

# Towards Next-Generation Question Answering Over Knowledge Graphs Systems via Accurate Benchmarking and Large-Scale Training

Abdelghny Orogat  
School of Computer Science  
Carleton University

## Abstract

Knowledge graphs (KGs) serve as pivotal information reservoirs, but their intricate structures and the need for expertise in structured query languages limit their accessibility mainly to experts. While the vast potential of KGs in Question Answering (QA) systems is acknowledged, their complexity often requires users to possess specific and advanced querying skills.

Since 2010, the advent of QA systems has ushered in a new era where users can employ natural language for queries, simplifying the process. The creation of numerous benchmarking datasets aids in training these modern QA systems, yet choosing the ideal dataset remains challenging. Current datasets, like *QALD*, come with their own set of limitations, emphasizing the urgency to refine them as KGs gain prominence in sectors such as healthcare and finance. Notably, these benchmarks often lack quantitative comparisons and employ diverse creation methods. Additionally, while knowledge graphs continually evolve, benchmarks remain static, failing to capture the latest updates and nuances.

In light of these challenges, our research offers a comprehensive solution for both benchmarking and training QA systems that are specifically optimized for KGs. Central to our approach is the introduction of CBench [4, 1]. CBench is an innovative, extensible benchmarking suite devised to provide in-depth analyses of existing QA benchmarks. It delves into the intricate linguistic, syntactic, and structural attributes of the questions and queries contained within these benchmarks. Our assessments revealed notable variations across benchmarks in relation to these properties, making the selection of a mere subset of them an unreliable practice for QA system evaluation. CBench stands out by providing not just conventional metrics but also a granular analysis of the linguistic and structural attributes of both answered and unanswered questions. This detailed insight empowers QA system developers by highlighting areas where their systems excel and pinpointing where improvements are needed.

Furthermore, our research introduces Maestro [3, 2], an avant-garde benchmark generation system specifically tailored for QA over KGs. Maestro is equipped to

produce a benchmark for any KG, provided the KG itself and, if available, a relevant text corpus that encompasses the KG’s domain. The benchmarks crafted by Maestro are exhaustive, encapsulating all salient properties of natural language questions and structured queries documented in existing literature, given that the targeted KG embodies these properties. A standout feature of Maestro is its ability to generate superior-quality natural language questions with diverse phrasings, rivaling those generated manually, thereby ensuring a more robust evaluation of QA systems.

## Future Work

Building on Maestro’s ability to generate annotated questions and queries, there’s potential for integration with large language models (LLMs). By utilizing Maestro’s output, including questions, annotated counterparts, and structured queries, we aim to train an LLM to create a superior QA system. The combination of Maestro’s detailed benchmarks and LLM capabilities promises to set new standards in the QA domain over KGs.

## References

- [1] A. Orogat and A. El-Roby. CBench: Demonstrating Comprehensive Evaluation of Question Answering Systems over Knowledge Graphs Through Deep Analysis of Benchmarks. *Proceedings of the VLDB Endowment (PVLDB)*, 14(12), 2021.
- [2] A. Orogat and A. El-Roby. SmartBench: Demonstrating Automatic Generation of Comprehensive Benchmarks for Question Answering over Knowledge Graphs. *Proceedings of the VLDB Endowment (PVLDB)*, 15(12), 2022.
- [3] A. Orogat and A. El-Roby. Maestro: Automatic generation of comprehensive benchmarks for question answering over knowledge graphs. *Proceedings of the ACM on Management of Data*, 1(2):1–24, 2023.
- [4] A. Orogat, I. Liu, and A. El-Roby. CBench: Towards Better Evaluation of Question Answering Over Knowledge Graphs. *Proceedings of the VLDB Endowment (PVLDB)*, 14(8), 2021.