Preface of the First International TEXT2SPARQL Challenge (TEXT2SPARQL'25)

Edgard Marx[1], Paulo Viviurka do Carmo[1], Marcos Gôlo[2] and Sebastian Tramp[3]

[1]*Leipzig University of Applied Science, Germany*
[2]*University of São Paulo, Brazil*
[3]*eccenca GmbH, Hainstr. 8, 04109 Leipzig, Germany (corresponding editor)*

Introduction

This preface presents information about the challenge, accepted papers, a detailed discussion of the new benchmark datasets, evaluation metrics, and ranking procedure in the following sections. The TEXT2SPARQL challenge invited researchers to participate by sharing an endpoint capable of translating natural language questions into SPARQL queries. Our procedure consisted of three steps: registration, ask, and evaluation. Figure 1 illustrates the challenge procedure with the three steps.

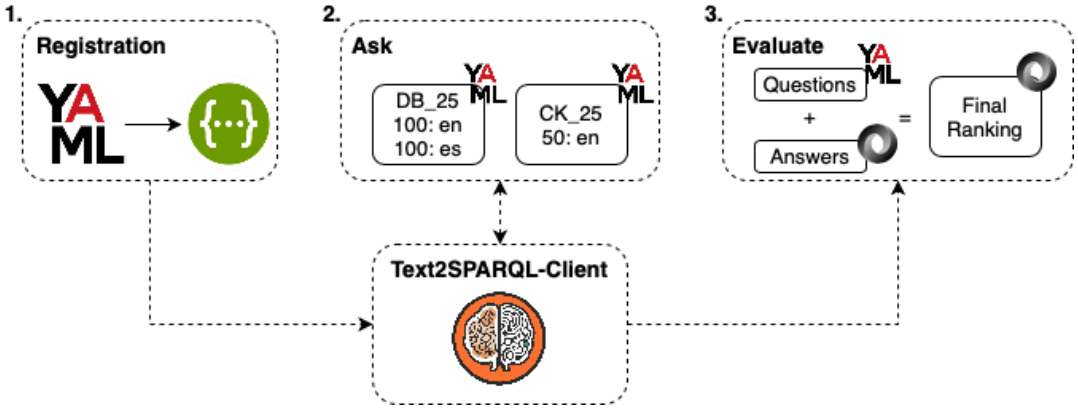

Figure 1: TEXT2SPARQL challenge pipeline with three steps: registration, ask, and evaluation.

Challengers had to register upfront by providing information on an HTTP(S) service, details about the institution, and the participants. The service required from the participants was an endpoint that accepted a GET request with two request parameters: the identifier of the dataset and a natural language question. It also needed to return an object containing the dataset, question, and generated query. A Command-Line Interface (CLI) client was developed that can read a .yml questions file and ask questions to a given endpoint. The client can be accessed at[1]. During the five-day evaluation phase, the organization sent 250 GET requests to each endpoint and recorded the service responses. Finally, using the recorded responses generated with the ask command in the client for each endpoint, an evaluation command for the CLI was developed that takes the question files and recorded responses to calculate the performance metrics. After presenting their strategies at the workshop, the teams submitted papers detailing their strategy. At the end of the review process, six papers were accepted:

First International TEXT2SPARQL Challenge, Co-Located with Text2KG at ESWC25, June 01, 2025, Portorož, Slovenia.
edgard.marx@htwk-leipzig.de (E. Marx); paulo.carmo@htwk-leipzig.de (P. V. d. Carmo); marcosgolo@usp.br (M. Gôlo); sebastian.tramp@eccenca.com (S. Tramp)
https://aksw.org/SebastianTramp (S. Tramp)
0000-0003-4707-2864 (S. Tramp)

