

LLM-Leaderboard for Knowledge Graph related Tasks - Finding the Best Language Model with Data from LLM-KG-Bench

Lars-Peter Meyer^{1,2,*}, Felix Brei¹, Johannes Frey³ and Marvin Hofer³

¹*Institute for Applied Informatics at Leipzig University, Goerdelerring 9, 04109 Leipzig, Germany*

²*Chemnitz Technical University, Germany*

³*Leipzig University, Germany*

Abstract

Large Language Models (LLMs) can benefit from and help with the interaction with RDF Knowledge Graphs. But how can one check what is possible and which LLM to choose? LLM-KG-Bench offers help to evaluate several RDF-related capabilities of many LLMs. Because the number of results is steadily increasing, we propose here a leaderboard for a summarized and configurable overview.

Keywords

Knowledge Graph, RDF, Large Language Model, LLM-Leaderboard, LLM-KG-Bench

1. Introduction

The capabilities of large language models (LLMs) to work with RDF Knowledge Graphs (KGs) sometimes offer new possibilities (see e.g., [1, 2, 3]). Many RDF-related capabilities of LLMs can get evaluated with the LLM-KG-Bench framework [4]¹. In previous work [5, 6, 7, 8, 4, 9] more than 25 benchmark tasks with automated evaluation capabilities were collected. To our knowledge, it is the biggest collection of RDF-related automated benchmark tasks so far. The results get collected in a public GitHub repository² and sparked further scientific evaluations (e.g., [10, 11]). But as the number of results is growing, an overview is missing. Thus, we present here a new leaderboard automatically generated from existing and new results generated with LLM-KG-Bench, including newly defined scores for RDF related capabilities.

2. Architecture

The LLM-KG-Bench Framework collects scores and information values together with LLM interaction data and more log information and stores them in several files. For generating the leaderboard, we are interested in score information stored in JSON files. Score data is defined here as a value from [0.0...1.0] with 1.0 being the best and 0.0 the worst score.

Building upon existing framework functions and code, we added a new aggregation command to generate files for a leaderboard overview based on given result files. This way the process can get automated in a public Git repo and static pages served with services like, e.g., github pages.

Posters, Demos, Blue Sky, and Tutorials at SEMANTiCS 2026, Sep 2026, Ghent, Belgium

*Corresponding author.

✉ LPMeyer@infai.org (L. Meyer); brei@infai.org (F. Brei); frey@informatik.uni-leipzig.de (J. Frey); hofer@informatik.uni-leipzig.de (M. Hofer)

ORCID 0000-0001-5260-5181 (L. Meyer); 0009-0008-5245-6655 (F. Brei); 0000-0003-3127-0815 (J. Frey); 0000-0003-4667-5743 (M. Hofer)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹LLM-KG-Bench Repository: <https://github.com/AKSW/LLM-KG-Bench>

²Repository with results: <https://github.com/AKSW/LLM-KG-Bench-Results>

3. Components and Configuration

The leaderboard contains the following parts:

- a summary table of combined scores, giving a short overview about several capability areas
- an overview of all tasks with one selected score column per task and one row per model
- capability plots (introduced in [4]) based on combined scores
- detailed statistics on all score data collected for all models. As this data is quite big, only its CSV file is linked from the leaderboard page.

For each model, the combined scores get computed as configurable weighted average values of task score values. We newly defined RDF capability scores and reused the SPARQL-related scores as used in [9]. The following list gives an overview of the definition of the combined scores used:

R-Syn-1: RDF-Syntax-Capability on first answer, based on 0_combined scores of RdfSyntaxFixList tasks.

R-Syn-max: RDF-Syntax-Capability on best answer, based on max_combined scores of RdfSyntaxFixList tasks.

R-Sem: RDF-Semantics-Capability based on tasks RdfConnectionExplainStatic and RdfFriendCount.

S-Syn-1: SPARQL-Syntax-Capability on first answer, based on 0_combined score of SparqlSyntaxFixingListLcQuad task and 0_answerParse of Text2SparqlExecEvalList tasks.

S-Syn-max: SPARQL-Syntax-Capability on best answer, based on max_combined score of SparqlSyntaxFixingListLcQuad task and max_answerParse of Text2SparqlExecEvalList tasks.

