NEURAL INFORMATION PROCESSING SYSTEMS

# 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]

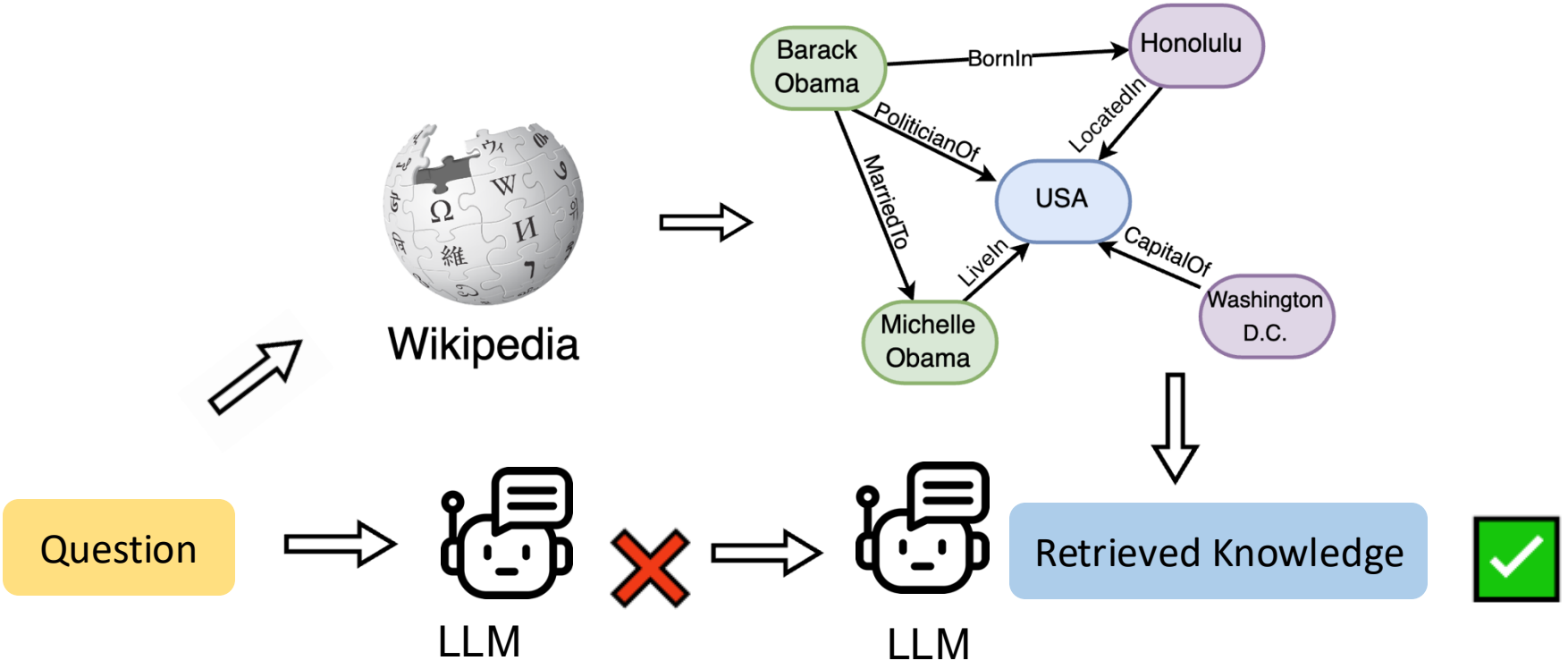
*[1]Rensselaer Polytechnic Institute, [2]University of Toronto, [3]Pennsylvania State University, [4]AT&T Chief Data Office, [5]Griffith University*

# KG-RAG

- **Knowledge Graph Retrieval-Augmented Generation (KG-RAG)** combines LLMs with structured knowledge graphs to achieve **accurate, grounded, and explainable reasoning.**

# KG-RAG

- **Knowledge Graph Retrieval-Augmented Generation (KG-RAG)** combines LLMs with structured knowledge graphs to achieve **accurate, grounded, and explainable reasoning.**