[1]https://github.com/AKSW/text2sparql-client

1. **Team INFAI:** ARUQULA - An LLM based Text2SPARQL Approach using ReAct and Knowledge Graph Exploration Utilities
2. **Team MIPT:** AIRI team in Text2SPARQL challenge: Text-To-SPARQL Executor for question-answering over knowledge graphs
3. **Team IIS:** Graf von Data: A Knowledge Graph Question Answering Agent for Organisational Usage
4. **Team AIFB:** Leveraging Data Shapes in Large Language Model Contexts for Question Answering on Public and Private Knowledge Graphs
5. **Team WSE:** Text-to-SPARQL Goes Beyond English: Multilingual Question Answering Over Knowledge Graphs through Human-Inspired Reasoning
6. **Team DBPedia:** Question Answering over DBpedia with Fine-tuned Autoregressive Models

New Benchmark Datasets

The TEXT2SPARQL challenge introduced 250 new question/query pairs over two new benchmark datasets. A DBpedia benchmark with English and Spanish queries from the 2015-10 core, dubbed DB25, and a corporate dataset with a showcase ontology made from scratch to demonstrate the eccenca Corporate Memory capabilities, dubbed CK25. For the DBpedia 200 question-query pairs were created by automatically modifying pairs from QALD 1-8[2] and LCQuaD 1.0[3]. These queries were then rewritten using GPT [1] and manually checked and modified to improve syntax and semantics, as shown in Figure 2. After the human check stage, 35 pairs were deemed initially valid, and 165 were then further checked and modified until 100 question/query pairs were reached. Finally, these questions were then translated into Spanish. For the corporate dataset, 50 questions/query pairs were manually curated, considering classic stakeholders. For details on this new dataset, refer to [2]. It is essential to mention that for both endpoints, we tried to use different SPARQL querying strategies (e.g., ASK, GROUP BY, ORDER BY) in order to balance the endpoints' evaluation.

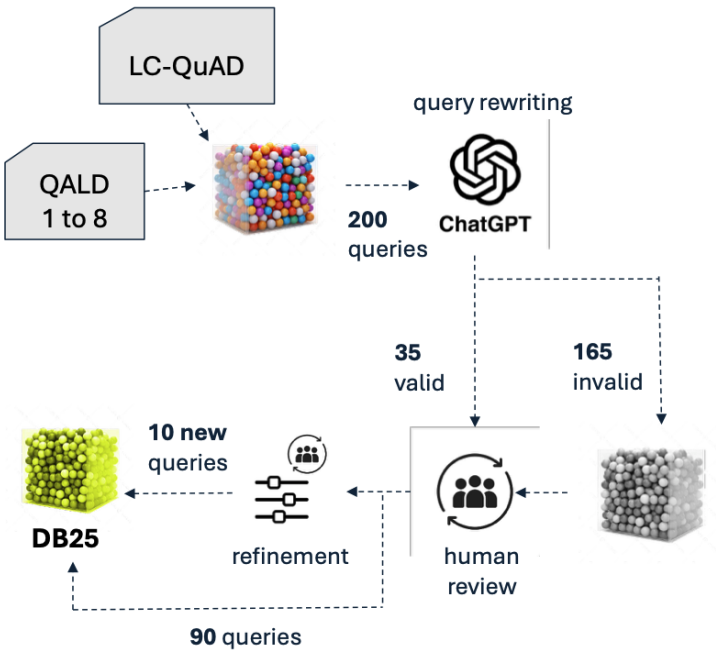

Figure 2: Process of formulating the 200 question/query pairs for DB25.

[2]https://github.com/ag-sc/QALD
[3]https://github.com/AskNowQA/LC-QuAD

Evaluation Metrics and Ranking

The pipeline presented in Figure 3 was used to evaluate the teams. We used Pytrec_eval [3], an information retrieval evaluation tool, to compute information retrieval measures. The challenge team obtained information about the true question/query pairs from the YAML datasets and predicted queries. Then, the challenge team sends these queries to the endpoints to retrieve an answer saved in JSON format. The result is transformed into the Pytrec_eval standard format, consisting of true and predicted lists. Finally, the performance of Pytrec was evaluated by comparing the two lists, which enabled us to calculate the metrics used for the final ranking post-processing.

