

# A Gerbil Learns to Explain<sup>\*</sup>

Quannian Zhang<sup>1,\*</sup>, Hardik Shetty<sup>1</sup>, Michael Röder<sup>1</sup>, Nikit Srivastava<sup>1</sup>,  
N'Dah Jean Kouagou<sup>1</sup> and Axel-Cyrille Ngonga Ngomo<sup>1</sup>

<sup>1</sup>Data Science Group (DICE), Heinz Nixdorf Institute, Paderborn University, Germany

## Abstract

Benchmarking is an indispensable approach for evaluating and comparing the performance of different systems. However, the results are typically summarized using a small set of metrics, and their analysis often requires manual effort and may introduce bias. In this demo, we present an extension of the GERBIL QA benchmarking framework that automatically generates human-readable explanations for the evaluation results of question answering systems. Our approach analyzes benchmarking outcomes and learns concepts that distinguish correctly answered questions from incorrectly answered ones. These concepts are further verbalized by a Large Language Model to produce human-readable explanations. With the help of our explanations, users are allowed to gain deeper insights into system behavior beyond standard evaluation metrics. A live version of the demo is publicly available at <https://gerbil-qa.aksw.org/gerbil/config>.

## Keywords

Benchmarks, Explainability and interpretability, Description logics

## 1. Introduction

Comparable benchmarks are key for the improvement of solutions across disciplines with quantifiable results. This insight has led to the development of a multitude of benchmarking frameworks (e.g., [1, 2, 3]) and online leaderboards in recent years.<sup>1</sup> Among these, GERBIL QA [1] is a framework that focuses on benchmarking question answering (QA) systems.

However, benchmarks are of little use if the results they provide do not lead to actionable insights [4]. We address this challenge in our recent work by proposing an approach for automatically explaining benchmarking results [5]. Our goal of this demonstration is to showcase the practical application of this approach by integrating our newly developed generation of explanations into the open-source GERBIL QA benchmarking framework to enable real-world users to benefit from the generated explanations. Our demonstration includes the evaluation of a question answering system and the generation of human-readable explanations summarizing the strengths of the evaluated system. The GERBIL QA benchmarking framework with our explanation generation in the backend is available online at <https://gerbil-qa.aksw.org/gerbil/config>.

## 2. System Overview

Our demonstration extends GERBIL QA and comprises four main components depicted in Figure 1:

1. The GERBIL QA benchmarking framework proposed by Usbeck et al. [1] is an online platform that enables users to benchmark question answering systems in a repeatable and comparable way abiding by the FAIR data principles [6]

---

*K-CAP 2025 Companion Volume*

<sup>\*</sup>Corresponding author.

✉ [quannian@mail.uni-paderborn.de](mailto:quannian@mail.uni-paderborn.de) (Q. Zhang); [Shettysh13@gmail.com](mailto:Shettysh13@gmail.com) (H. Shetty); [michael.roeder@uni-paderborn.de](mailto:michael.roeder@uni-paderborn.de) (M. Röder); [nikit.srivastava@uni-paderborn.de](mailto:nikit.srivastava@uni-paderborn.de) (N. Srivastava); [ndah.jean.kouagou@upb.de](mailto:ndah.jean.kouagou@upb.de) (N.J. Kouagou); [axel.ngonga@uni-paderborn.de](mailto:axel.ngonga@uni-paderborn.de) (A. N. Ngomo)



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

<sup>1</sup>For example, [https://huggingface.co/spaces/open-llm-leaderboard/open\\_llm\\_leaderboard#/](https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard#/)

2. Our newly developed explanation service [5] learns a description logic concept describing commonalities of correctly answered questions that distinguish them from those questions that the benchmarked system could not answer.
3. Our system is based on a knowledge base that contains data about the benchmark datasets in RDF.
4. A Large Language Module—in our current implementation, GEMMA-3-27B [7]—is used to verbalize the previously learned description logic concept.

The workflow of our extension comprises three steps. First, after the user configured an experiment in GERBIL QA we measure the single F1 scores that the benchmarked QA system achieves for each single question. After that, we separate the questions into two sets. Questions with an F1 score of 0.5 or higher are treated as correctly answered questions while all other questions are added to the set of not correctly answered questions.

In the second step, the explanation service receives the name of the dataset that has been used for the evaluation and the two sets of questions as input. For each supported dataset, we generated a knowledge graph representing the benchmarking dataset. These graphs comprise a representation of each question as resource and additional data about the question that has been gathered from three sources:

1. Features that have been gathered from the natural language question, e.g., a dependency parse tree.
2. The expected answers for the question. In case the answer comprises entities from a reference knowledge graph, we add the concise bounded description (CBD) of these entities as well.
3. The SPARQL query that selects the expected answers from the reference knowledge graph and features of these queries, e.g., entities that are mentioned within the query are added with their CBD.

