CompUGE-Bench: Comparative Understanding and Generation Evaluation Benchmark for Comparative Question Answering

Ahmad Shallouf and Irina Nikishina and Chris Biemann
University of Hamburg
Germany
{name.surname}@uni-hamburg.de

Abstract

This paper presents CompUGE, a comprehensive benchmark designed to evaluate Comparative Question Answering (CompQA) systems. The benchmark is structured around four core tasks: Comparative Question Identification, Object and Aspect Identification, Stance Classification, and Answer Generation. It unifies multiple datasets and provides a robust evaluation platform to compare various models across these sub-tasks. We also create additional all-encompassing CompUGE datasets by filtering and merging the existing ones. The benchmark for comparative question answering sub-tasks is designed as a web application available on HuggingFace Spaces.[1,2]

1 Introduction

Nowadays, people are frequently confronted with a wide array of decisions, ranging from mundane tasks, such as selecting a meal, to more significant choices, such as determining a career path or making investment decisions. For instance, when selecting a movie to watch, many would ask, "Which one is better, *Harry Potter* or *The Lord of the Rings*?" Comparative Question Answering (CompQA) in the field of Natural Language Processing (NLP) aims to address exactly these types of questions. The task involves comparing two or more entities across different aspects and providing an answer backed by logical reasoning and argumentation. CompQA systems (Panchenko et al., 2019; Chekalina et al., 2021; Shallouf et al., 2024) help users make informed decisions by retrieving, processing, and generating comparative information.

In previous work (Bondarenko et al., 2022a; Shallouf et al., 2024), four key tasks are identified as essential for building an effective CompQA system: Comparative Question Identification (CQI), Object and Aspect Identification (OAI), Stance Classification (SC), and Answer Generation (AG). Each of these tasks plays a pivotal role in the system's ability to generate meaningful comparative answers. However, one of the main challenges in developing these systems is the diversity of available datasets (four for CQI, three for OAI, and two for SC). They vary in structure, labels, and coverage, making it challenging to compare models consistently using these datasets separately.

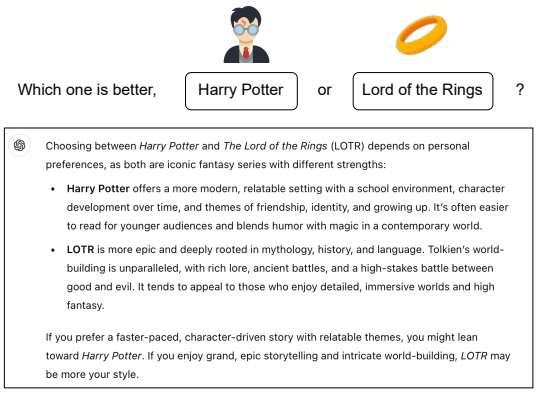


Figure 1: Example of answering comparative questions using ChatGPT.

To address this challenge, we introduce **CompUGE-Bench**, a unified benchmark designed to evaluate CompQA systems across these four tasks. **CompUGE-Bench** combines datasets from multiple sources to allow fair model comparison. It is a standardized platform to evaluate their CompQA solutions, promoting progress in the field.

Therefore, we formulate the following research questions: *(RQ1) What datasets should be used for creating CompUGE Bench?* and *(RQ2) How can a web-based benchmark be effectively designed for comparative question answering sub-tasks?*

Our contributions are as follows:

- We design a web-based benchmark, making it publicly available for submitting new results.

[1] https://huggingface.co/spaces/uhhlt/CompUGE-Bench

[2] https://youtu.be/rnf6HW1Y7mc

Proceedings of the 31st International Conference on Computational Linguistics: System Demonstrations, pages 189–198
January 19–24, 2025.

- We select and merge datasets for each task, bringing them into a unified structure.
- We conduct extensive experiments, providing baselines for future research.

CompUGE is available as a web application[3,4]; the source code for the benchmark, experiments, and analysis are available under an MIT License.[5].

2 Related Work

In this section, we do not describe the papers describing the datasets we utilize for constructing our benchmark, but focus on other existing approaches for Comparative QA and the adjacent tasks.

One of the significant contributions to the field is the Touché competition series (Bondarenko et al., 2021, 2022b, 2023), organized as part of the Conference and Labs of the Evaluation Forum (CLEF). It focuses on argument retrieval and comparative argumentation, providing a platform for researchers to develop systems capable of retrieving and ranking arguments on diverse topics, including those requiring comparative reasoning.

Schildwächter et al. (2019) investigate methods for answering comparative questions beyond traditional search results, and present CAM — Comparative Argumentative Machine, a specialized system that can handle the nuances of comparative queries. Chekalina et al. (2021) develop a similar system for answering comparative questions, highlighting the importance of handling predicates and aspects in comparisons. Inspired by CAM, Maslova et al. (2023) develop a comparative question system for Russian, while Nikishina et al. (2024) explore the ability of both CAM and RuCAM to process comparative questions in both languages, addressing the challenge of language diversity in user queries.

