B	6.88	12.32	7.23	12.88	6.95	12.34	7.23	13.72	6.88	11.96
	Mistral-7B-Instruct-v0.2	24.98	32.34	26.51	34.38	25.36	33.20	26.60	34.85	24.98	32.72
	GPT-4o-mini	32.34	42.49	34.11	44.39	32.74	42.91	34.11	44.86	32.34	42.35
	GPT-4	42.38	51.37	44.95	54.49	43.01	52.32	45.13	55.33	42.38	51.48
	DeepSeek-Chat	42.48	51.84	45.51	55.24	43.17	52.64	45.60	55.79	42.48	51.78
	GPT-4o	45.29	54.21	47.91	57.46	45.89	54.93	48.01	57.83	45.29	54.11
	GPT-4.1	**47.43**	**56.96**	**50.05**	**59.96**	**48.03**	**57.72**	**50.05**	**60.33**	**47.43**	**56.95**
RAG-based	RoG (LLaMA2-7B)	20.10	27.28	21.32	28.92	17.75	24.26	17.65	24.79	20.10	27.16
	GCR (LLaMA-3.1 + GPT-4o)	49.37	58.96	52.46	62.84	49.30	58.88	50.61	60.76	49.37	59.18
	ToG (GPT-4o)	49.65	59.89	52.55	63.02	50.38	60.73	53.01	64.23	49.65	59.76
	PoG (GPT-4o)	**50.67**	**60.18**	**54.03**	**63.95**	**51.47**	**61.30**	**54.31**	**64.78**	**50.67**	**60.34**
LLM-SP	LLaMA2-7B (w/ SP)	69.79	73.79	73.12	77.76	70.43	74.55	72.81	77.67	69.79	73.76
	GPT-4o (w/ SP)	**82.46**	**84.89**	**89.62**	**92.22**	**84.07**	**86.75**	**89.81**	**93.23**	**82.46**	**84.95**

Key Findings

- **Retrieval is the major bottleneck: LLM-SP models dominate** (e.g., GPT-4o: **54.2%** → **84.9%** with supporting subgraph).
- **Multi-hop reasoning failure:** Even the strongest LLMs struggle to **retrieve and follow correct reasoning paths.**
- **Semantic evaluation matters: LASM > EM** across all models, showing that many answers marked incorrect under exact-match are actually correct.

Takeaways

➢ Existing KGQA benchmarks contain **quality issues**, including **inaccurate ground truth answers**, **low-quality or ambiguous questions**, and a **rigid exact-match evaluation** that penalizes semantically correct responses.

➢ **KGQAGen** provides **a scalable, KG-grounded, and SPARQL-verified framework** that systematically overcomes these limitations and enables the construction of reliable, multi-hop benchmarks.

➢ Experiments show that **high-quality retrieval remains the central bottleneck** in current KG-RAG systems, even when paired with advanced LLMs.

NEURAL INFORMATION PROCESSING SYSTEMS

GitHub

Hugging Face

paper

code

dataset