Designing an open-source LLM interface and social platforms for collectively driven LLM evaluation and auditing

Anonymous Author(s)

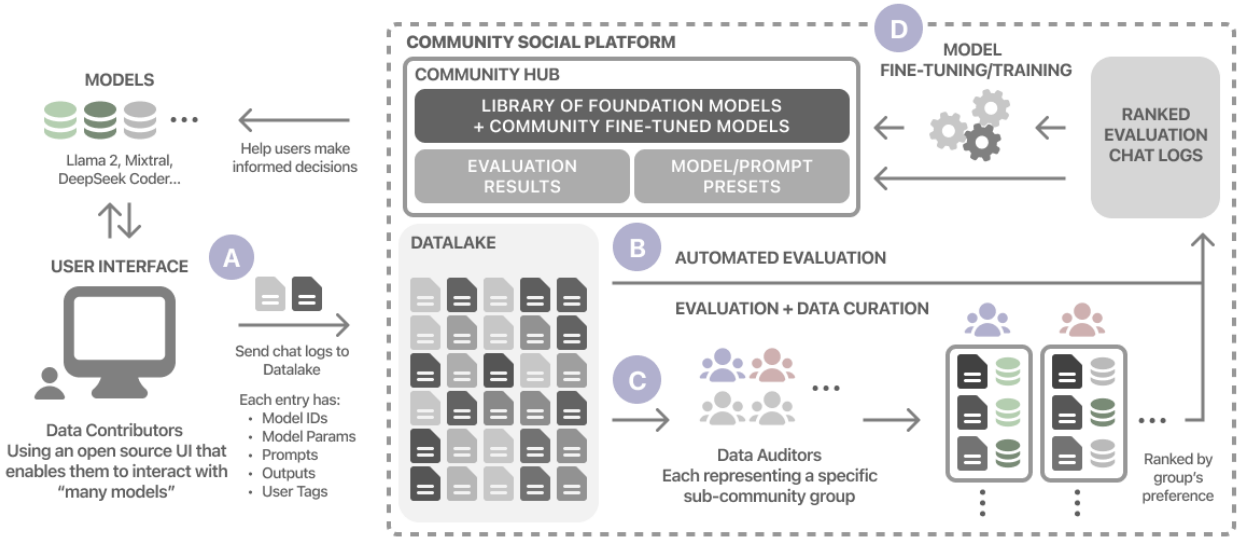

Figure 1: Overview of our social platform for LLM evaluation and auditing. (A) Users interact with open-source user interface, which allows them to easily switch between many models and submit their chat logs including model details and user information to community datalake. (B) Automated evaluation pipeline analyzes collected chat logs for comparative analysis based on user feedback, enabling a preliminary quantitative assessment of model performance and satisfaction. (C) Crowdsourcing in specialized sub-communities conducts double-blind evaluations of chat logs, eliciting feedback on desired model behaviors, facilitating qualitative analysis, and aligning model development with sub-community preferences. (D) Aggregated evaluations from automated and crowdsourced pipelines inform the development of community-specific model configurations and provide curated datasets, enabling tailored LLM solutions. Users receive comprehensive evaluation results and recommendations for effective model and prompt configurations, fostering continuous improvement based on real-world usage and community feedback.

ABSTRACT

In the emerging landscape of large language models (LLMs), the imperative for robust evaluation and auditing mechanisms is paramount to ensure their ethical deployment and alignment with user needs. This workshop paper proposes a novel framework for the human-centered evaluation and auditing of LLMs, centered around an open-source chat user interface (UI) that facilitates direct interaction with a wide range of models. This approach allows for a collection of rich datasets critical for nuanced evaluation from a diverse spectrum of user interactions. Building on this foundation, we propose a social platform designed to leverage the collective intelligence of its users through crowdsourcing, enabling the evaluation and auditing of LLMs across various domains. This platform supports a dual-layered evaluation pipeline: an automated preliminary assessment based on user feedback and a deeper, community-driven analysis within domain-specific subcommunities. The culmination of this process informs the development of tailored model configurations and curated datasets, ensuring that LLMs serve the specific needs of different user groups. By combining an open-source UI with a socially-driven evaluation platform, our approach fosters a community-centric ecosystem for continuous LLM improvement, emphasizing transparency, inclusivity, and alignment with human values.