Regarding Stance Detection task, Kang et al. (2023) explore how LLMs can classify the stance of comparative sentences. By leveraging the vast knowledge and contextual understanding of LLMs, their approach improves the accuracy of preference predictions derived from natural language inputs.

As for answer generation, comparative opinion summarization is a closely related task. Iso et al. (2022) present a method for generating summaries that encapsulate comparative opinions from multiple sources. Their approach focuses on collaborative decoding to produce summaries that highlight key differences and similarities between entities. This work aligns with our answer generation task, as both aim to distill essential comparative information into concise summaries, however, their dataset tackles summaries for each object separately.

Comparative Question Identification

Dataset	Total	Comp	Non-Comp
Webis 2020	14100	1431	13569
Webis 2022	9876	4938	4938
Beloucif et al. (2022)	796	387	409
Mintaka	20000	2000	18000
CompUGE	37684	7565	30119

Object and Aspect Identification

Dataset	Total
Beloucif et al. (2022)	2332
Webis-2022	3530
Chekalina et al. (2021)	3004
CompUGE	5862

Stance Classification

Dataset	Total	Better	Worse	Neutral	None
CompSent-19	7199	1,364	593	-	5242
Webis-2022	950	69	287	324	276
Webis-2022*	144	14	46	-	84
CompUGE	7343	1378	639	-	5326

Table 1: Datasets statistics for each task. Asterisk (*) stands for the dataset after filtration of unavailable sentences with non-disclosure agreements.

3 Tasks and Datasets

Comparative Question Answering involves several interconnected tasks, each requiring specific datasets for training and evaluation. In this section, we delve deeper into the datasets associated with the primary tasks: Comparative Question Identification (CQI), Object and Aspect Identification (OAI), and Stance Classification (SC). We highlight the internal structures of these datasets, the differences among them, and the challenges faced in merging them into a unified benchmark. The statistics for each dataset is presented in Table 1.

3.1 Comparative Question Identification

Comparative Question Identification (CQI) is a binary classification task aiming to determine whether a given question is comparative or not. Figure 2 presents examples of comparative and non-comparative questions. Existing datasets for CQI include *Webis 2020* (Bondarenko et al., 2020), *Webis 2022* (Bondarenko et al., 2022a), *Beloucif*

[3] https://huggingface.co/spaces/uhhlt/CompUGE-Bench
[4] https://youtu.be/rnf6HW1Y7mc
[5] https://github.com/uhh-lt/compuge

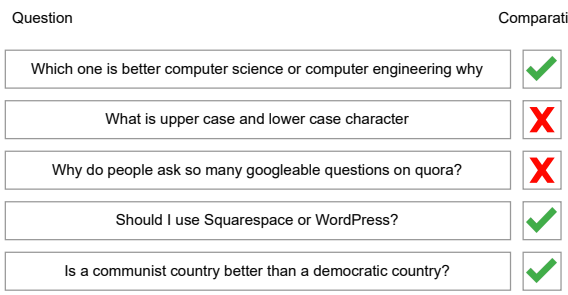


Figure 2: Examples of comparative and non-comparative questions.

Dataset (Beloucif et al., 2022), *Mintaka* (Sen et al., 2022). We describe them in the next paragraphs.

Webis 2020 (Bondarenko et al., 2020): This dataset has significant class imbalance which poses challenges for model training, often requiring techniques like resampling or class weighting to address the skewed distribution. The questions are primarily sourced from web search queries and are short and colloquial. They often lack context and may contain misspellings or abbreviations, reflecting real-world user queries.

Webis 2022 (Bondarenko et al., 2022a): Unlike Webis 2020, the questions in Webis 2022 are more diverse and include additional annotations, such as the objects being compared. The balanced class distribution aids in training models without the need for class balancing techniques. However, there is an overlap of approximately 2,700 questions between Webis 2020 and Webis 2022, which can lead to data leakage if not properly managed.

Beloucif Dataset (Beloucif et al., 2022): Notably, only the test set is publicly available; the training set is not accessible, which complicates direct comparisons with models trained on this dataset. The questions in the Beloucif dataset are carefully curated and may include more complex linguistic structures, making them potentially more challenging for models trained on other datasets.

Mintaka (Sen et al., 2022): The questions are labeled with their respective types and include additional metadata such as language, difficulty level, and domain. Unlike the other datasets, Mintaka is artificially created and designed to cover a wide range of question types. The comparative questions may follow a specific template or structure, which might differ from the more naturally occurring questions in the Webis datasets.

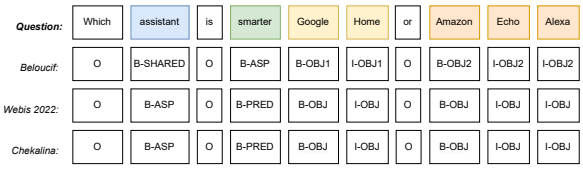


Figure 3: Example of Object and Aspect Identification annotation schemata for different datasets.