KG-RAG General Workflow

# Where Are We Now?

- While many works keep improving KG-RAG models,
  we asked: **how do we actually evaluate our progress?**

| Author [Citation] | Model | WebQSP | CWQ | Year | Short Introduction |
|---|---|---|---|---|---|
| Mavromatis et al. [34] | ReaRev | ✓ | ✓ | 2022 | LLM + GNNs refine reasoning on incomplete graphs. |
| Yu et al. [71] | DECAF | ✓ | ✓ | 2022 | Joint answer/logical form decoding from free-text retrieval. |
| Sun et al. [55] | ToG | ✓ | ✓ | 2023 | LLM agent explores KGs via beam search for deep, interpretable reasoning. |
| Luo et al. [32] | RoG | ✓ | ✓ | 2023 | Relation-grounded KG paths guide LLM reasoning with explanations. |
| Luo et al. [31] | ChatKBQA | ✓ | ✓ | 2023 | LLM-generated logical forms, improved with KG retrieval. |
| Wang et al. [63] | KD-CoT | ✓ | ✓ | 2023 | External KG knowledge injected into CoT reasoning. |
| Liu et al. [30] | DualR | ✓ | ✓ | 2024 | GNN for structural reasoning, frozen LLM for semantic reasoning. |
| Luo et al. [33] | GCR | ✓ | ✓ | 2024 | KG-Trie constrains LLM decoding for logic-faithful KG reasoning. |
| Dong et al. [13] | Effi-QA | ✓ | ✓ | 2024 | Iterative LLM planning, KG exploration, and self-reflection for QA. |
| Mavromatis et al. [35] | GNN-RAG | ✓ | ✓ | 2024 | GNN-based subgraph reasoning with LLM in RAG pipeline. |
| Xu et al. [67] | READS | ✓ | ✓ | 2024 | LLM decomposes KGQA into retrieval, pruning, inference. |
| Li et al. [26] | DoG | ✓ | ✓ | 2024 | LLM generates "well-formed chains" via constrained decoding. |
| Fang et al. [17] | KARPA | ✓ | ✓ | 2024 | LLM pre-plans, matches KG paths, reasons in training-free manner. |
| Xu et al. [69] | GoG | ✓ | ✓ | 2024 | LLM agent selects, generates, reasons on incomplete KGs. |
| Zhan et al. [72] | RARoK | ✓ | ✓ | 2024 | RAG-augmented CoT for complex medical KGQA. |
| Li et al. [27] | SubgraphRAG | ✓ | ✓ | 2024 | MLP + triple-scoring for efficient subgraph extraction. |
| Fang et al. [16] | DARA | ✓ | | 2024 | LLM decomposes and grounds formal KG queries. |
| Hu et al. [22] | GRAG | ✓ | | 2024 | Text-to-graph, retrieves/prunes subgraphs for RAG. |
| Xiong et al. [66] | Interactive-KBQA | ✓ | ✓ | 2024 | LLM agent generates SPARQL via multi-turn KB interaction. |
| Dehghan et al. [12] | EWEK-QA | ✓ | ✓ | 2024 | Web retrieval + KG triple extraction for citation-based QA. |
| Chen et al. [9] | PoG | ✓ | ✓ | 2024 | Self-correcting LLM planner for decomposed KGQA. |
| Wen et al. [65] | CLEAR-KGQA | ✓ | ✓ | 2025 | Interactive clarification and Bayesian inference for ambiguity. |
| Tan et al. [57] | Path-Over-Graphs | ✓ | ✓ | 2025 | LLM agent explores/prunes multi-hop KG paths. |
| Wang et al. [64] | ReKnoS | ✓ | ✓ | 2025 | Aggregates "super-relations" for LLM forward/backward reasoning. |
| Xu et al. [68] | MemQ | ✓ | ✓ | 2025 | Memory module separates LLM reasoning from KG tool use. |
| Gao et al. [18] | FRAG | ✓ | ✓ | 2025 | Modular KG-RAG adapts retrieval to query complexity. |
| Shen et al. [50] | RwT | ✓ | ✓ | 2025 | LLM-guided MCTS refines KG reasoning chains. |
| Solanki et al. [51] | Efficient-G-Retriever | ✓ | | 2025 | Attention-based subgraph retriever for LLM-aligned RAG. |
| Tang et al. [58] | GGI-MAB | ✓ | ✓ | 2025 | Multi-armed bandit adapts RAG retrieval for KGQA. |
| Zhang et al. [75] | TrustUGA | ✓ | | 2025 | Unified Condition Graph, two-level LLM querying. |

Around 30 papers released between 2022 and 2025.