CCS CONCEPTS

• **Human-centered computing** → **Interaction design**; **Natural language interfaces**; • **Computing methodologies** → **Natural language processing**.

KEYWORDS

Large Language Models, User Interface, Evaluation, Crowdsourcing, Data Curation, Data Ethics, Community-driven AI

1 INTRODUCTION

The rapid advancement of Large Language Models (LLMs) suggests a critical need for interactive human-centered evaluation and auditing methods that capture how well these systems work [6]. Thus far, LLM evaluation has relied heavily on benchmarks that summarize LLM performance but often fail to capture the nuanced requirements of different user contexts and needs, as every evaluation set represents some set of perspectives of users and contexts, making them limited in some ways. On top of the core challenge that creating an evaluation set requires prioritizing views and values over others, there are other concerns with the current evaluation paradigm regarding issues of dataset leakage and the resulting overfitting to evaluation data [17].

One avenue for enabling improved LLM evaluation and auditing is to design new interfaces and platforms that allow users to opt-in to sharing "in the wild" data. Some efforts to integrate evaluation interfaces with crowdsourcing platforms already exist. For instance, Chatbot Arena [15] from Lmsys draws on gamification patterns from social computing [12]. However, there are a number of downsides to existing implementations: designing an ecologically valid environment is challenging and users who participate more heavily may be much more heavily weighted than others. This lack of representation is especially pertinent for using crowdsourced data to evaluate and audit LLM usage in domains like workplace and education settings, where the context of use is crucial.

In this workshop paper, we describe early stage research aimed at applying an open-source user interface (UI) and a social platform to the problem of LLM evaluation and auditing. We describe the potential for designing a social platform that allows communities or individuals who operate local, decentralized LLMs to opt in to share data for evaluation and auditing. We discuss both aspects of the local LLM interface that are relevant to auditing, as well as propose an initial set of design goals for a social platform that provides incentives for high-quality data sharing and describe in-progress work to achieve these design goals. We emphasize a core challenge that will become increasingly salient to both HCI and ML researchers: as "local LLM" offerings proliferate and offer users a chance to decentralize and opt out of data collection, how can we provide hubs that foster symbiotic relationships? In other words, what interface and platform design considerations are needed so that individual LLM users, communities of users, and the LLM research community all stand to benefit?

Our work draws heavily on social computing and crowdsourcing with the aim of incorporating many user perspectives. By gathering a richer, more representative dataset that reflects the varied ways in which different communities interact with LLMs, we can attain a more nuanced understanding of model performance across different contexts and use cases. This approach with open-source user interfaces also offers the potential to integrate evaluation more seamlessly into the workflows of people using LLMs for work or personal reasons, allowing for the continuous evolution of evaluation metrics in line with changing user needs and preferences.

1.1 Contributions

This workshop paper makes two primary contributions. First, we describe how open-source, extensible, and locally capable interfaces (as opposed to interfaces operated through private services) for LLMs can help support evaluation. Second, we describe how social platforms that incorporate elements of crowdsourcing and data sharing might enable highly pluralistic evaluation studies and audits (i.e. evaluations with more perspectives and values represented) Additionally, we raise the ethical challenges surrounding data collection and the potential for peoples' values to conflict that are inherent in this approach.

Our paper aims to support a paradigm shift towards LLM evaluation that emphasizes insights and frameworks from HCI and social computing. With appropriately designed interfaces and social platforms, the research community can benefit from online communities that opt-in to share data in the form of new datasets for evaluation and auditing that represent a broader set of user needs and human values.

In the subsequent section, we delve into the importance of leveraging user interface interactions for evaluating LLMs. Here, the emphasis is placed on the significance of utilizing an open-source user interface, which not only facilitates seamless interaction with many models but also serves as a cornerstone for employing crowdsourcing as a tool for overcoming the existing limitations in LLM evaluation methodologies. Critically, there are several lines of work from open-source software contributors that already exist, and we describe how the research community might leverage these efforts.

