Human-Centric Evaluation Framework for Comparing Large Language Models

Darius Hennekeuser*

Bonn-Rhein-Sieg University of Applied Sciences, darius.hennekeuser@h-brs.de

Erik Dethier

Bonn-Rhein-Sieg University of Applied Sciences, erik.dethier@h-brs.de

Daryoush Vaziri

Bonn-Rhein-Sieg University of Applied Sciences, Daryoush.vaziri@h-brs.de

Gunnar Stevens

University of Siegen, gunnar.stevens@uni-siegen.de

This demonstration presents a human-centered evaluation tool that compares different Large Language Models (LLMs) based on their performance in individual user tasks. Unlike existing automatic evaluation methods, which can lack context-specific accuracy, our tool aims to emphasize user experience and practical applicability. Developed from preliminary participant observations, the tool allows users to input prompts and receive responses from multiple LLMs simultaneously, which they can then evaluate using a Likert scale and comments. This approach ensures unbiased and task-relevant assessments, enhancing the selection process for suitable LLMs. Future improvements will focus on user-configurable settings. This work aims to bridge the gap between general evaluation metrics and the unique needs of individual users, providing a more nuanced and practical evaluation of LLMs.

CCS CONCEPTS • **Human-centered computing** → human computer interaction (HCI) • **Computing methodologies** → Artificial intelligence → Natural language processing

Additional Keywords and Phrases: Large Language Models, LLM, Evaluation, Natural Language Processing, NLP

1 INTRODUCTION

Evaluating the performance and accuracy of Large Language Models (LLMs) is a crucial task in computer science to verify and ensure their applicability in a given context [1].

In recent years, the excitement surrounding LLMs due to the latest breakthroughs in artificial intelligence (AI) has resulted in a proliferation of LLMs in various languages and based on different training corpora. However, factors such as language, knowledge base, accurate user understanding, and many other characteristics are crucial for determining the suitability of natural language processing (NLP) systems, such as LLMs, for specific application contexts [2].

* Corresponding author

In both practice[1] and literature, there are numerous evaluation methods that provide automatic assessments performed by machines. These include using LLMs to evaluate other LLMs or conducting multiple-choice and single-choice tests [1, 3–7]. These tests are useful for evaluating the basic capabilities of language models and making them fundamentally comparable [1, 8]. However, the application context in a user's reality is so unique that these evaluation metrics are too general to provide accurate assessments for different use cases such as medical consultation [9], education [10], justice [11], and private household [12]. For instance, the intensively used MTBench dataset covers eight common categories of user requirements: writing, role-playing, extraction, reasoning, mathematics, coding, knowledge I (STEM), and knowledge II (humanities/social sciences) [6]. While the data set evaluates the capabilities of language models in various categories, there is still no concrete reference to the individual tasks of individual users, which should be decisive for the selection of the language model. Here, human-computer interaction (HCI) should advocate a human-centered approach that supports an evaluation in which the user is able to make an assessment himself.

In addition, the predominant evaluation metrics are mostly developed for the English language and are merely translated into other languages, which neglects a good evaluation of the capabilities in other languages [3, 4, 6, 13–15].

There are existing tools that involve user evaluations, such as the 'chatbot arena' from LMSYS. However, these tools compare models anonymously and rely on user feedback to generate a general leaderboard based on human evaluations. This approach does not benefit users with their specific tasks but instead provides feedback for overall model rankings[2].

Therefore, this paper demonstrates a prototype of an evaluation tool we developed for human assessment of LLMs in specific application contexts. The tool was created based on several participant observations conducted with users testing LLM-based applications for their use cases. During these observations, we noticed that users often relied on self-created Excel tools for their evaluations. Here, users collect their results from different models one after the other in a table in which the results of the models are compared and evaluated.

The following prototype of our tool described below has already been tested by users in several workshops.

