ChatBS: An Exploratory Sandbox for Bridging Large Language Models with the Open Web

John S. Erickson[1], Henrique Santos[1,*], Jamie McCusker[1], Sola Shirai[1], Deborah L. McGuinness[1] and James Hendler[1]

[1]*Tetherless World Constellation, Rensselaer Polytechnic Institute, Troy, NY, 12180, United States*

Abstract

The recent widespread public availability of generative large language models (LLMs) has drawn much attention from the academic community to run experiments in order to learn more about their strengths and drawbacks. From prompt engineering and fine-tuning to fact-checking and task-solving, researchers have pursued several approaches to try to take advantage of these tools. As some of the most powerful LLMs are "closed" and only accessible through web APIs with prior authorization, combining LLMs with the open web is still a challenge. In this evolving landscape, tools that can facilitate the exploration of the capabilities and limitations of LLMs are desirable, especially when connecting with traditional web features such as search and structured data. This article presents ChatBS, a web-based exploratory sandbox for LLMs, working as a front-end for prompting LLMs with user inputs. It provides features such as entity resolution from open knowledge graphs, web search using LLM outputs, as well as popular prompting techniques (e.g. multiple submissions, "step-by-step"). ChatBS has been extensively used in Rensselaer Polytechnic Institute's Data INCITE courses and research, serving as key tool for utilizing LLMs outputs at scale in these contexts.

Submission Type: Demo.
Demo URL: https://inciteprojects.idea.rpi.edu/chatbs/app/chatbs/

1. Introduction

Generative large language models (LLMs) have become pervasive in recent years [1], permitting free access to virtually any web user. They are inarguably revolutionary in their ability to generate human-like text with unprecedented scale and speed. Largely led by the industry, these LLMs are trained using vast amounts of web-scraped data, and little is known about exactly what data and how it is used in their training. This proliferation has ignited considerable interest within the academic community, prompting researchers to conduct experiments aimed at exploring the strengths and limitations of these powerful tools (e.g. morality [2], kbqa [3]). However, a key challenge arises due to the fact that some of the most formidable LLMs are considered "closed" and are only accessible through web APIs with prior authorization. This presents an obstacle to effectively combining content from LLMs with the open web, an issue that remains a central concern in the evolving landscape of web science. The development of tools that can facilitate the exploration of the capabilities and limitations of LLMs, especially

when integrated with traditional web features like search and structured data, has become increasingly desirable.

ChatBS serves as a front-end for prompting LLMs with user inputs, offering functionalities to support experimentation and research. On the input side, ChatBS allows users to quickly incorporate popular prompting techniques, such as making multiple submissions and employing a "step-by-step" approach. On the output side, ChatBS includes a web search using LLM-generated content and entity resolution against open knowledge graphs. We provide an overview of ChatBS, its functionalities, and its impacts, emphasizing its potential to bridge the gap between LLMs and the open web.

2. ChatBS: An Exploratory Sandbox

ChatBS was conceived as an accessible, extensible platform for interactive experimentation with LLMs such as, initially, GPT 3.5 and, more recently, GPT 4. It was designed to empower users to easily conduct sophisticated prompt experiments that might be difficult or impossible through the normal ChatGPT user interface, and would normally be done by directly accessing an API (esp. the OpenAI API) programmatically. ChatBS features include:

- LLM Selection: Users can easily select between models from the GTP-4 and GPT-4o families. In the future, OpenAI and other models will be added or dropped depending on availability. We plan to diversify model availability by adding openly available models, such as the recently-released Llama 3.1 model [4], as options.
- System Prompt Customization: To support robust prompt experimentation allowing users to pre-define the context, scope, guardrails, or output format for the model to use during an interaction.
- User Prompt Customization.
- Chain-of-Thought Facilitation: Standardized prompt modifications to request models to *explain step-by-step.*
- Explanation enhancement: Standardized user prompt append to request references to support generated content.
- Repeated Prompt Submission: To support scalability in the generation of datasets with multiple responses to the same standardized prompt.
- Relationship Detection, Entity Resolution, and RDF Generation: ChatBS parses and links the generated answers to entities in Wikidata, so users can contrast claims.
- Dataset generation: To document experiments and to make the results more shareable and replicable, both the main ChatBS results and the RDF results (if requested) are downloadable as JSON(-LD) files.