Merging these datasets into a unified benchmark for CQI is non-trivial due to the following factors: class imbalance of Webis 2020 dataset; the overlap of around 2,700 questions between Webis 2020 and Webis 2022; different criteria for what constitutes a comparative question (e.g. Webis 2022 classifies sentence like *"what was highest temperature in nigeria ever"* as comparative); style and complexity of the questions (e.g. Mintaka's artificially created questions are much easier to identify).

3.2 Object and Aspect Identification

Object and Aspect Identification (OAI) is a sequence labelling task focused on identifying the objects and aspects (attributes or features) being compared in a question or sentence. Figure 3 illustrates an example of OAI including different annotation schemata that we also describe below.

Beloucif Dataset (Beloucif et al., 2022) introduces four labels for tokens: *OBJECT-1*, *OBJECT-2 ASPECT*, and *SHARED*. The *SHARED* label is used for tokens that are common to both objects being compared. The annotations are in the BIO (Begin-Inside-Outside) format, sentences are tokenized, and each token is assigned a label.

Webis 2022 (Bondarenko et al., 2022a) uses three entity types: *OBJECT*, *ASPECT*, and *PREDICATE*. The *PREDICATE* label represents comparative predicates or verbs that express the comparison. The annotations may not be compatible with the labels used in Beloucif et al. (2022) due to the different roles assigned to tokens.

Chekalina 2021 (Chekalina et al., 2021) focuses on comparative sentences rather than questions, with annotations for *OBJECT*, *ASPECT*, and *PREDICATE* in BIO-format. It contains sentences from comparative texts, and the annotations include longer texts and more complex linguistic structures.

The main challenge in merging these datasets for OAI is the difference in annotation schemata. The use of *SHARED*, *OBJECT-1* and *OBJECT-2* in the Beloucif dataset versus *PREDICATE* and *OBJECT*

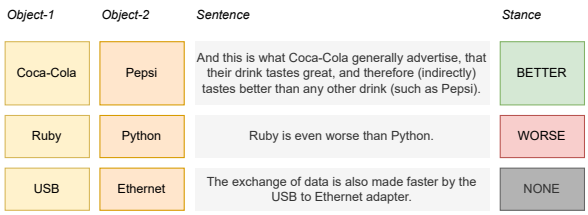


Figure 4: Examples of Stance Classification labels.

in Webis 2022 and Chekalina 2021 creates inconsistency in labels. Another problem is the input type: Beloucif and Webis 2022 focus on questions, while Chekalina 2021 includes sentences from comparative texts. The difference between questions and sentences affects the language structure and the way entities are expressed. Finally, the tokenization differences across datasets can complicate the merging process during the post-processing stage.

3.3 Stance Classification

Stance classification involves determining whether one object is better, worse, or neutral compared to the other in a sentence. Figure 4 provides an example for each label type.

CompSent-19 (Panchenko et al., 2019): contains sentences with *better*, *worse*, or *neutral* labels for each sentence. The dataset focuses on comparative sentences extracted from web, and the objects are explicitly identified as separate columns.

Webis 2022 (Bondarenko et al., 2022a): introduces an additional class, making it a four-class classification problem (*better*, *worse*, *neutral*, *none*). The sentences in this dataset may be longer and more complex, and 806 entries had to be discarded due to non-disclosure agreements or excessive length, resulting in 144 sentences left.

The main challenges in merging these datasets are the inconsistent labels and a large difference between sentence lengths (97 tokens for CompSent-19 and 1624 tokens for Webis 2022).

3.4 Answer Generation

Answer Generation is the task of generating a concise summary or answer that compares two objects based on a set of comparative sentences. Figure 5 illustrates this task with the example from (Chekalina et al., 2021), which includes a human-written answer from Yahoo!Answers[6]. As we have only one dataset for this task, merging datasets is not applicable.

[6] https://en.wikipedia.org/wiki/Yahoo_Answers

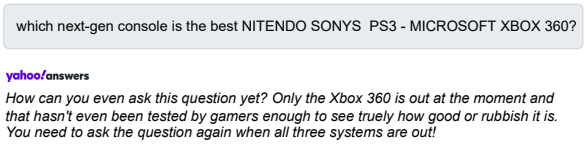


Figure 5: Example of a comparative answer from Chekalina et al. (2021).

4 CompUGE Datasets Creation

For each task, existing datasets are brought into a unified structure. The datasets are merged using all possible permutations to create comprehensive training sets. Then, we then train four Transformer Encoder models on each dataset combination: **DistilBERT-base-uncased fine-tuned on English** (Sanh et al., 2019), **DistilBERT-base-uncased** (Sanh et al., 2019), **RoBERTa-base** (Liu et al., 2019), **DeBERTa-base** (He et al., 2020). These models were chosen for their balance of performance and computational efficiency in (Shallouf et al., 2024). They also represent a variety of Encoder architectures and sizes, which is beneficial for assessing model robustness across tasks.

Each model is then tested on every test set for that task. All model predictions alongside key metrics (accuracy, precision, recall, F1-score) are stored. We averaged the metrics between all four models. Finally, we analyzed the resulting metrics and prediction files alongside the structures of the datasets to select the best dataset combinations.