S-Sem-R: SPARQL-Semantic-Read-Capability, based on combinedF1 score of Sparql2AnswerListOrga-turtle task.

S-Sem-W-1: SPARQL-Semantic-Write-Capability on first answer, based on 0_combined scores of Text2SparqlExecEvalList tasks.

S-Sem-W-max: SPARQL-Semantic-Write-Capability on best answer, based on max_combined scores of Text2SparqlExecEvalList tasks.

4. Results

The leaderboard is published and available at <https://github.com/AKSW/LLM-KG-Bench-Leaderboard>. Additionally, the data files are published as hugging face dataset at <https://huggingface.co/datasets/lpmeyer/LLM-KG-Bench-Leaderboard>.

The leaderboard contains scores from evaluations of more than 45 LLMs. This includes about 40 evaluations from earlier work (especially [4]) as well as, additionally, a selection of 6 updated LLMs from the same main families: Claude 4.6 Opus, Claude 4.6 Sonnet, Gemini 3 Flash Preview, GPT 5.2-chat, GPT 5.4 2026/03 and Qwen 3.5-397B

The leaderboard gives a brief introduction followed by the first section introducing the combined scores and showing a sortable table with entries for each model and combined scores. This table spans 10 columns and 47 data rows. A part of this table is shown in fig. 1.

As the combined scores sometimes hide details, the next section of the leaderboard page shows another table giving the average main score values for each task and LLM. This table contains 27 columns (model name + 26 score columns) and again 47 data rows.

Model	total	R-Syn-1	R-Syn-Max
Qwen-3.5-397B	0.977 (± 0.133)	0.941 (± 0.215)	0.994 (± 0.015)
GPT5.2-chat	0.955 (± 0.178)	0.962 (± 0.168)	0.995 (± 0.014)
Gemini 3 Flash Preview	0.942 (± 0.200)	0.961 (± 0.168)	0.995 (± 0.014)
Claude 3.5 Sonnet	0.933 (± 0.222)	0.950 (± 0.175)	0.990 (± 0.022)
Deepseek-R1	0.931 (± 0.226)	0.955 (± 0.174)	0.991 (± 0.020)
Claude 3.5 Haiku	0.925 (± 0.232)	0.937 (± 0.203)	0.984 (± 0.029)
Qwen-3-235B	0.919 (± 0.246)	0.912 (± 0.260)	0.993 (± 0.017)

Figure 1: Screenshot from a part of the overview table. The full table contains 10 columns(model name + total score + 8 combined scores) and 47 data rows