Building upon the groundwork laid by the open-source user interface, the following section explores the role of social platforms in the realm of LLM evaluation. This discussion is not just confined to the symbiosis between user interfaces and crowdsourcing techniques but also extends to the concept of data curation via community-driven evaluation. We highlight the necessity of striking a balance between centralized and decentralized evaluation practices and explore future potentials and the challenges inherent in cultivating a social networking ecosystem dedicated to the nuanced evaluation of LLMs.

2 EVALUATING LLMS THROUGH USER INTERFACE INTERACTION

2.1 The Interface as a Evaluation Tool

The interface that somebody uses to interact with LLMs can massively impact how the underlying system is evaluated. The user interface, by its very nature, is intrinsically linked to the user experience. It serves as the primary medium through which users interact with LLMs, thereby playing a critical role in shaping their perceptions and overall satisfaction with the model. Every assessment of an LLM, implicitly or explicitly, involves an interface through which the model's capabilities are accessed and judged. Therefore, the interface becomes a fundamental aspect of the evaluation, offering a window into how users naturally engage with the model.

By integrating interface evaluation into the assessment of LLMs, we ensure that our understanding of these models is grounded in real-world usage, thereby enhancing the relevance and applicability of our findings. This section delves deeper into the rationale and

advantages of considering the user interface as an essential part of evaluating LLMs.

A key challenge in LLM evaluation is ensuring that the results are reflective of real-world usage and ecologically valid. Widely used open-source chat interfaces [1, 7, 10, 13] allow us to capture a more realistic picture of how LLMs are used in various contexts. For instance, the manner in which a user interacts with an LLM for educational purposes can vastly differ from how they might use it for creative writing or coding assistance in workplace settings [14].

The dynamic nature of user interfaces also supports continuous and evolving evaluation. As users interact with LLMs, their needs and expectations may change, and new challenges may emerge. A one-off evaluation process cannot capture this evolution. However, by leveraging user interfaces, we can continuously gather data and feedback, adapting the evaluation process to match the evolving landscape of LLM usage. This approach ensures that the evaluation remains relevant and aligned with current user needs and expectations.

The diversity of users interacting with LLMs through interfaces ensures that evaluations capture a wide range of perspectives and use cases. This diversity is critical for developing LLMs that are equitable, fair, and broadly applicable. By analyzing data from various users, we can identify and address biases or gaps in the model's performance, ensuring that the LLM is effective and appropriate for a wide array of users and scenarios.

2.2 Case Study of an Extensible Interface for Evaluation: Open WebUI

The evaluation of LLMs has reached a critical juncture where traditional metrics and benchmarks no longer suffice [17]. Open WebUI [13] is an open-source software (OSS) interface for local (e.g. Meta's downloadable Llama 2) and/or private (e.g. OpenAI's GPT) LLMs. Its design and functionality offer a novel paradigm that aligns more closely with the real-world usage and expectations of LLMs. In the subsequent paragraphs, we delve deeper into the various facets that make Open WebUI an invaluable tool in the evaluation landscape of LLMs.

The primary strength of open-source UIs like Open WebUI lies in their ability to obtain data from real-world interactions between users and LLMs. Open WebUI, being an open-source LLM UI that operates entirely locally, in contrast to platforms such as ChatGPT which run on centralized servers [8], offers end-users a similar experience to using ChatGPT that they're accustomed to. This local deployment capability allows Open WebUI to be used in a variety of settings, from high-security environments to remote locations with limited internet access. Its open-source nature not only democratizes the evaluation process but also fosters a community-driven approach to understanding and improving LLMs.

Additionally, Open WebUI enables users to interact with multiple LLMs in various configurations including OpenAI APIs within the same UI. This flexibility is crucial in evaluating the models' performance across different settings and use cases. It opens avenues for comparing different models under identical conditions, or the same model under varying conditions, providing a rich dataset for nuanced analysis. This level of customization and flexibility in evaluation cannot be achieved in traditional evaluation methods.