ChatBS was initially conceived as a compelling and fun way to fact-check LLM-generated answers to questions. The earliest versions of ChatBS (2023) demonstrated the ability to identify asserted relationships between entities within the answers; to resolve entities and relationships (predicates) to URIs; and to express these result graphs as RDF in JSON-LD format. For each obtained output, ChatBS employs a relationship discovery and entity linking service that
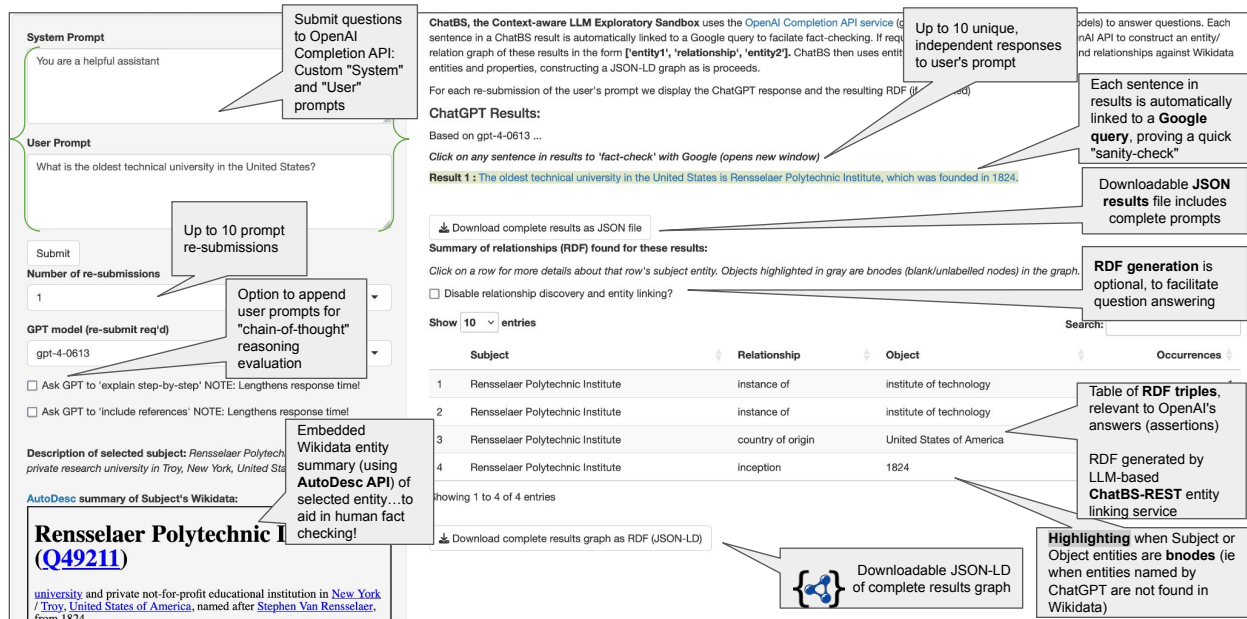
Figure 1: ChatBS user interface.

parses and analyses the generated content to produce a graph with the LLM's claims. For this, we implemented the approach in [5] which involves the use of pre-trained language models for entity and relationship recognition and naive entity linking to establish connection with Wikidata.

ChatBS, as a web application, is entirely accessible through any modern web browser. Its architecture comprises four basic elements: UI, OpenAI's continuation API, relationship discovery and entity-linking, and the Wikidata KG. The UI (Fig. 1) is developed as an R Shiny [6] app, providing a textual input for users to enter a natural language question about something (just like a user would do when using ChatGPT) and a parameter to set the desired number of answers for the same question (this feature was incorporated to stress the potential different answers, sometimes conflicting, LLMs can generate for the very same prompt). Original responses from the LLM are displayed on the right-hand side panel. Generated triples produced by the relationship discovery module, composing the graph representing the responses' claims, as a table, are optionally displayed by the use of a switch. In addition, a Wikidata snippet conveniently displays the Wikidata web page for any of the entities or relationships linked by the entity-linking module by clicking on any of them. This snippet can be leveraged by users to contrast claims in the triples table against crowd-sourced information in Wikidata, for instance. ChatBS's architecture is summarized in Fig. 2.

3. Impact and Conclusion

ChatBS has proven helpful in a variety of ways beyond its initial goal as an LLM fact-checker. As a clean, efficient LLM client optimized for experimentation, it has been especially useful for
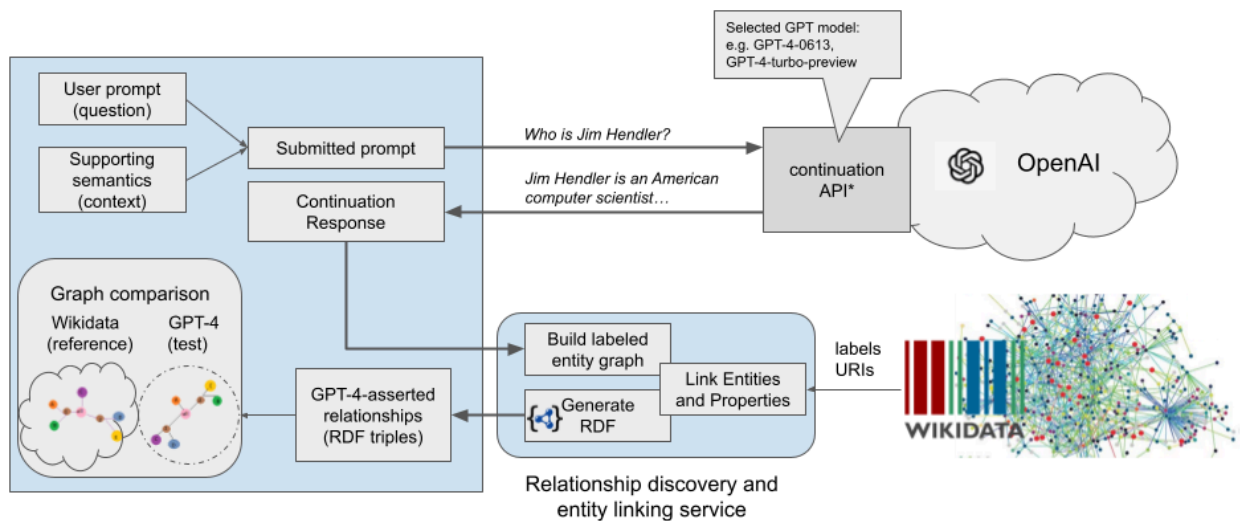
Figure 2: ChatBS architecture.