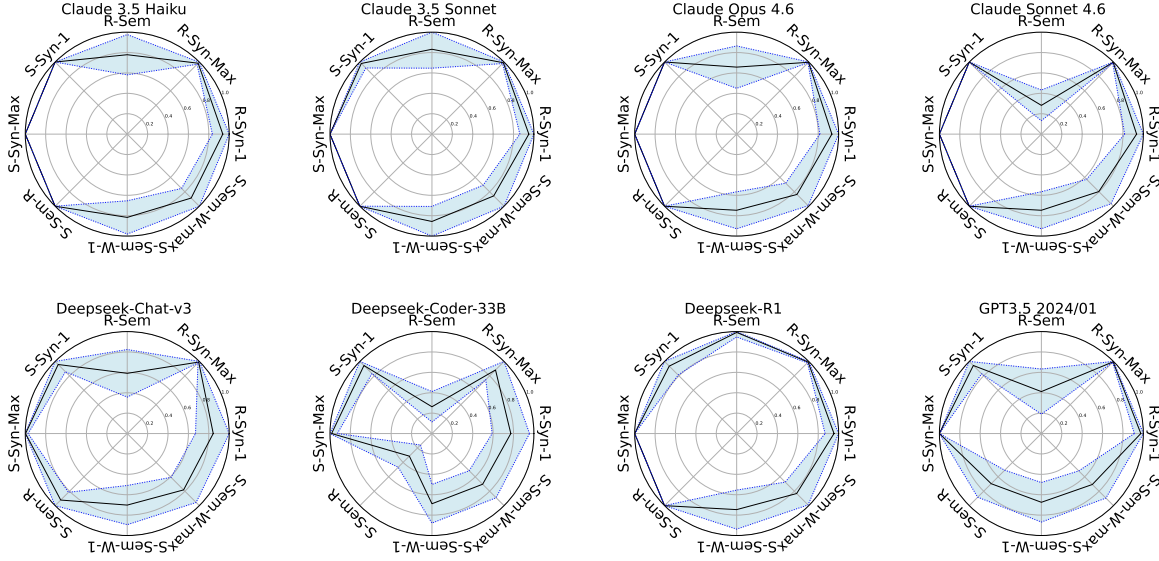


Figure 2: A part of the capability plot list. These capability plots are included for each evaluated LLM and give a graphical overview on the estimated RDF related capabilities of a LLM. The full list contains 47 capability plots.

The third section of the leaderboard page lists the capability plots [4, 9] for all LLMs. The capability plots give a short graphical overview on the 8 combined scores of a LLM. Figure 2 shows a screenshot of the first part of this list, the leaderboard contains 47 capability plots.

The fourth and last section of the leaderboard gives the links to the CSV source files for the tables shown on the page plus the longest data table, too big to show on a single page, containing all mean scores for all tasks and LLMs.

All data tables and plots are published in the GitHub repo, see the section [Online Resources](#).

5. Discussion and Outlook

This leaderboard combines the results of many tasks defined in previous work on LLM-KG-Bench about RDF and SPARQL-related tasks into concise CSV tables and plots shown on the leaderboard page. It gives readers the possibility to easily compare the evaluated LLMs according to either several combined scores or detailed results. This can be done either by sorting the tables accordingly or by detailed analysis using the CSV files provided or by using the data analysis tools on the Hugging Face Dataset.

As we were able to include evaluation data from so many LLMs, one can see quite diverse results. Interestingly, not always does the newest or biggest LLM score best on the RDF-related benchmark tasks. This underscores the need for RDF-related LLM benchmarks.

Further work could offer more data analysis tools directly on the leaderboard page. Especially if the

number of LLMs and tasks grows further, probably some filters would help.

But altogether, the leaderboard presented here provides the opportunity to extend LLM-KG-Bench in new RDF-related areas as well as evaluate more LLMs, and still keep an overview of the growing list of results.

Online Resources

- Leaderboard:
<https://aksw.github.io/LLM-KG-Bench-Leaderboard/>
- Leaderboard Data Repository:
<https://github.com/AKSW/LLM-KG-Bench-Leaderboard>
- Source Code:
<https://github.com/AKSW/LLM-KG-Bench/tree/feature/Leaderboard>
- Dataset Entry at Hugging Face:
<https://huggingface.co/datasets/lpmeyer/LLM-KG-Bench-Leaderboard>

Acknowledgments

We thank Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI) for running many experiments involving open LLMs on their GPU cluster. This work was partially supported by grants from the German Federal Ministry of Education and Research (BMBF) to the projects KupferDigital2 (13XP5230L), PROVIDER (13N17700) and ScaleTrust (16DTM312D).

Declaration on Generative AI

During the preparation of this work, the authors used LLMs and LanguageTool in order to: Grammar and spelling check, Paraphrase and reword. LLMs were also used to help coding. After using these services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