Unlike isolated testing environments, open-source UIs including Open WebUI enable users to engage with LLMs in their natural digital habitats – be it for work, education, or personal use. This real-world interaction data is invaluable, providing insights into how LLMs perform under diverse and often unpredictable conditions. By analyzing these interactions, we can gauge the practical utility, adaptability, and reliability of LLMs in actual usage scenarios. Furthermore, Open WebUI's local execution allows for the collection of data on model speed and performance across various hardware configurations. This aspect is crucial for stakeholders who need to understand the operational requirements and limitations of LLMs in different hardware environments. By providing insights into how LLMs perform on diverse hardware, Open WebUI enables a more holistic analysis, aiding in the optimization of LLMs for a wide range of applications.

One of the notable advantages of Open WebUI, which could also be easily extended and implemented to other open-source UI projects, is its streamlined data exportability. This feature significantly facilitates the use of its existing userbase for data collection in evaluations. Open WebUI enables a seamless aggregation of user interaction data, which is crucial for conducting evaluations that are both accurate and reflective of diverse real-world scenarios. Historically, such in-depth and varied data collection was exclusively achievable in industry environments, where companies could tap into their captive audiences with their centralized systems [2, 3, 8]. Open WebUI, by contrast, presents a more transparent and accessible option, breaking down barriers that previously limited research to industry confines via crowdsourcing. This not only empowers researchers to develop more robust evaluation metrics but also opens up a multitude of opportunities for experimentation and analysis with real-life usage data. The potential applications of this approach are vast, extending well beyond mere evaluation to explore uncharted territories in LLM application and performance.

2.3 Crowdsourcing as a Method for Evaluation

A fundamental advantage of crowdsourcing in LLM evaluation is its ability to capture a wide range of user interactions, encompassing diverse languages, cultural contexts, and application domains. Traditional evaluation methods often rely on a limited set of benchmarks or datasets, which may not adequately represent the vast array of potential LLM users. Open-source UIs, by facilitating user interactions from various backgrounds and with different needs, can gather data that is more representative of the global user base. This inclusive data collection is crucial for identifying and addressing biases in LLMs, ensuring that these models are fair and effective for a wide spectrum of users.

The crowdsourced approach also enables the rapid identification of issues and challenges within LLMs. In a traditional evaluation setting, the process of identifying flaws or biases is often slow and iterative, typically confined to the perspectives of a small group of developers or evaluators. In contrast, crowdsourcing allows for immediate feedback from a large and diverse user base. This feedback is not limited to technical performance but also includes the ethical and social implications of LLM responses. Users can report inappropriate or biased responses, unusual behavior, or other issues which can then be quickly addressed by developers.

2.4 Enhancing User Engagement in Crowdsourcing LLM Evaluation

The success of evaluating LLMs through user interfaces like Open WebUI, hinges critically on active stakeholder participation. This engagement is not merely about having users interact with the models; it is about involving them deeply in the evaluative process, thereby transforming them from passive consumers to active contributors. Thus, encouraging broad user participation is not just a supplementary goal; it's a fundamental requirement for the validity and effectiveness of this evaluation method.

2.4.1 Creating a Community-Centric Evaluation Ecosystem. The first step towards enhanced user engagement is the cultivation of a community around the LLMs. This community should be built on the principles of collaboration, shared learning, and mutual benefit. By participating in the evaluation process, users not only contribute to the improvement of the LLM but also gain insights into its capabilities and limitations. This two-way street creates a sense of ownership and responsibility among the users, which is crucial for sustained engagement.

2.4.2 Gamification and Recognition. Gamification elements can significantly boost user engagement [4, 12]. Incorporating elements like badges, leaderboards, or points for active contributors can create a more engaging and rewarding experience. Recognizing top contributors publicly within the community not only motivates them but also inspires others to participate actively.

2.4.3 Facilitating Peer Learning and Sharing. Encouraging users to share their best practices, interesting findings, and custom prompts can foster a learning environment within the community [5]. This peer-to-peer interaction ensures that users are not just contributing data for evaluation but are also learning from each other, making their participation more rewarding.