the *ad hoc* replication of LLM-based research results, for evaluating potential research methods, including prompt engineering strategies, and as a research tool unto itself. ChatBS has been employed in the classroom to support the reproducibility of published research by students within Rensselaer Data INCITE [7] (e.g. DecodingTrust [8] and BBQ [9]). In addition, it has been used to support original publications in diverse domains, such as finance [10] and health [11].

Acknowledgments

References

[1] J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, S. Zhong, B. Yin, X. Hu, Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond, ACM Trans. Knowl. Discov. Data 18 (2024) 160:1–160:32. doi:10.1145/3649506.

[2] L. Jiang, J. D. Hwang, C. Bhagavatula, R. L. Bras, J. Liang, J. Dodge, K. Sakaguchi, M. Forbes, J. Borchardt, S. Gabriel, Y. Tsvetkov, O. Etzioni, M. Sap, R. Rini, Y. Choi, Can Machines Learn Morality? The Delphi Experiment, 2022. doi:10.48550/arXiv.2110.07574, arXiv:2110.07574 [cs].

[3] Y. Tan, D. Min, Y. Li, W. Li, N. Hu, Y. Chen, G. Qi, Can ChatGPT Replace Traditional KBQA Models? An In-Depth Analysis of the Question Answering Performance of the GPT

LLM Family, in: T. R. Payne, V. Presutti, G. Qi, M. Poveda-Villalón, G. Stoilos, L. Hollink, Z. Kaoudi, G. Cheng, J. Li (Eds.), The Semantic Web – ISWC 2023, Lecture Notes in Computer Science, Springer Nature Switzerland, Cham, 2023, pp. 348–367. doi:10.1007/978-3-031-47240-4_19.

[4] Introducing Llama 3.1: Our most capable models to date, 2024. URL: https://ai.meta.com/blog/meta-llama-3-1/.

[5] J. McCusker, LOKE: Linked Open Knowledge Extraction for Automated Knowledge Graph Construction, 2023. doi:10.48550/arXiv.2311.09366, arXiv:2311.09366 [cs].

[6] W. Chang, J. Cheng, J. Allaire, C. Sievert, B. Schloerke, Y. Xie, J. Allen, J. McPherson, A. Dipert, B. Borges, shiny: Web Application Framework for R, 2024. URL: https://shiny.posit.co/, r package version 1.8.1.9000, https://github.com/rstudio/shiny.

[7] The Rensselaer Data Informatics Challenges in Technology Education Pipeline (Data INCITE), 2024. URL: https://idea.rpi.edu/research/projects/data-incite.

[8] B. Wang, W. Chen, H. Pei, C. Xie, M. Kang, C. Zhang, C. Xu, Z. Xiong, R. Dutta, R. Schaeffer, S. T. Truong, S. Arora, M. Mazeika, D. Hendrycks, Z. Lin, Y. Cheng, S. Koyejo, D. Song, B. Li, DecodingTrust: a comprehensive assessment of trustworthiness in GPT models, in: Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23, Curran Associates Inc., Red Hook, NY, USA, 2024, pp. 31232–31339.

[9] A. Parrish, A. Chen, N. Nangia, V. Padmakumar, J. Phang, J. Thompson, P. M. Htut, S. Bowman, BBQ: A hand-built bias benchmark for question answering, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Findings of the Association for Computational Linguistics: ACL 2022, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 2086–2105. doi:10.18653/v1/2022.findings-acl.165.

[10] M. Lashuel, G. Kurdistan, A. Green, J. S. Erickson, O. Seneviratne, K. P. Bennett, LLM-Based Code Generation for Querying Temporal Tabular Financial Data, In review (2024).

[11] K. Bhanot, J. S. Erickson, K. P. Bennett, MortalityMinder: Visualization and AI Interpretations of Social Determinants of Premature Mortality in the United States, Information 15 (2024) 254. doi:10.3390/info15050254, number: 5 Publisher: Multidisciplinary Digital Publishing Institute.