- [1] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, X. Wu, Unifying large language models and knowledge graphs: A roadmap, *IEEE Transactions on Knowledge and Data Engineering (TKDE)* (2024). doi:10.1109/TKDE.2024.3352100.
- [2] S. Liu, S. Semnani, H. Triedman, J. Xu, I. D. Zhao, M. Lam, Spinach: Sparql-based information navigation for challenging real-world questions, in: *Findings of the Association for Computational Linguistics: EMNLP 2024*, Association for Computational Linguistics, 2024, pp. 15977–16001. doi:10.18653/v1/2024.findings-emnlp.938.
- [3] F. Brei, L. Bühmann, J. Frey, D. Gerber, L.-P. Meyer, C. Stadler, K. Bulert, ARUQULA - an LLM based Text2SPARQL approach using ReAct and knowledge graph exploration utilities, in: S. Tramp, M. Gôlo, E. Marx, P. V. do Carmo (Eds.), *Proceedings of the First International TEXT2SPARQL Challenge, Co-Located with Text2KG at ESWC25*, June 01, 2025, Portorož, Slovenia, volume 4094 of *CEUR Workshop Proceedings*, 2025, pp. 49–63. URL: <https://ceur-ws.org/Vol-4094/paper4.pdf>.
- [4] L.-P. Meyer, J. Frey, D. Heim, F. Brei, C. Stadler, K. Junghanns, M. Martin, LLM-KG-Bench 3.0: A Compass for Semantic Technology Capabilities in the Ocean of LLMs, Springer Nature Switzerland, 2025, pp. 280–296. doi:10.1007/978-3-031-94578-6_16.
- [5] L.-P. Meyer, J. Frey, K. Junghanns, F. Brei, K. Bulert, S. Gründer-Fahrer, M. Martin, Developing a scalable benchmark for assessing large language models in knowledge graph engineering, in: N. Keshan, S. Neumaier, A. L. Gentile, S. Vahdati (Eds.), *Proceedings of the Posters and Demo Track*

- of the 19th International Conference on Semantic Systems (SEMANTICS 2023), volume 3526 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023. URL: <https://ceur-ws.org/Vol-3526/paper-04.pdf>.
- [6] J. Frey, L. Meyer, N. Arndt, F. Brei, K. Bulert, Benchmarking the abilities of large language models for RDF knowledge graph creation and comprehension: How well do llms speak turtle?, in: M. Alam, M. Cochez (Eds.), *Proceedings of the Workshop on Deep Learning for Knowledge Graphs (DL4KG 2023) co-located with the 21th International Semantic Web Conference (ISWC 2023)*, Athens, November 6-10, 2023, volume 3559 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023. URL: <https://ceur-ws.org/Vol-3559/paper-3.pdf>.
 - [7] J. Frey, L.-P. Meyer, F. Brei, S. Gruender, M. Martin, Assessing the evolution of LLM capabilities for knowledge graph engineering in 2023, in: *The Semantic Web: ESWC 2024 Satellite Events*, Springer Nature Switzerland, 2025, pp. 51–60. doi:10.1007/978-3-031-78952-6_5.
 - [8] L.-P. Meyer, J. Frey, F. Brei, N. Arndt, Assessing SPARQL capabilities of large language models, in: E. Vakaj, S. Iranmanesh, R. Stamartina, N. Mihindukulasooriya, S. Tiwari, F. Ortiz-Rodríguez, R. Mcgranaghan (Eds.), *Proceedings of the 3rd International Workshop on Natural Language Processing for Knowledge Graph Creation co-located with 20th International Conference on Semantic Systems (SEMANTiCS 2024)*, volume 3874 of *CEUR Workshop Proceedings*, 2024, pp. 35–53. URL: <https://ceur-ws.org/Vol-3874/paper3.pdf>.
 - [9] L.-P. Meyer, J. Frey, F. Brei, D. Heim, S. Gründer-Fahrer, S. Todorovikj, C. Stadler, M. Schröder, N. Arndt, M. Martin, Evaluating large language models for RDF knowledge graph-related tasks — the LLM-KG-Bench-framework 3, *Semantic Web – Interoperability, Usability, Applicability 17* (2026). doi:10.1177/22104968261461351.
 - [10] D. Heim, L.-P. Meyer, M. Schröder, J. Frey, A. Dengel, How do scaling laws apply to knowledge graph engineering tasks? the impact of model size on large language model performance, in: *ESWC 2025 Workshops and Tutorials Joint Proceedings*, volume 3977 of *CEUR Workshop Proceedings*, 2025. URL: <https://ceur-ws.org/Vol-3977/elmke-3.pdf>.
 - [11] S. Todorovikj, L.-P. Meyer, M. Martin, Characterizing knowledge graph tasks in LLM benchmarks using cognitive complexity frameworks, in: D. Chaves-Fraga, I. Heibi, D. Garijo, D. Collarana, A. Salatino, S. Vahdati (Eds.), *Proceedings of SEMANTiCS 2025 Poster Track*, volume 4064 of *CEUR Workshop Proceedings*, 2025. URL: <https://ceur-ws.org/Vol-4064/PD-paper5.pdf>.