Given this pre-generated knowledge graph and the provided sets of questions, we use a concept learner to learn a description logic concept that describes the correctly answered questions and what distinguishes them from the incorrectly answered ones. More details about this step are presented in our research paper [5].

In the third step, the learned description logics concept is verbalized by a Large Language Model (LLM). The resulting human-readable explanation is returned to the GERBIL QA system and included into the experiment results.

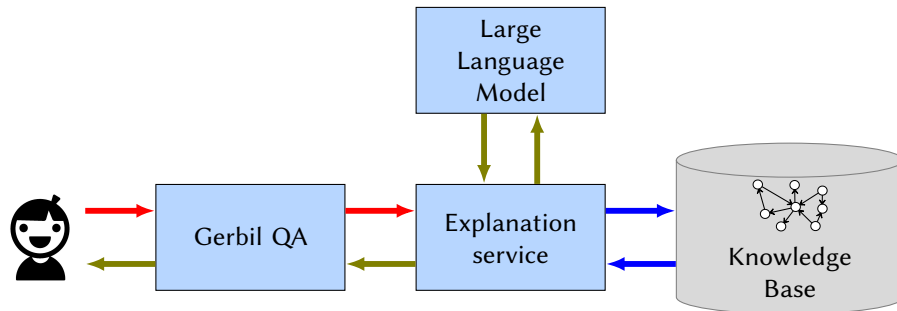


**Figure 1:** Overview of the three steps of our system, where red, blue, and green arrows represent the first, second, and third step, respectively.

GERBIL Experiment Configuration

New Experiment

Experiment Type: QA

Dataset: QALD10 Test Multilingual

Or upload another dataset: Name: Type something

+ Select file...

System: None selected

Or add another webservice via URI: Name: Type something

URI: Type something

Add system

Or add another webservice via URI: Name: Type something

URI: Type something

Add system

Or upload a file with answers: Name: Type something

QALD10 Test Multilingual

+ Select file...

✖ QALD10-Deeppavlov(qald10\_deeppavlov\_en\_Result.json) (undefined)(QALD10 Test Multilingual)

Language: Type something: e.g. fr

Disclaimer: ☒ I have read and understand the disclaimer.

Run Experiment

(a) Experiment configuration interface

GERBIL Experiment

Experiment URI: <https://gerbil-qa.aksw.org/gerbil/experiment?id=202601300003>

Type: QA

Matching: Me - strong entity match

| Annotator                    | Dataset                  | Language |             | Micro F1                                                                                                                                                                                                                                                                  | Micro Precision | Micro Recall | Macro F1 | Macro Precision | Macro Recall | Error Count | avg | Macro F1 | QALD Timestamp      | GERBIL version |
|------------------------------|--------------------------|----------|-------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------|--------------|----------|-----------------|--------------|-------------|-----|----------|---------------------|----------------|
| QALD10-Deeppavlov (uploaded) | QALD10 Test Multilingual | en       |             | 0.1013                                                                                                                                                                                                                                                                    | 0.0798          | 0.1388       | 0.1596   | 0.1638          | 0.1715       | 0           | 0   | 0.2581   | 2026-01-30 10:29:52 | 0.3.0          |
| QALD10-Deeppavlov (uploaded) | QALD10 Test Multilingual | en       | Explanation | <p>This QAS can answer questions where the answer is a specific thing (identified by an IRI) that is <b>not</b> a human, a fictional character, or a work of art.</p> <p>Essentially, it answers questions about <i>real-world, non-person, non-fiction</i> entities.</p> |                 |              |          |                 |              |             |     |          |                     |                |
| QALD10-Deeppavlov (uploaded) | QALD10 Test Multilingual | en       | C2KB        | 0.5368                                                                                                                                                                                                                                                                    | 0.6494          | 0.4574       | 0.4687   | 0.4665          | 0.4942       | 0           |     |          | 2026-01-30 10:29:52 | 0.3.0          |
| QALD10-Deeppavlov (uploaded) | QALD10 Test Multilingual | en       | P2KB        | 0.6612                                                                                                                                                                                                                                                                    | 0.7552          | 0.588        | 0.5671   | 0.5535          | 0.5914       | 0           |     |          | 2026-01-30 10:29:52 | 0.3.0          |
| QALD10-Deeppavlov (uploaded) | QALD10 Test Multilingual | en       | RE2KB       | 0.1749                                                                                                                                                                                                                                                                    | 0.239           | 0.1379       | 0.2168   | 0.2331          | 0.21         | 0           |     |          | 2026-01-30 10:29:52 | 0.3.0          |

(b) Result page with evaluation metrics and explanations

**Figure 2:** Overview of the GERBIL QA benchmarking workflow, a citable link to the example experiment results is available at <https://gerbil-qa.aksw.org/gerbil/experiment?id=202601300003>.