4.1 Comparative Question Identification

We do seven permutations for training (three individual datasets, three pairwise merges, and one merge of all datasets) and tested on all four datasets (including Beloucif's test set).

Results Table 2 shows the averaged accuracy across models when trained on different dataset combinations and tested on each dataset. When testing on Beloucif et al. (2022), the best performance is achieved by training on all datasets combined, yielding an average accuracy of 0.8, and it is clearly visible that Beloucif et al. (2022) is the most challenging one. Table 3 provides detailed metrics for models tested on this dataset. Based on these results, we decide to merge Mintaka, Webis 2020, and Webis 2022 for training and use Beloucif for testing in the benchmark.

Training Data	Beloucif	Mintaka	Webis 20	Webis 22
All Datasets	0.80	0.99	0.97	0.97
Mintaka	0.63	0.99	0.89	0.68
Mintaka + Webis 20	0.70	0.99	0.97	0.93
Mintaka + Webis 22	0.74	0.99	0.94	0.97
Webis 20	0.72	0.75	0.97	0.88
Webis 20 + Webis 22	0.74	0.97	0.97	0.98
Webis 22	0.72	0.96	0.94	0.97

Table 2: Averaged accuracy across models for CQI.

Training Data	Accuracy	Precision	Recall	F1-Score
All Datasets	0.80	0.85	0.80	0.79
Mintaka	0.63	0.66	0.63	0.60
Mintaka + Webis 20	0.70	0.73	0.70	0.68
Mintaka + Webis 22	0.74	0.82	0.74	0.73
Webis 20	0.72	0.74	0.72	0.71
Webis 20 + Webis 22	0.74	0.82	0.74	0.73
Webis 22	0.72	0.81	0.72	0.70

Table 3: Averaged metrics for models tested on Beloucif et al. (2022).

4.2 Object and Aspect Identification

We conduct experiments with 7 dataset combinations for training and test on all three datasets.

Results Table 4 presents the averaged F1-scores across models when trained on different dataset combinations and tested on each dataset. When excluding Chekalina et al. (2021) from training and testing, we observe better alignment between Webis 2022 and Beloucif datasets. Table 6 in Appendix A shows the averaged F1-scores without training or testing on Chekalina et al. (2021) alone. Based on these observations, we decide to merge the processed version of Webis 2022 with Beloucif and exclude Chekalina 2021 from the main OAI benchmark. The processed version of Webis 2022 relabels all *PREDICATE* entities to *ASPECT*, and in Beloucif, we removed sentences containing the *SHARED* label.

4.3 Stance Classification

For this task, we use two datasets (CompSent-19 and the processed version of Webis 2022) and one merged dataset, resulting in three training permutations. The Webis 2022 dataset required significant preprocessing: entries with non-disclosure agreements were removed, extremely long sentences were discarded, and the four classes were reduced to three by merging *NO-STANCE* and *NEUTRAL* into *NEUTRAL* resulting in 144 sentences.

Results Table 5 shows the averaged F1-scores across models when trained on different dataset combinations and tested on each dataset. Training on the merged dataset improved performance on Webis 2022, increasing the F1-score from 0.42 to 0.53. However, the performance on CompSent-19 remained high regardless of whether Webis 2022 was included in training. Based on these results, we merge CompSent-19 with the processed version of Webis 2022 for the benchmark.

Training Data	All	Beloucif	Webis	Chekalina
All Datasets	0.81	0.76	0.82	0.83
Chekalina + Webis	0.75	0.59	0.80	0.84
Webis + Beloucif	0.68	0.76	0.82	0.46
Chekalina + Beloucif	0.65	0.75	0.36	0.84
Webis	0.63	0.61	0.81	0.47
Beloucif	0.51	0.76	0.35	0.45
Chekalina	0.49	0.30	0.21	0.84

Table 4: Averaged F1-scores across models for OAI.

Training Data	CompSent-19	Webis 2022
All Datasets	0.89	0.53
CompSent-19	0.89	0.42
Webis 2022	0.42	0.36

Table 5: Averaged F1-scores for Stance Classification.

5 System Design and Architecture

The CompUGE benchmark system is designed with a modular architecture, consisting of three main components.

PostgreSQL Database serves as the central repository for storing datasets, model submissions, evaluation results, and leaderboards. It ensures data integrity and supports concurrent access by multiple users.

FastAPI Backend Server acts as the intermediary between the frontend and the database. It handles API requests from the frontend, processes data, runs evaluation scripts, and communicates with the database. The backend is built using FastAPI[7], a modern, high-performance web framework for building APIs with Python.

Angular Frontend provides an interactive web interface for users to interact with the benchmark. Users can explore available tasks and datasets, submit their model results for evaluation, and view leaderboards. The frontend is developed using Angular[8], a popular web application framework.