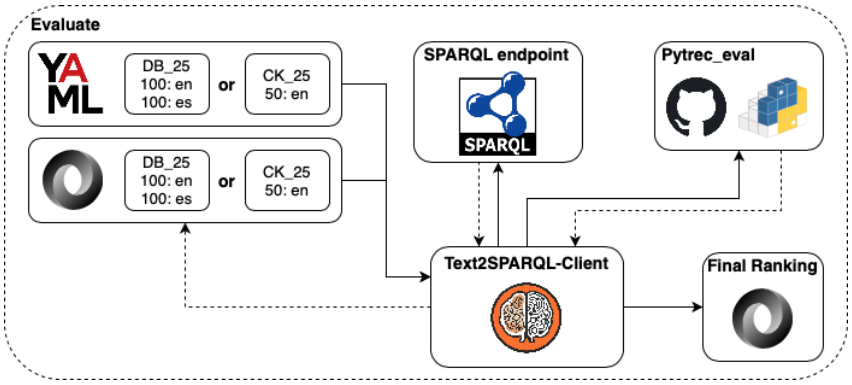

Figure 3: Text2SPARQL-client evaluate command pipeline.

This challenge explores precision, recall, and f_1 metrics. Precision and recall are defined in Equations 1 and 2, in which the precision is defined as the proportion of retrieved documents that are relevant to the user, and the recall is defined as the proportion of relevant documents that were retrieved. f_1 is an evenly weighted harmonic mean of precision and recall, as shown in Equation 3.

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}, \tag{1}$$

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}, \tag{2}$$

$$f_1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \tag{3}$$

There are queries in both datasets where the order matters, as indicated by the flag *RESULT_ORDER_MATTERS* in the YAML files. In these cases, we calculate *nDCG*, a normalized measure for the Discounted Cumulative Gain (*DCG*) metric. The *DCG* compares a position p of where the document was retrieved and penalizes the value based on a logarithmically proportional reduction of the relevance *rel*, as shown in Equation 4. To calculate the $nDCG_p$, where p represents a relevant document ranked in the set, the DCG_p value was used, which is subsequently divided by the ideal ($IDCG_p$). The final *nDCG* score is then obtained by averaging the $nDCG_p$ scores of all retrieved documents.

$$DCG_{\text{p}} = \sum_{i=1}^{p} \frac{rel_i}{\log_2(i+1)} \tag{4}$$

Finally, the organizers create a new metric by considering the averages of the f_1 measure for every question, except those flagged as *RESULT_ORDER_MATTERS*, as shown in Equation 5. All these steps then guarantee a final value between 0 and 1, which considers the maximum retrieval of relevant documents and the order in which the documents were retrieved, where q is the number of questions.

$$f_1\_nDCG = \frac{1}{q} \sum_{i=1}^{q} \begin{cases} nDCG_i & \text{if } RESULT\_ORDER\_MATTERS, \\ f_{1_i} & \text{otherwise.} \end{cases} \tag{5}$$

Text2SPARQL baseline

In recent years, large language models (LLMs) have become a central tool in text mining tasks due to their ability to understand, synthesize, and generate natural language with high accuracy [4]. In tasks involving the translation of natural language into formal representations, such as generating SPARQL queries from natural language questions (text-to-SPARQL), LLMs have demonstrated SoTA results [5]. These models can be explored in various ways, such as using pre-trained versions to more sophisticated approaches involving task-specific fine-tuning. As part of the challenge proposed in this workshop, open-source LLMs were used as baselines, evaluating their performance in controlled and reproducible settings to provide a solid foundation for comparison among participants.

There is a need for datasets that accurately and diversely represent the target task, enabling effective fine-tuning of language models. In the context of this challenge, the focus is on generating SPARQL queries that target the DBpedia ontology. Over the past years, several datasets have been proposed for the text-to-SPARQL task using DBpedia as a reference. Among them, we highlight four main sources employed in our preparation: QALD1-9[4], LC-QuAD 1.0[5], Paraqa[6], and Question-Sparql[7]. These datasets were merged into a unified corpus to train our models robustly. Only queries in English and Spanish were used in these datasets, as these languages were the focus of the challenge. The organizers applied a preprocessing pipeline that involved filtering out inconsistent, duplicate, or non-executable SPARQL queries when tested against the DBpedia endpoint adopted in the challenge. This process ensured that the training data reliably reflected the constraints and characteristics of the target knowledge base.