### 3. Demonstration Walkthrough

The demonstration starts with a user configuring an experiment in the GERBIL QA web interface, the benchmarking process requires the following two inputs: 1. the benchmark dataset used for the

experiment, and 2. the results generated by the question answering (QA) system. Figure 2a shows the configuration of an experiment example in the GERBIL QA interface. At the moment, our explanation service supports three QALD datasets—QALD9 plus DBpedia, QALD9 plus Wikidata, and QALD10 [8, 9]. Hence, our demonstration focuses on these datasets.

For each experiment, GERBIL QA provides a citeable link [1] that can be used to load the result page of the experiments. This page—shown in Figure 2b—contains a result table that summarizes the evaluation results for the questions answering task as well as several subtasks, like entity linking or relation extraction (See [1] for a detailed description). In the newly added second line of the table, the human-readable explanation generated by our explanation service is provided.

## 4. Conclusion

We present a demonstration of the integration of an explainable benchmarking approach for question answering into GERBIL QA. Our demo showcases how such an integration complements traditional evaluation metrics with human-readable explanations. As future work, we plan to extend the number of knowledge bases to support other established QA benchmarking datasets.

## Declaration on Generative AI

During the preparation of this work, the author(s) used gemma-3-27b in order to: verbalize the learned description logic concept to human-readable explanations in Figure 2b. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication’s content.

## References

- [1] R. Usbeck, M. Röder, M. Hoffmann, F. Conrads, J. Huthmann, A.-C. Ngonga-Ngomo, C. Demmler, C. Unger, Benchmarking question answering systems, *Semantic Web* 10 (2019) 293–304.
- [2] M. Röder, D. Kuchelev, A.-C. Ngonga Ngomo, HOBBIT: A platform for benchmarking Big Linked Data, *Data Science* 3 (2020) 15–35. doi:10.3233/DS-190021.
- [3] M. A. N. Pour, A. Algergawy, F. Amardeilh, R. Amini, O. Fallatah, D. Faria, I. Fundulaki, I. Harrow, S. Hertling, P. Hitzler, M. Huschka, L. Ibanescu, E. Jiménez-Ruiz, N. Karam, A. Laadhar, P. Lambrix, H. Li, Y. Li, F. Michel, E. Nasr, H. Paulheim, C. Pesquita, J. Portisch, C. Roussey, T. Saveta, P. Shvaiko, A. Splendiani, C. Trojahn, J. Vataščinová, B. Yaman, O. Zamazal, L. Zhou, Results of the ontology alignment evaluation initiative 2021, in: *Proceedings of the 16th International Workshop on Ontology Matching co-located with the 20th International Semantic Web Conference (ISWC 2021)*, 2021, pp. 62–108. URL: [https://ceur-ws.org/Vol-3063/oeai21\\_paper0.pdf](https://ceur-ws.org/Vol-3063/oeai21_paper0.pdf).
- [4] K. J. Rohlfing, P. Cimiano, I. Scharlau, T. Matzner, H. M. Buhl, H. Buschmeier, E. Esposito, A. Grimmer, B. Hammer, R. Häb-Umbach, et al., Explanation as a social practice: Toward a conceptual framework for the social design of ai systems, *IEEE Transactions on Cognitive and Developmental Systems* 13 (2020) 717–728.
- [5] Q. Zhang, M. Röder, N. Srivastava, N. J. Kouagou, A.-C. Ngonga Ngomo, Explainable benchmarking through the lense of concept learning, in: *Proceedings of the 13th Knowledge Capture Conference 2025, K-CAP ’25*, Association for Computing Machinery, New York, NY, USA, 2025, p. 139–147. URL: <https://doi.org/10.1145/3731443.3771359>. doi:10.1145/3731443.3771359.
- [6] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al., The fair guiding principles for scientific data management and stewardship, *Scientific data* 3 (2016) 1–9.
- [7] G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière, L. Rouillard, T. Mesnard, G. Cideron, J. bastien Grill, S. Ramos, E. Yvinec, M. Casbon, E. Pot, I. Penchev, G. Liu, F. Visin, K. Kenealy, L. Beyer, X. Zhai, A. Tsitsulin, R. Busa-Fekete, A. Feng,