# Datasets Quality Issues

A manual audit of **16 popular KGQA** datasets reveals an average factual correctness rate of only **57%.**

| Dataset | KG | Year | Sample / Total | Correctness (%) |
|---|---|---|---|---|
| WebQSP [70] | Freebase | 2016 | 100 / 1639 | 52.00 |
| CWQ [56] | Freebase | 2018 | 300 / 3531 | 49.33 |
| ComplexQuestions [5] | Freebase | 2016 | 60 / 800 | 63.33 |
| GraphQuestions [52] | Freebase | 2016 | 60 / 2607 | 70.00 |
| QALD-9 [37] | DBpedia | 2018 | 60 / 150 | 61.67 |
| MetaQA [76] | WikiMovies | 2018 | 60 / 39093 | 20.00 |
| SimpleDBpediaQA [4] | DBpedia | 2018 | 60 / 8595 | 43.33 |
| CSQA [49] | Wikidata | 2018 | 60 / 27797 | 65.00 |
| LC-QuAD 1.0&2.0 [60, 14] | DBpedia/Wikidata | 2017/2019 | 60 / 7046 | 38.34 |
| FreeBaseQA [23] | Freebase | 2019 | 60 / 3996 | 98.67 |
| CFQ [25] | Freebase | 2020 | 60 / 239357 | 71.67 |
| GrailQA [20] | Freebase | 2020 | 60 / 13231 | 30.00 |
| QALD-9-Plus [46] | DB/Wikidata | 2022 | 60 / 136 | 63.33 |
| KQAPro [8] | FB15k+Wikidata | 2022 | 60 / 11797 | 66.67 |
| Dynamic-KGQA [10] | YAGO | 2025 | 60 / 40000 | 45.00 |

- Low-Quality or Ambiguous **Questions**
- Inaccurate Ground Truth **Answers**

# Question Issues

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

**ID: GrailQAPlus-2101990009000**
**Question:** Which automotive designer designed NA?
**Answer:** Koichi Hayashi, Tom Matano, Bob Hall
**Issue:** The meaning of "NA" is ambiguous and could refer to multiple models or entities.

# Question Issues

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

**ID: CSQA-7**
**Question:** Which sex does Wolfgang Brandstetter belong to?
**Answer:** male
**Issue:** This is a 1-hop lookup question with no requirement for reasoning ability.

# Question Issues

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

**ID: CWQ-2621_7360e892294860c6ef7ad9a10e540e1b**
**Question:** Which author wrote editions of "Notes from My Travels 's husband"?
**Answer:** Brad Pitt
**Issue:** The question is ill-formed and refers to a non-existent book.

# Answer Issues

- **Inaccurate Ground Truth Answers**
  - Incorrect annotations
  - Outdated answers
  - Incomplete annotations

**ID: WebQTest-273**
**Question:** When did Michael Jordan return to the NBA?
**Answer:** 1984
**Issue:** 1984 is the year of his NBA debut, not his return. The correct answer should be 1995.

# Answer Issues

- **Inaccurate Ground Truth Answers**
  - Incorrect annotations
  - Outdated answers
  - Incomplete annotations