The model Qwen 2.5 was selected, a high-performance open-source LLM, for fine-tuning [6]. The Unsloth library was used, which implements an efficient fine-tuning strategy based on QLoRA (Quantized Low-Rank Adaptation) [7]. Training was conducted for a total of 100 steps, using a learning rate of 0.001, which was chosen to strike a balance between training time and result quality. During the generation phase, the organizers evaluated the model's performance using four different temperature values (0.01, 0.25, 0.5, and 0.75) to assess the impact of variability on SPARQL query generation. Our results showed that intermediate temperature values, particularly 0.25 and 0.5, outperform the other results. These experimental settings introduced a moderate level of diversity that helped the model produce more accurate and contextually appropriate queries without compromising the syntactic correctness of the SPARQL language. The 1 table presents our baseline results.

Table 1
F1-score on the DBpedia challenge benchmark test set.

LLM models	pre-trained	fine-tuned
qwen 2.5 7B	0.129	0.320
qwen 2.5 14B	0.264	0.410
qwen 2.5 32B	0.152	0.409

Our best baseline result was achieved using the fine-tuned Qwen 2.5 14B model. In comparison, our worst baseline relied on the pretrained Qwen 2.5 7B model. The results demonstrate that fine-tuning significantly improved model performance, leading to the generation of more accurate and contextually appropriate SPARQL queries. We highlight that smaller fine-tuned models outperform larger pretrained models. All three models and the constructed dataset[8] are publicly available on Hugging Face:

- **Text2SPARQL-S** refers to the small version (7B), which requires approximately 6 GB of GPU memory. Model: https://huggingface.co/aksw/text2sparql-S;
- **Text2SPARQL-M** denotes the medium version (14B), which requires approximately 11 GB of GPU memory. Model: https://huggingface.co/aksw/text2sparql-M;

[4]https://github.com/ag-sc/QALD
[5]https://github.com/AskNowQA/LC-QuAD
[6]https://huggingface.co/datasets/Orange/paraqa-sparqltotext
[7]https://huggingface.co/datasets/julioc-p/Question-Sparql
[8]https://huggingface.co/datasets/aksw/Text2SPARQL-Raw

- **Text2SPARQL-L** corresponds to the large version (32B), which uses approximately 20 GB of GPU memory. Model: https://huggingface.co/aksw/text2sparql-L.

Text2SPARQL Awards

Table 2 presents the average f_1_*nDCG* results for every scenario and team. Upon purely analyzing the results, it becomes apparent that some interesting behaviors emerge across the scenarios. Firstly, it is interesting that INFAI's endpoint was the best performer for Corporate and DBpedia English. However, INFAI's endpoint achieved a significantly lower performance for DBpedia Spanish, resulting in it becoming slightly worse than WSE for the overall ranking. Also, considering the DBpedia Spanish results, we can see that WSE achieved the most significant margin win from all scenarios, with almost a 13-point difference to AIFB in second place. One essential aspect to mention is that the results from AIFB and all three endpoints from DBPEDIA (marked with a *) were achieved after modifying the generated queries with the necessary prefixes. All other endpoints were executed without any modifications.

Table 2
f_1_*nDCG* average on the evaluation scenarios. Best is **bold**, second best is underlined, and third best is in *italic*.