Each component is containerized using Docker for ease of deployment and scalability. The database and backend server are deployed on an

[7] https://fastapi.tiangolo.com
[8] https://angular.io

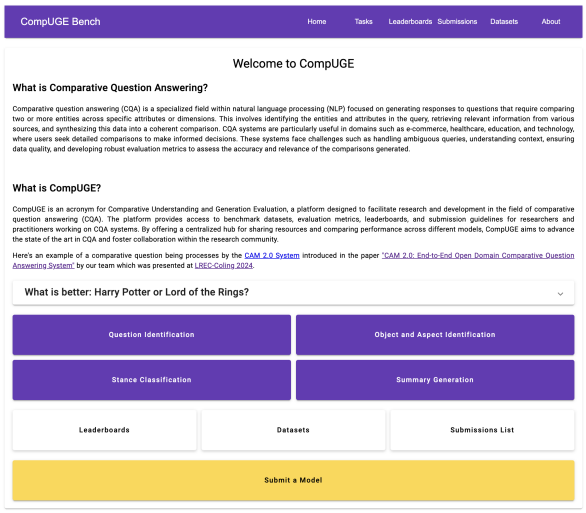


Figure 6: CompUGE Benchmark Start Page.

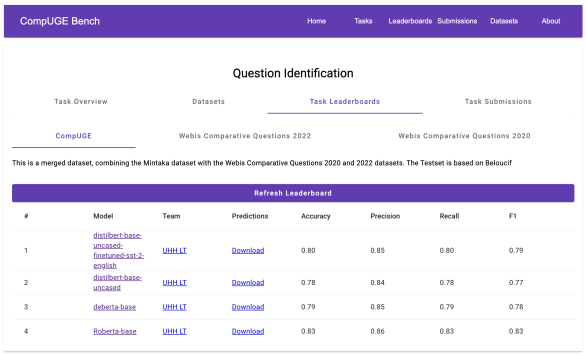


Figure 7: Example of the leaderboard display for the CQI task.

internal server, ensuring data security and compliance with institutional policies. The frontend is hosted on HuggingFace Spaces[9], making the benchmark easily accessible to the research community.

5.1 Modular Design

The system's modular design allows for easy expansion and maintenance. Key features include:

- **Expandable List of Tasks**: More Comparative QA tasks can be added to the system without affecting existing functionalities.
- **Datasets Association**: Each task can have multiple associated datasets, all adhering to the same data format.
- **Leaderboards**: For each task and dataset combination, there is a corresponding leaderboard that tracks model performances using standardized metrics.

[9] https://huggingface.co/spaces

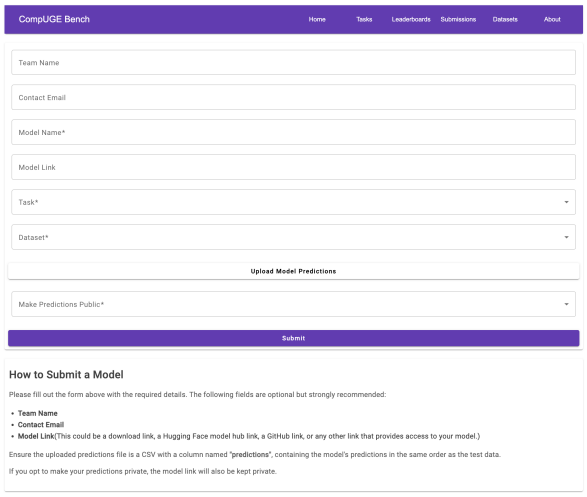


Figure 8: New Submission interface, with a simple guide on how to submit a model to CompUGE

All datasets for a given task share the same data format, and their leaderboards use consistent evaluation metrics. This design choice simplifies the process for researchers to benchmark their models across different datasets and tasks. Figure 6 demonstrates the start page of the benchmark, while 7 showcases a leaderboard for one of the tasks.

The submission page in Figure 8 allows users to submit the results of their models for evaluation. The form requires users to provide key details: team, contact email, model name, etc. The model link could point to a download link, a Hugging Face model hub, a GitHub repository, or any other online location that provides access to the model. The model's predictions are updated as a CSV file, ensuring the predictions are in the same order as the test data and that the file contains a column named "predictions". Users are also asked whether they want to make their predictions public. If users choose to keep their predictions private, the model link will also be kept confidential. After filling in the necessary details, the submission can be finalized by clicking the Submit button. More screenshots can be seen in Appendix B.

6 Conclusion

CompUGE provides a structured and comprehensive benchmark for evaluating comparative question answering systems. By integrating datasets from multiple sources and evaluating models across distinct sub-tasks, it offers a robust platform for future research. The benchmark is available on Hugging Face Spaces, and its source code is open-sourced under the MIT License.

Limitations

The main limitations of the paper are as follows:

- All our experiments with different Comparative Question Answering tasks are done using Encoder Transformer models. We do not run experiments using LLMs, as more time- and resource-consuming models. Our main idea was to provide baselines and select the best datasets for the benchmark, not to test all existing models. However, we leave testing Generative Transformer models for future work.
- Due to limited resources available, our benchmark allows only result file upload to the server. This may lead to unfair and non-reproducible results. We try to approach it by asking the user to provide the path or the name of the used model, leaving server model evaluation for future work.