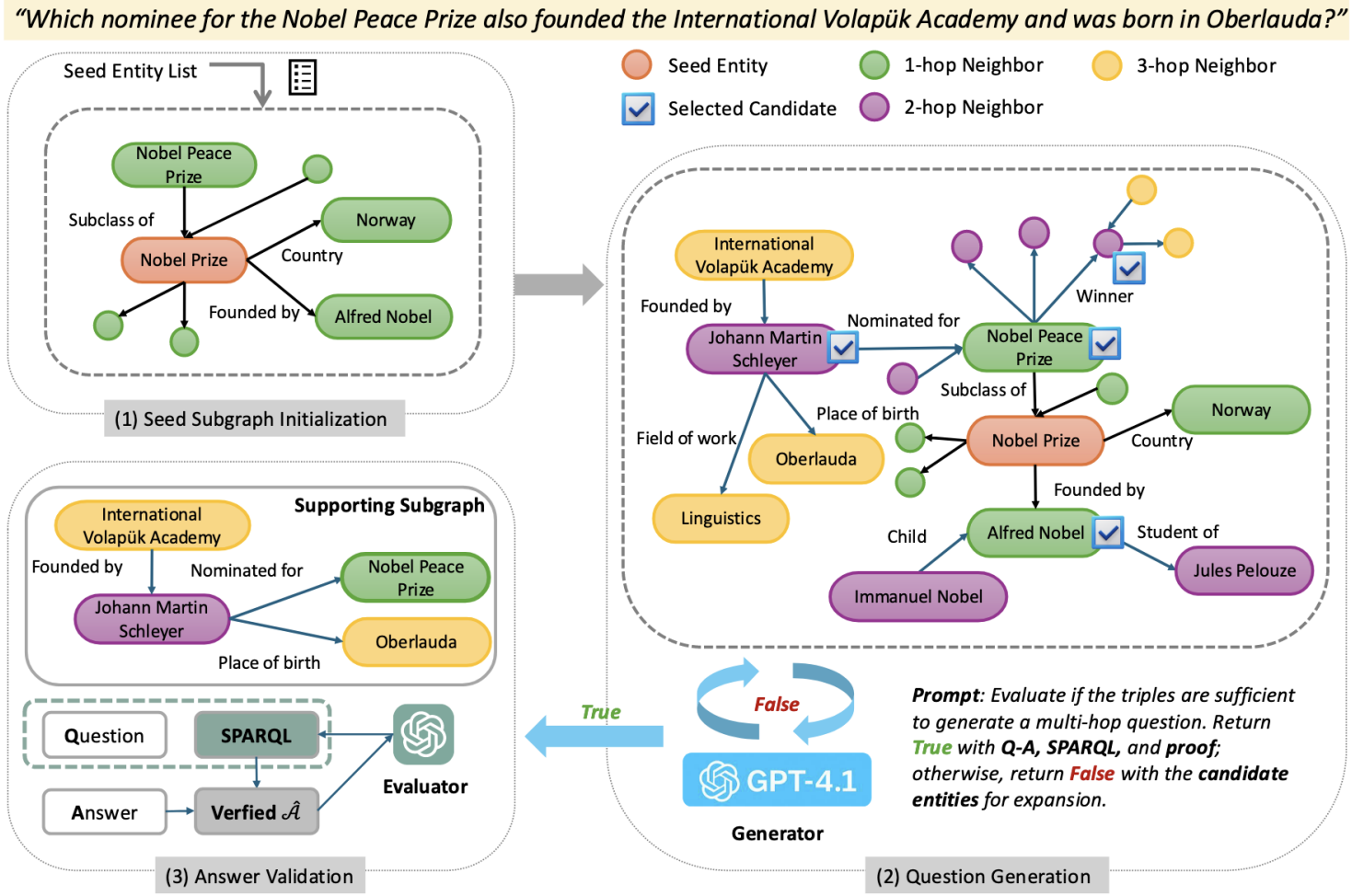
**ID: WebQTest-182**
**Question:** Who is Khloe Kardashian's husband?
**Answer:** Lamar Odom
**Issue:** The answer is outdated; Khloe Kardashian and Lamar Odom divorced in 2016.

# Answer Issues

- **Inaccurate Ground Truth Answers**
  - Incorrect annotations
  - Outdated answers
  - Incomplete annotations

**ID: CWQ-124_7360e892294860c6ef7ad9a10e540e1b**
**Question:** What movie did the author who published editions for *Notes from My Travels* direct?
**Answer:** Unbroken
**Issue:** The annotation is incomplete. The book *Notes from My Travels* was written by Angelina Jolie, who has directed multiple films, including *In the Land of Blood and Honey* (2011), *Unbroken* (2014), *By the Sea* (2015), and *First They Killed My Father* (2017).

# KGQAGen Framework

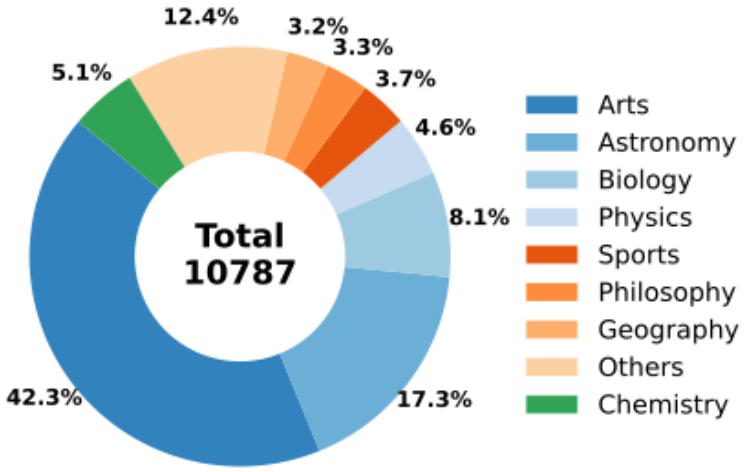



- **Reliable:** Built from real Wikidata and SPARQL-verified.
- **Scalable:** Generated with LLMs using KG context at scale.

# KGQAGen-10k Sample Dataset

- After applying KGQAGen framework in Wikidata knowledge graph, we finally obtain **10,787 verified instances published as KGQAGen-10k.**

- We conducted a manual audit, which found that 289 out of 300 examples **(96.3% )** were correct.




Topic Coverage

# See you at our poster:

*Fri 5 Dec 4:30 p.m.*

*Exhibit Hall C/D/E #110*

Paper:

NEURAL INFORMATION PROCESSING SYSTEMS

Dataset:

Hugging Face