2.4.4 Open Question: Effective Strategies for Sustained Engagement. The strategies outlined above lay the groundwork for fostering active and meaningful user engagement in the evaluation of LLMs through Open WebUI. However, the question remains: how can we implement these strategies most effectively? This is not just a matter of logistical planning but also of understanding the diverse motivations and constraints of potential users. What incentives will be most compelling? How can we balance the need for high-quality, meaningful interactions with the desire to involve as many users as possible? These questions lead us to the next section, which will explore potential solutions and blueprints for incentivizing participation, ensuring that users are not only willing but also eager to contribute to this crucial evaluative endeavor.

3 SOCIAL PLATFORMS FOR EVALUATION AND BEYOND

The integration of social platforms in the evaluation of LLMs marks a significant shift in our approach towards understanding and improving these technologies. This section delves into the multifaceted role of social platforms in LLM evaluation, emphasizing the dynamic interplay between user interfaces, crowdsourcing methods, and community-driven data curation. First, we highlight the delicate balance required between centralized and decentralized evaluation ecosystems, advocating for a middle path that leverages the strengths of both to foster open research and diverse perspectives. Subsequently, We explore the existing landscape, future possibilities, and the inherent challenges of building a social network-like ecosystem for LLM evaluation.

3.1 Navigating the Centralization-Decentralization Spectrum

Our proposed social platform operates on the principle of finding an optimal balance between centralization and decentralization of the overarching context in which LLMs are operated (see e.g. discussion of the complexity of decentralization as a concept and rhetorical strategy[9]). This balance is crucial for capturing a broad and diverse range of data and perspectives, which is essential for effective evaluations of LLMs.

A centralized approach, while efficient in data collection, may fail to capture the nuanced needs and contexts of a diverse user base. A purely centralized approach also risks homogenizing the data collected, primarily representing the active user base while sidelining the perspectives of potential users and non-users. This narrow data capture might fail to reflect the diverse contexts in which LLMs operate, limiting the model's ability to generalize across different populations and use cases. Moreover, centralization often places barriers to open science, making it challenging for the academic community to access, utilize, and contribute to the dataset, thereby stifling collaborative innovation and transparency in LLM development.

On the other end of the spectrum, decentralization offers a distributed model of data collection and management, where control and ownership are spread across a wider array of participants. This approach naturally facilitates a broader capture of data and perspectives, as it empowers users from varied backgrounds to contribute their unique interactions and feedback. However, the challenge with decentralization lies in aggregating sufficient data to achieve meaningful insights and evaluations. Without adequate participation, the data collected may be too sparse to inform robust LLM evaluations or to understand complex user interactions comprehensively.

Our platform aims to mediate this spectrum by encouraging broad participation and allowing users control over their data, thus ensuring a rich dataset that encompasses a wide array of interactions and perspectives. To ensure the platform's success in attracting and retaining a diverse user base, it is imperative to create incentives for participation. Recognizing contributors for their valuable insights or data that leads to model improvements can motivate ongoing engagement. Additionally, the platform must appeal to a variety of users, from casual conversationalists to professionals in need of reliable decision-support tools. This diversity is vital for evaluating LLMs across different scenarios, ensuring the development of models that are robust, reliable, and broadly applicable.

3.2 The Current Landscape and Future Vision

Currently, Open WebUI's social platform stands at the forefront of integrating LLM interaction with the dynamic of social platforms, creating an ecosystem where users can actively participate in the evaluation and enhancement of LLMs [13]. This integration not only enhances user engagement but also fosters a collective space

for learning, sharing, and customizing model interactions. The platform's features, such as sharing model configurations along with chat logs and engaging in collaborative prompt engineering, demonstrate the potential of this hybrid model to significantly impact the field of LLM evaluation.

Looking forward, we imagine a social platform that not only supports interaction with LLMs but also enables users to actively engage and contribute in the evaluation and development process. This future platform would function almost like a social network, but with a focused purpose: to collaboratively improve LLMs. A crucial aspect of this vision is acknowledging that community involvement in curating and evaluating datasets leads to more refined, representative, and community-led methodologies [11]. This engagement deeply enriches the AI evaluation process, ensuring the development of technologies that are not only technologically advanced but also in tune with the diverse needs and perspectives of the user community. By embracing community-driven evaluation and dataset curation, our social platform can leverage the collective wisdom and diversity of its users to foster a more inclusive, accurate, and effective LLM evaluation ecosystem.