Ethical Statement

This work was conducted in compliance with ethical standards in AI research. All datasets used in the study are publicly available, and no private data was utilized. The benchmark is designed to support reproducible research and transparent model evaluation.

References

Meriem Beloucif, Seid Muhie Yimam, Sarah Stahlhacke, and Chris Biemann. 2022. Classification and identification of elements in comparative questions. In *Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC 2022)*, pages 3771–3779.

Alexander Bondarenko, Yamen Ajjour, Vivien Dittmar, Niklas Homann, Pavel Braslavski, and Matthias Hagen. 2022a. Towards understanding and answering comparative questions. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining (WSDM 2022)*, pages 66–74.

Alexander Bondarenko, Pavel Braslavski, Michael Völske, Robin Aly, Maik Fröbe, Alexander Panchenko, Chris Biemann, Benno Stein, and Matthias Hagen. 2020. Comparative web search questions. In *Proceedings of the 13th International Conference on Web Search and Data Mining (WSDM 2020)*, pages 52–60.

Alexander Bondarenko, Maik Fröbe, Johannes Kiesel, Ferdinand Schlatt, Valentin Barriere, Brian Ravenet, Léo Hemamou, Simon Luck, Jan Heinrich Reimer, Benno Stein, Martin Potthast, and Matthias Hagen. 2023. Overview of Touché 2023: Argument and Causal Retrieval. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 14th International Conference of the CLEF Association (CLEF 2023)*, volume 14163 of *Lecture Notes in Computer Science*, pages 507–530, Berlin Heidelberg New York. Springer.

Alexander Bondarenko, Maik Fröbe, Johannes Kiesel, Shahbaz Syed, Timon Gurcke, Meriem Beloucif, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. 2022b. Overview of Touché 2022: Argument Retrieval. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 13th International Conference of the CLEF Association (CLEF 2022)*, volume 13390 of *Lecture Notes in Computer Science*, Berlin Heidelberg New York. Springer.

Alexander Bondarenko, Lukas Gienapp, Maik Fröbe, Meriem Beloucif, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. 2021. Overview of Touché 2021: Argument Retrieval. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 12th International Conference of the CLEF Association (CLEF 2021)*, volume 12880 of *Lecture Notes in Computer Science*, pages 450–467, Berlin Heidelberg New York. Springer.

Valeriya Chekalina, Alexander Bondarenko, Chris Biemann, Meriem Beloucif, Varvara Logacheva, and Alexander Panchenko. 2021. Which is better for deep learning: Python or matlab? answering comparative questions in natural language. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (EACL 2021)*, pages 302–311.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Hayate Iso, Xiaolan Wang, Stefanos Angelidis, and Yoshihiko Suhara. 2022. Comparative opinion summarization via collaborative decoding. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3307–3324, Dublin, Ireland. Association for Computational Linguistics.

Inwon Kang, Sikai Ruan, Tyler Ho, Jui-Chien Lin, Farhad Mohsin, Oshani Seneviratne, and Lirong Xia. 2023. Llm-augmented preference learning from natural language. *CoRR*, abs/2310.08523.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Maria Maslova, Stefan Rebrikov, Anton Artsishevski, Sebastian Zaczek, Chris Biemann, and Irina Nikishina. 2023. Rucam: Comparative argumentative machine for the russian language. In *AIST*, volume 14486 of *Lecture Notes in Computer Science*, pages 78–91. Springer.

Irina Nikishina, Alexander Bondarenko, Sebastian Zaczek, Onno Lander Haag, Matthias Hagen, and Chris Biemann. 2024. Extending the comparative argumentative machine: Multilingualism and stance detection. In *RATIO*, volume 14638 of *Lecture Notes in Computer Science*, pages 317–334. Springer.

Alexander Panchenko, Alexander Bondarenko, Marc Franzek, Matthias Hagen, and Chris Biemann. 2019. Categorizing comparative sentences. In *Proceedings of the 6th Workshop on Argument Mining (ArgMining 2019)*, pages 136–145.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Maike Schildwächter, Alexander Bondarenko, Juliane Zenker, Matthias Hagen, Chris Biemann, and Alexander Panchenko. 2019. Answering comparative questions: Better than ten-blue-links? In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, pages 361–365.

Parth Sen, Alham Fikri Aji, and Amir Saffari. 2022. Mintaka: A complex, natural, and multilingual dataset for end-to-end question answering. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING 2022)*, pages 1604–1619.

Ahmad Shallouf, Irina Nikishina, and Chris Biemann. 2024. Cam 2.0: End-to-end open domain comparative question answering system. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.

A Additional Results

Training Data	Tested on	Overall	Object	Aspect
Webis	Webis	0.84	0.82	0.86
All Datasets	Webis	0.84	0.83	0.85
Webis + Beloucif	Webis	0.84	0.82	0.86
Webis + Beloucif	Beloucif	0.79	0.84	0.60
All Datasets	Beloucif	0.78	0.85	0.55
Beloucif	Beloucif	0.77	0.83	0.53

Table 6: Averaged F1-scores for Webis and Beloucif combinations as well as all datasets.