2 THE EVALUATION INTERFACE OF THE TOOL

As observed, users commonly use tables to assess LLM responses to their prompts. Our demonstrated system's user interface (UI) is designed to assist users in evaluating LLMs for specific tasks. The UI is supposed to aid in decision-making by providing comprehensive support for selecting the most suitable LLM to integrate into the user's task, thereby enhancing the overall user experience. Our proposed UI based on preliminary results from workshops and user tests can be found in Figure 1.

Our system's UI is composed of four main components: an input mask (1), multiple chat modules (2), the task title/description (3) and an evaluation section (4) for the LLMs of each chat module.

The input mask allows users to send task-specific prompts. These prompts are simultaneously sent to multiple LLMs (currently three), with the responses logged into the respective chat modules. This concurrent approach enables users to compare responses from different LLMs side-by-side, rather than sequentially.

To prevent bias, the order of the LLMs in the chat modules is randomized, and their names are concealed until after the user has evaluated the responses. This ensures unbiased evaluations, free from the influence of brand recognition or recurring order patterns.

[1] https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard
[2] https://chat.lmsys.org/

At the top of the UI, users can set a title and a description for the task they want to evaluate using the LLMs. Both the title and the description must be defined before sending each prompt, as they are essential for the evaluation report described in the following section 3.

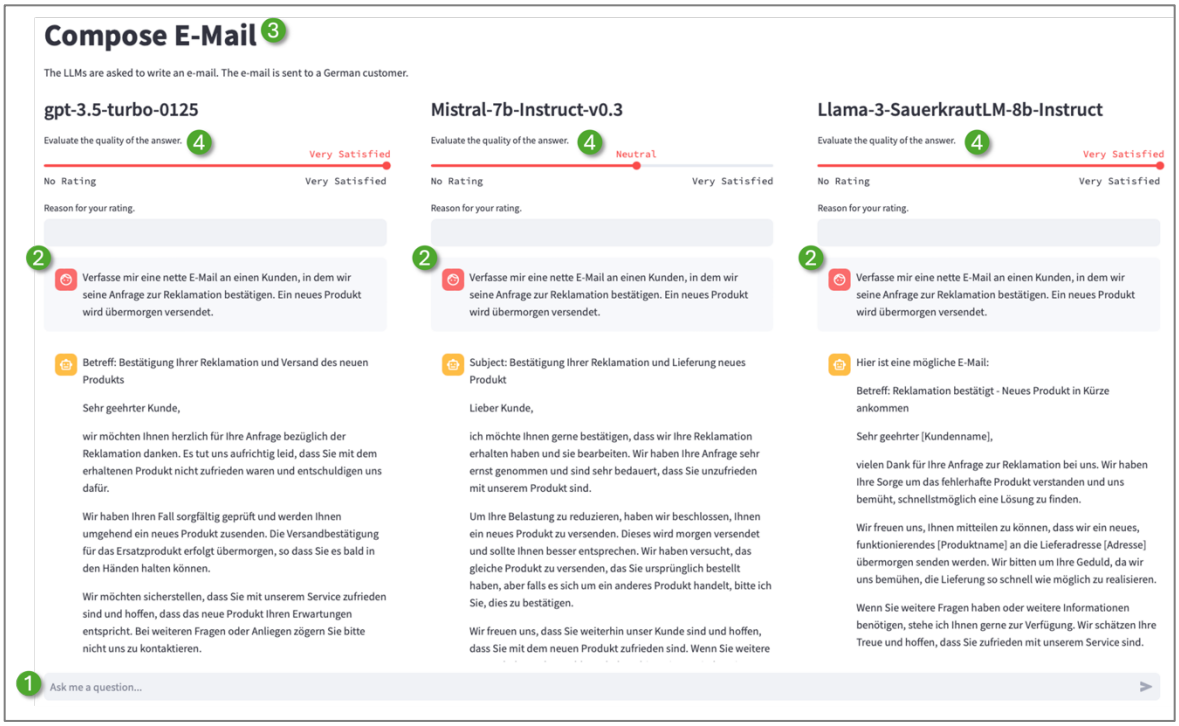

Figure 1: Evaluation interface of the system (screenshot)

Each LLM response can be evaluated as soon as it appears. Users can rate the answers using a slide control with a five-point Likert scale: ‘Very Dissatisfied’, ‘Dissatisfied’, ‘Neutral’, ‘Satisfied’, and ‘Very Satisfied’. Additionally, users can provide comments explaining their ratings and reasoning. These evaluations are logged for each prompt and response and are utilized in the evaluation report described in the following section.