3.3 Crowdsourcing Methodology for Evaluation

Our social platform introduces a dual-track evaluation pipeline designed to harness the power of crowdsourcing for the nuanced assessment of LLMs. This novel approach combines automated evaluation with deep, community-driven auditing to offer a comprehensive evaluation framework that is both scalable and adaptable to diverse user needs and preferences.

Automated Evaluation Pipeline: At the foundation of our evaluation framework lies the automated pipeline, which utilizes the raw chat logs collected from user interactions with LLMs. Each interaction is tagged with models utilized and its IDs, allowing for an organized comparative analysis based on user feedback, exemplified by mechanisms such as "thumbs up" or "thumbs down" ratings or response regeneration history. This system facilitates a preliminary, quantitative assessment of model performance and user satisfaction, acting as a crucial initial filter within our comprehensive evaluation framework. The automated pipeline's efficiency lies in its ability to quickly aggregate and analyze vast amounts of feedback, providing a baseline understanding of model strengths and areas for improvement.

Community-Driven Evaluation Pipeline: Building upon the automated evaluation, we further refine our evaluation process through a robust, crowdsourced pipeline. This pipeline leverages the intrinsic value of the platform's diverse subcommunities—groups with specialized interests or expertise, such as those focusing on medical, legal, or programming domains. By engaging these groups in a double-blind review of chat logs and model interactions, we solicit deeper, qualitative insights into model performance. Critical questions posed to these communities, such as "Do you want your model to behave more in this way?", facilitate a more nuanced evaluation. This method not only assesses model adequacy but also aligns future model development with the specific needs and preferences of different user groups, fostering a highly tailored and community-centric approach to LLM improvement.

Roles and Responsibilities:

(1) **Data Contributors:** These users are at the heart of our evaluation ecosystem, directly interacting with LLMs and contributing invaluable interaction logs. This role is crucial for generating the raw data that feeds both evaluation tracks. Contributors can enhance their submissions with tags and annotations, providing crucial metadata that enriches the dataset and guides subsequent evaluations and model refinements.
(2) **Data Auditors:** Running parallel to the contributions are the Data Auditors tasked with curating the dataset. They review submissions based on quality, relevance, and adherence to ethical standards, employing upvotes or downvotes to signal the value of each contribution. This process not only maintains the dataset's integrity but also democratizes the evaluation process, enabling a community consensus on the standards and benchmarks for LLM performance.

This dual-track approach offers a dynamic and flexible framework for LLM evaluation, combining the scalability of automated processes with the depth and contextuality of human judgment. By leveraging the collective intelligence of our platform's subcommunities, we enable a more democratic and inclusive evaluation process. This not only enhances the quality of the dataset but also ensures that model development is continually informed by real-world feedback and evolving user needs. This crowdsourcing methodology not only incentivizes participation by offering users a tangible impact on model development but also aligns with the broader goal of creating more reliable, ethical, and user-centric LLMs.

3.4 Data Curation as a Byproduct of Evaluation

Data curation emerges as a natural byproduct of the evaluation process on social platforms designed for LLM auditing. This highlights an often overlooked opportunity to enhance model performance through the strategic collection and organization of evaluation data. As users interact with and evaluate LLMs, their inputs, feedback, and the contexts of their interactions generate a wealth of data that, if properly curated, can significantly inform and refine model training processes.

The act of evaluation itself, especially when informed by diverse user experiences and insights, leads to the generation of highly relevant and contextualized data sets. These datasets not only reflect a wide range of user needs and preferences but also embody the nuances of language and interaction patterns across different domains and demographics. By curating this data, the platform can create rich, annotated resources that provide invaluable insights for the fine-tuning of LLMs.