Training Data	Overall	Object	Aspect
Chekalina	0.86	0.92	0.79
Chekalina + Webis	0.86	0.91	0.79
Chekalina + Beloucif	0.86	0.91	0.80
All Datasets	0.85	0.89	0.80
Webis	0.50	0.45	0.56
Webis + Beloucif	0.46	0.41	0.53
Beloucif	0.40	0.49	0.25

Table 7: Averaged F1-scores tested on Chekalina et al. (2021).

B CompUGE Benchmark Details

We include several screenshots of the CompUGE benchmark system to illustrate its user interface and functionalities. Figures 9 to 11 showcase different parts of the system.

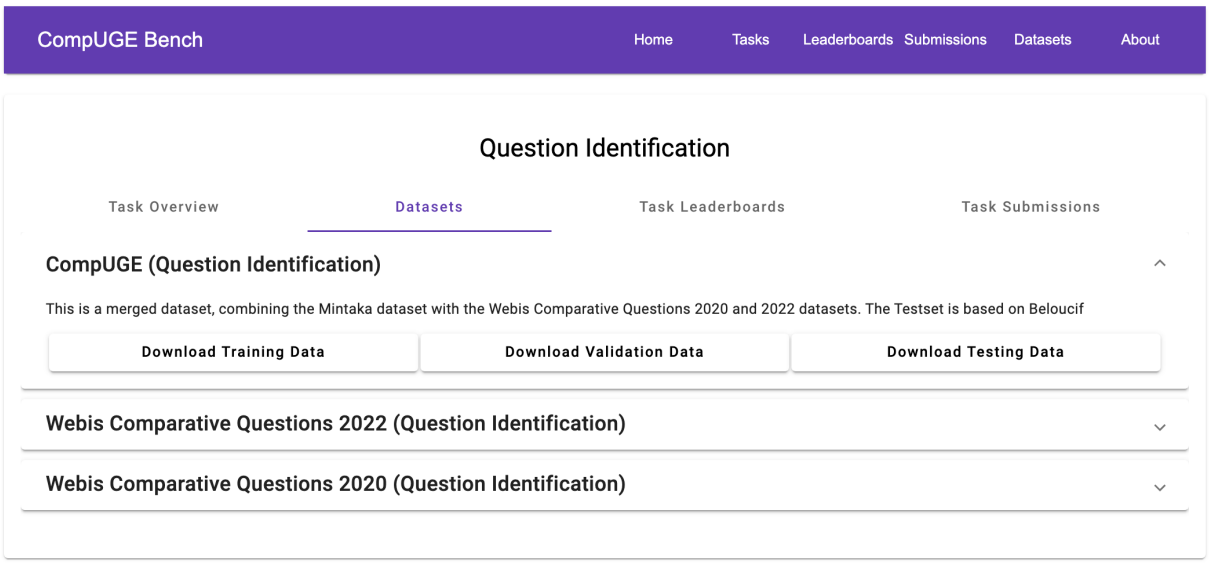


Figure 9: The datasets tab in a Task Page (Question Identification), provides drop down for each dataset, containing description and download buttons for splits