Currently, the types of LLMs and the number of chat modules (and therefore the number of LLMs to be evaluated) can only be adjusted in the system’s backend, not through the UI.

3 THE EVALUATION REPORT FROM THE TOOL

In this current prototype, the evaluation report still has to be generated manually from the backend by the administrator.

As shown in Figure 2, the report first summarizes all evaluated tasks and their numerical evaluation, which was recorded using the five-point Likert scale, in a simple line diagram (A). This allows the user to see at a glance which model performs best in the general trend, and for which tasks it is stronger or weaker.

After this, a table is provided for each task performed (B), showing the numerical rating (β), the comment of the user (γ) and the answer of the model (δ) below the task title and the original prompt (α) for each model used. This format allows

the user to obtain an aggregated, numerical comparison, but also to compare the answers in detail. With the table providing the answer comparison in detail, we stay close to the excel tools we have observed in the field.

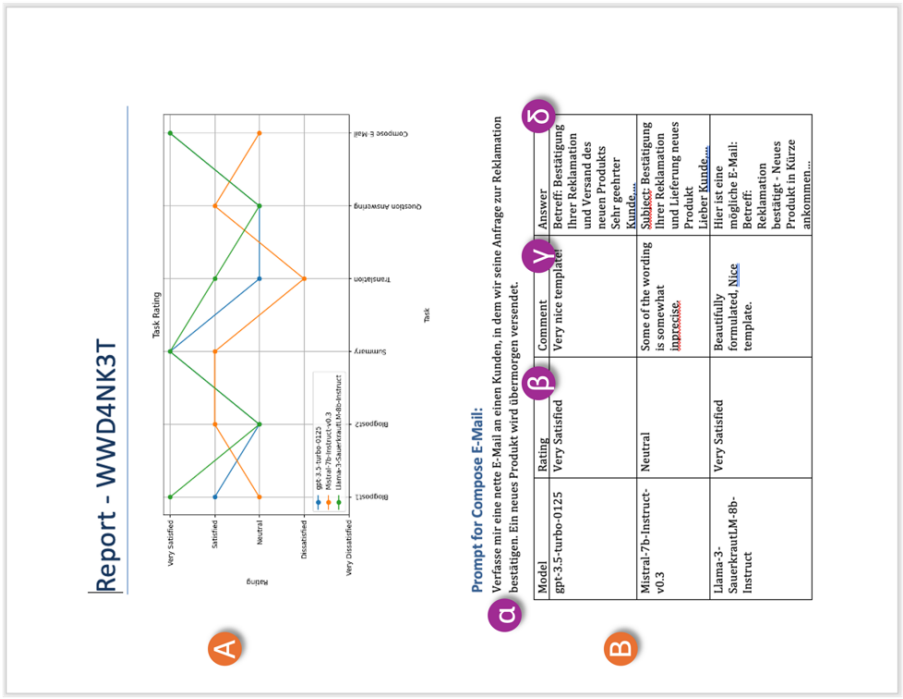

Figure 2: evaluation report (screenshot)

4 FUTURE WORK

We have demonstrated how a user-centered evaluation tool for testing different language models could be designed based on preliminary results. However, the current prototype of this tool requires manual adjustments in the backend to configure aspects such as the integrated LLM, the number of models compared, and the model type. A production system should ideally allow users to configure these settings themselves, including the selection and number of models.

Practically, several measures should be taken to improve the overall user experience and capabilities of the system. Users should be able to work with external materials by employing techniques such as retrieval augmented generation [16], having an internet connection, and using function calling (which allows the LLM to execute predefined functions). These enhancements will enable more realistic testing for users.