Effective data curation requires a deliberate approach to the design of evaluation activities and the UI of the platform. Encouraging users to provide detailed feedback, ask questions, and share their interaction contexts helps in collecting more granular and actionable data. Furthermore, integrating prompts or questions that guide users in reflecting on specific aspects of their interaction with LLMs can enrich the dataset with targeted insights on model behavior, user expectations, and potential improvements.

This approach not only facilitates the direct improvement of LLMs through enhanced training datasets [16] but also contributes to the broader field of AI research by generating publicly available

datasets that capture a wide array of human-AI interactions. Such datasets are invaluable for the development of models that are more aligned with human needs, ethical standards, and societal values.

3.5 Challenges and Ethical Considerations

As we chart this new territory, several challenges emerge. Ensuring the authenticity of user contributions, incentivizing stakeholders to participate from across user distribution to mitigate representational harm, establishing effective moderation systems to maintain data quality, determining the appropriate weighting of community votes, and building user interfaces intuitive enough to facilitate the entire process are critical considerations. Moreover, the ethical implications of data collection, privacy concerns, and the need for mechanisms to anonymize sensitive information must be addressed rigorously.

The integration of social platforms in the evaluation of LLMs represents a significant evolution in our approach towards understanding and improving these complex systems. By leveraging the collective intelligence and diverse experiences of a broad user base, we can develop more nuanced, context-aware, and ethically aligned models. This paradigm shift not only enriches the evaluation process but also democratizes it, fostering a more inclusive and participatory ecosystem for LLM development.

REFERENCES

[1] Yuvanesh Anand, Zach Nussbaum, Adam Treat, Aaron Miller, Richard Guo, Ben Schmidt, GPT4All Community, Brandon Duderstadt, and Andriy Mulyar. 2023. GPT4All: An Ecosystem of Open Source Compressed Language Models. arXiv:2311.04931 [cs.CL]

[2] Anthropic. 2024. Claude. https://claude.ai/

[3] Google. 2024. Gemini - chat to supercharge your ideas. https://gemini.google.com/

[4] Juho Hamari, Jonna Koivisto, and Harri Sarsa. 2014. Does gamification work?–a literature review of empirical studies on gamification. In *2014 47th Hawaii international conference on system sciences*. Ieee, 3025–3034.

[5] Noriko Hara. 2008. *Communities of practice: Fostering peer-to-peer learning and informal knowledge sharing in the work place*. Vol. 13. Springer Science & Business Media.

[6] Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, Rose E. Wang, Minae Kwon, Joon Sung Park, Hancheng Cao, Tony Lee, Rishi Bommasani, Michael Bernstein, and Percy Liang. 2024. Evaluating Human-Language Model Interaction. arXiv:2212.09746 [cs.CL]

[7] oobabooga. 2024. oobabooga/text-generation-webui. https://github.com/oobabooga/text-generation-webui original-date: 2022-12-21T04:17:37Z.

[8] OpenAI. 2024. ChatGPT. https://chat.openai.com/

[9] Nathan Schneider. 2019. Decentralization: an incomplete ambition. *Journal of cultural economy* 12, 4 (2019), 265–285.

[10] SillyTavern. 2024. SillyTavern/SillyTavern. https://github.com/SillyTavern/SillyTavern original-date: 2023-02-09T10:19:24Z.

[11] Zirui Cheng Jiwoo Kim Meng-Hsin Wu Tongshuang Wu Kenneth Holstein Haiyi Zhu Tzu-Sheng Kuo, Aaron Halfaker. 2024. Wikibench: Community-Driven Data Curation for AI Evaluation on Wikipedia. *arXiv preprint arXiv:2402.14147* (2024).

[12] Luis Von Ahn and Laura Dabbish. 2008. Designing games with a purpose. *Commun. ACM* 51, 8 (2008), 58–67.

[13] Open WebUI. 2024. open-webui/open-webui. https://openwebui.com/

[14] Lixiang Yan, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and Dragan Gašević. 2024. Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology* 55, 1 (2024), 90–112.

[15] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems* 36 (2024).

[16] Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: Less Is More for Alignment. arXiv:2305.11206 [cs.CL]

[17] Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. 2023. Don't Make Your LLM an Evaluation Benchmark Cheater. arXiv:2311.01964 [cs.CL]