- N. Sachdeva, B. Coleman, Y. Gao, B. Mustafa, I. Barr, E. Parisotto, D. Tian, M. Eyal, C. Cherry, J.-T. Peter, D. Sinopalnikov, S. Bhupatiraju, R. Agarwal, M. Kazemi, D. Malkin, R. Kumar, D. Vilar, I. Brusilovsky, J. Luo, A. Steiner, A. Friesen, A. Sharma, A. Sharma, A. M. Gilady, A. Goedeckemeyer, A. Saade, A. Feng, A. Kolesnikov, A. Bendebury, A. Abdagic, A. Vadi, A. György, A. S. Pinto, A. Das, A. Bapna, A. Miech, A. Yang, A. Paterson, A. Shenoy, A. Chakrabarti, B. Piot, B. Wu, B. Shahriari, B. Petrini, C. Chen, C. L. Lan, C. A. Choquette-Choo, C. Carey, C. Brick, D. Deutsch, D. Eisenbud, D. Cattle, D. Cheng, D. Paparas, D. S. Sreepathihalli, D. Reid, D. Tran, D. Zelle, E. Noland, E. Huizenga, E. Kharitonov, F. Liu, G. Amirkhanyan, G. Cameron, H. Hashemi, H. Klimczak-Plucińska, H. Singh, H. Mehta, H. T. Lehri, H. Hazimeh, I. Ballantyne, I. Szpektor, I. Nardini, J. Pouget-Abadie, J. Chan, J. Stanton, J. Wieting, J. Lai, J. Orbay, J. Fernandez, J. Newlan, J. yeong Ji, J. Singh, K. Black, K. Yu, K. Hui, K. Vodrahalli, K. Greff, L. Qiu, M. Valentine, M. Coelho, M. Ritter, M. Hoffman, M. Watson, M. Chaturvedi, M. Moynihan, M. Ma, N. Babar, N. Noy, N. Byrd, N. Roy, N. Momchev, N. Chauhan, N. Sachdeva, O. Bunyan, P. Botarda, P. Caron, P. K. Rubenstein, P. Culliton, P. Schmid, P. G. Sessa, P. Xu, P. Stanczyk, P. Tafti, R. Shivanna, R. Wu, R. Pan, R. Rokni, R. Willoughby, R. Vallu, R. Mullins, S. Jerome, S. Smoot, S. Girgin, S. Iqbal, S. Reddy, S. Sheth, S. Pöder, S. Bhatnagar, S. R. Panyam, S. Eiger, S. Zhang, T. Liu, T. Yacovone, T. Liechty, U. Kalra, U. Evci, V. Misra, V. Roseberry, V. Feinberg, V. Kolesnikov, W. Han, W. Kwon, X. Chen, Y. Chow, Y. Zhu, Z. Wei, Z. Egyed, V. Cotruta, M. Giang, P. Kirk, A. Rao, K. Black, N. Babar, J. Lo, E. Moreira, L. G. Martins, O. Sanseviero, L. Gonzalez, Z. Gleicher, T. Warkentin, V. Mirrokni, E. Senter, E. Collins, J. Barral, Z. Ghahramani, R. Hadsell, Y. Matias, D. Sculley, S. Petrov, N. Fiedel, N. Shazeer, O. Vinyals, J. Dean, D. Hassabis, K. Kavukcuoglu, C. Farabet, E. Buchatskaya, J.-B. Alayrac, R. Anil, Dmitry, Lepikhin, S. Borgeaud, O. Bachem, A. Joulin, A. Andreev, C. Hardin, R. Dadashi, L. Hussenot, Gemma 3 technical report, 2025. URL: <https://arxiv.org/abs/2503.19786>. arXiv: 2503.19786.
- [8] A. Perevalov, D. Diefenbach, R. Usbeck, A. Both, Qald-9-plus: A multilingual dataset for question answering over dbpedia and wikidata translated by native speakers, in: 2022 IEEE 16th International Conference on Semantic Computing (ICSC), IEEE, 2022, pp. 229–234.
- [9] R. Usbeck, X. Yan, A. Perevalov, L. Jiang, J. Schulz, A. Kraft, C. Möller, J. Huang, J. Reineke, A.-C. N. Ngomo, M. Saleem, A. Both, QALD-10 – The 10th Challenge on Question Answering over Linked Data, Under review in the Semantic Web Journal (2023). URL: <https://www.semantic-web-journal.net/system/files/swj3357.pdf>.