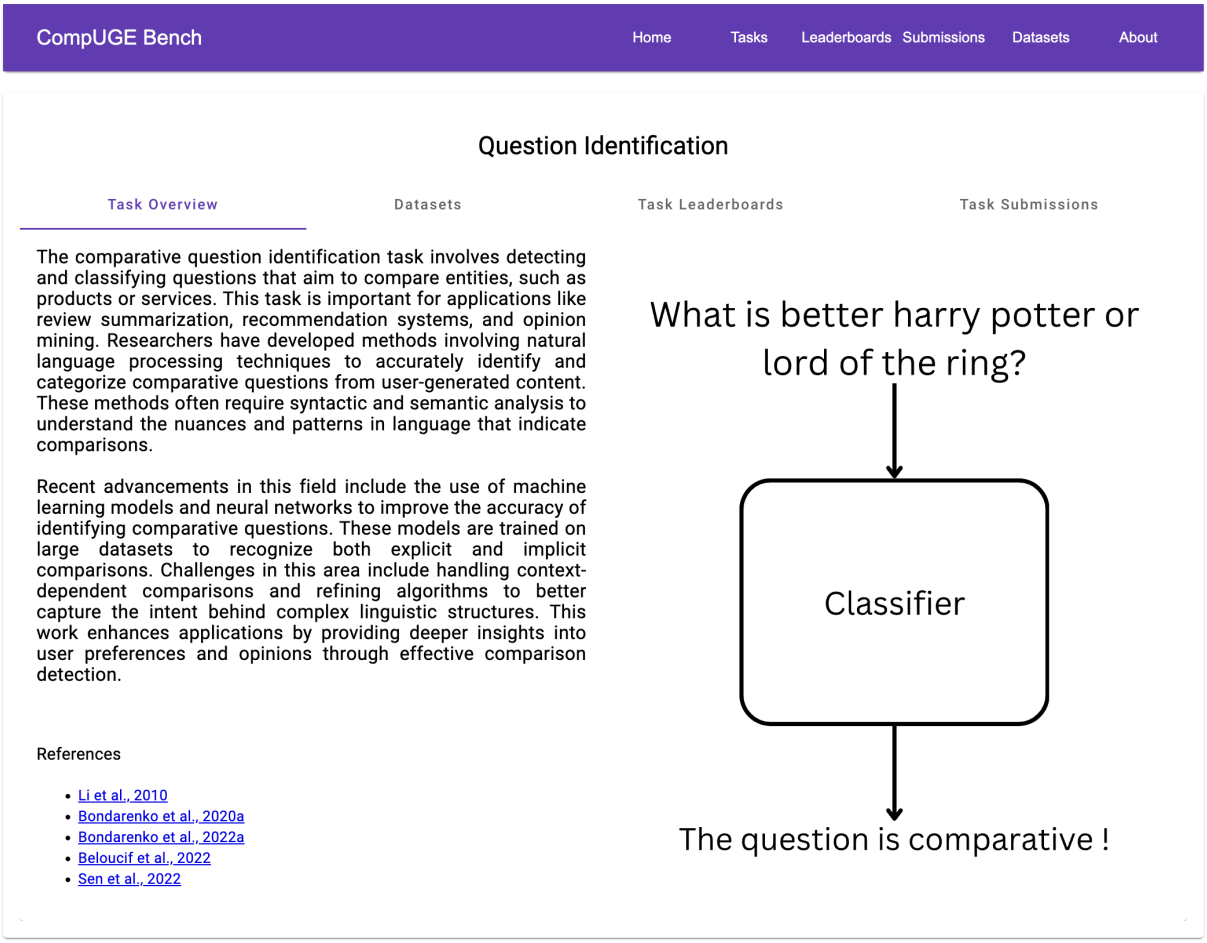
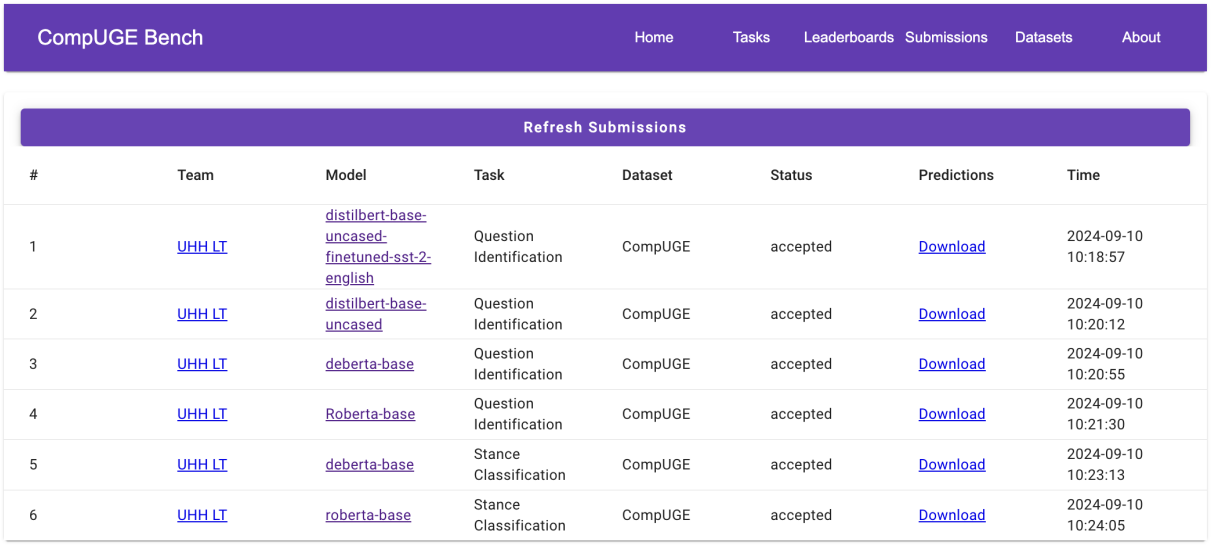


Figure 10: Task Page, which provides access to an overview of the task, Datasets, Task specific leaderboards and Task specific Submissions



#	Team	Model	Task	Dataset	Status	Predictions	Time
1	UHH LT	distilbert-base-uncased-finetuned-sst-2-english	Question Identification	CompUGE	accepted	Download	2024-09-10 10:18:57
2	UHH LT	distilbert-base-uncased	Question Identification	CompUGE	accepted	Download	2024-09-10 10:20:12
3	UHH LT	deberta-base	Question Identification	CompUGE	accepted	Download	2024-09-10 10:20:55
4	UHH LT	Roberta-base	Question Identification	CompUGE	accepted	Download	2024-09-10 10:21:30
5	UHH LT	deberta-base	Stance Classification	CompUGE	accepted	Download	2024-09-10 10:23:13
6	UHH LT	roberta-base	Stance Classification	CompUGE	accepted	Download	2024-09-10 10:24:05

Figure 11: Overall Submissions list, which provides information on weather a submission was accepted, and for public submissions gives access to submitted predictions, contact email of submitter and model link