Although we developed the current prototype based on field observations with users and successfully tested it in several workshops, future work should extensively investigate user interactions with such a system. It is also crucial to identify additional evaluation factors that are relevant to users.

ACKNOWLEDGMENTS

We thank the people we were able to observe in their evaluation work of various language models in the context of studies. We are grateful for the funding from the Federal Ministry of Economics and Climate Protection of the Federal Republic of Germany.

REFERENCES

1. Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P.S., Yang, Q., Xie, X.: A Survey on Evaluation of Large Language Models. ACM Trans. Intell. Syst. Technol. 15, 1–45 (2024). https://doi.org/10.1145/3641289
2. Hennekeuser, D., Vaziri, D., Golchinfar, D., Stevens, G.: What I Don't Like about You?: A Systematic Review of Impeding Aspects for the Usage of Conversational Agents. Interacting with Computers. iwae018 (2024). https://doi.org/10.1093/iwc/iwae018
3. Sakaguchi, K., Bras, R.L., Bhagavatula, C., Choi, Y.: WinoGrande: an adversarial winograd schema challenge at scale. Commun. ACM. 64, 99–106 (2021). https://doi.org/10.1145/3474381
4. Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., Choi, Y.: HellaSwag: Can a Machine Really Finish Your Sentence? In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 4791–4800. Association for Computational Linguistics, Florence, Italy (2019)
5. Du, X., Liu, M., Wang, K., Wang, H., Liu, J., Chen, Y., Feng, J., Sha, C., Peng, X., Lou, Y.: Evaluating Large Language Models in Class-Level Code Generation. In: Proceedings of the IEEE/ACM 46th International Conference on Software Engineering. pp. 1–13. ACM, Lisbon Portugal (2024)
6. Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Zhang, H., Gonzalez, J.E., Stoica, I.: Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, https://arxiv.org/abs/2306.05685, (2023)
7. Chiang, C.-H., Lee, H.: Can Large Language Models Be an Alternative to Human Evaluations? In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 15607–15631. Association for Computational Linguistics, Toronto, Canada (2023)
8. Noguer I Alonso, M.: A Framework for the Evaluation of Large Language Models. SSRN Journal. (2023). https://doi.org/10.2139/ssrn.4649866
9. Ayers, J.W., Poliak, A., Dredze, M., Leas, E.C., Zhu, Z., Kelley, J.B., Faix, D.J., Goodman, A.M., Longhurst, C.A., Hogarth, M., Smith, D.M.: Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. JAMA Intern Med. 183, 589 (2023). https://doi.org/10.1001/jamainternmed.2023.1838
10. AlAfnan, M.A., Samira Dishari, Marina Jovic, Koba Lomidze: ChatGPT as an Educational Tool: Opportunities, Challenges, and Recommendations for Communication, Business Writing, and Composition Courses. JAIT. (2023). https://doi.org/10.37965/jait.2023.0184
11. Nay, J.J., Karamardian, D., Lawsky, S.B., Tao, W., Bhat, M., Jain, R., Lee, A.T., Choi, J.H., Kasai, J.: Large language models as tax attorneys: a case study in legal capabilities emergence. Phil. Trans. R. Soc. A. 382, 20230159 (2024). https://doi.org/10.1098/rsta.2023.0159
12. King, E., Yu, H., Lee, S., Julien, C.: Sasha: Creative Goal-Oriented Reasoning in Smart Homes with Large Language Models. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 8, 1–38 (2024). https://doi.org/10.1145/3643505
13. Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., Tafjord, O.: Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge, https://arxiv.org/abs/1803.05457, (2018)
14. Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., Steinhardt, J.: Measuring Massive Multitask Language Understanding, https://arxiv.org/abs/2009.03300, (2020)
15. Lin, S., Hilton, J., Evans, O.: TruthfulQA: Measuring How Models Mimic Human Falsehoods. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 3214–3252. Association for Computational Linguistics, Dublin, Ireland (2022)
16. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., and Lin, H. (eds.) Advances in Neural Information Processing Systems. pp. 9459–9474. Curran Associates, Inc. (2020)