Team	Corporate	DBpedia en	DBpedia es	DBpedia	Overall
WSE	.3216	.5245	**.5374**	**.5310**	**.4264**
INFAI	**.4472**	**.5485**	.2075	.3780	.4117
IIS-Q	.4431	.4173	.3014	.3594	*.3914*
IIS-L	*.3707*	.3769	.2907	.3338	.3544
MIPT	.2189	.4466	*.3668*	.4067	.3123
AIFB*	.0000	*.4595*	.4485	.4540	.2279
DBPEDIA-SC*	-	.1080	.0880	.0980	-
DBPEDIA-CL*	-	.0656	.0500	.0578	-
DBPEDIA-CG*	-	.0600	.0356	.0478	-
Challenge Baseline	-	-	-	*.4100*	-

Considering datasets and languages, we have four categories for the Text2SPARQL challenge awards:

1. **Corporate**
   - **1st** INFAI: Daniel Gerber, Lorenz Bühmann, Lars-Peter Meyer, Felix Brei, Claus Stadler
   - **2nd** IIS-Q: Daniel Henselmann, Rene Dorsch, and Andreas Harth
   - **3rd** IIS-L: Daniel Henselmann, Rene Dorsch, and Andreas Harth
2. **DBpedia English**
   - **1st** INFAI: Daniel Gerber, Lorenz Bühmann, Lars-Peter Meyer, Felix Brei, Claus Stadler
   - **2nd** IIS-Q: Daniel Henselmann, Rene Dorsch, and Andreas Harth
   - **3rd** AIFB: Jan Wardenga and Tobias Käfer
3. **DBpedia Spanish**
   - **1st** WSE: Aleksandr Perevalov and Andreas Both
   - **2nd** AIFB: Jan Wardenga and Tobias Käfer
   - **3rd** MIPT: Oleg Somov, Daniil Berezin, and Roman Avdeev
4. **Overall**
   - **1st** WSE: Aleksandr Perevalov and Andreas Both
   - **2nd** INFAI: Daniel Gerber, Lorenz Bühmann, Lars-Peter Meyer, Felix Brei, Claus Stadler
   - **3rd** IIS-Q: Daniel Henselmann, Rene Dorsch, and Andreas Harth

General Chairs

- Edgard Marx, Leipzig University of Applied Sciences (HTWK), Germany
- Sebastian Tramp, eccenca GmbH, Germany
- Diego Moussallem, Paderborn University, Germany

Assistant Committee

- Paulo Viviurka do Carmo, Leipzig University of Applied Sciences (HTWK), Germany
- Marcos Paulo Silva Gôlo, University of São Paulo, Brazil

Program Committee

- Adrian Brasoveanu, Modul University Vienna, Austria
- Aidan Hogan, DCC, Universidad de Chile
- Axel Ngonga, University of Paderborn, Germany
- Andreas Both, HTWK, Germany
- Gong Cheng, Nanjing University, China
- Gustavo Publio, Schwarz IT, Germany
- Muhammad Saleem, University of Paderborn, Germany
- Ricardo Usbeck, Leuphana Universität Lüneburg, Germany
- Ricardo Marcondes Marcacini, USP, Brazil
- Sanju Tiwari, Sharda University, India

Acknowledgements

The editors would like to thank the advisory team, authors, program committee, and other organizers for their ongoing support in making this event a success.

References

[1] OpenAI, GPT-4 Technical Report, 2024. URL: https://arxiv.org/abs/2303.08774. arXiv:2303.08774.
[2] S. Tramp, R. Pietzsch, The CK25 Corporate Knowledge Reference Dataset for Benchmarking Text 2 SPARQL Question Answering Approaches, in: The 1st GOBLIN Workshop on Knowledge Graph Technologies, DBpedia Association, 2025, pp. –. URL: https://github.com/eccenca/ck25-dataset.
[3] C. Van Gysel, M. de Rijke, Pytrec_eval: An extremely fast python interface to trec_eval, in: SIGIR, ACM, 2018, pp. 1–10.
[4] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, et al., A survey on evaluation of large language models, ACM Transactions on Intelligent Systems and Technology (2023).
[5] A. Perevalov, A. Both, A.-C. Ngonga Ngomo, Multilingual question answering systems for knowledge graphs–a survey, Semantic Web 15 (2024) 2089–2124.
[6] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, et al., Qwen technical report, arXiv preprint arXiv:2309.16609 (2023).
[7] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, Qlora: Efficient finetuning of quantized llms, Advances in neural information processing systems 36 (2023).