How to Reflect Diverse People's Perspectives in Large-Scale LLM-based Evaluations?

Yoonjoo Lee
School of Computing, KAIST
Daejeon, Republic of Korea
yoonjoo.lee@kaist.ac.kr

Tae Soo Kim
School of Computing, KAIST
Daejeon, Republic of Korea
taesoo.kim@kaist.ac.kr

Juho Kim
School of Computing, KAIST
Daejeon, Republic of Korea
juhokim@kaist.ac.kr

ABSTRACT

Despite the remarkable capability of Large Language Models (LLMs) to process a broad spectrum of inputs and outputs in various tasks, existing evaluation methodologies, including both benchmarks and human evaluations, struggle to capture the nuanced performance of LLMs or are not scalable. Recent initiatives have leveraged LLMs as evaluators to achieve more scalable and generalizable assessments of model performance, yet this approach raises concerns about the inclusiveness of the perspectives represented in these evaluations. To encompass the perspectives of diverse individuals, we propose a preliminary workflow for simulating the evaluations of a wide range of individuals through LLMs by iterating between constructing user models of representative users and expanding the diversity of the representatives. This method aims to create a more inclusive evaluation framework that can help LLM modelers or system builders to gain a more comprehensive sense of their model's performance across diverse inputs, tasks, and users.

1 INTRODUCTION

As Large Language Models (LLMs) become more capable, their ability to process a wide array of inputs and outputs allows them to undertake diverse tasks and engage in complex interactions with diverse users. For instance, while researchers previously proposed models that could perform conversational question-answering (ConvQA) in highly specialized types of text [2], a single state-of-the-art LLM can provide relevant responses for diverse domains and diverse audiences spanning various age groups, interests, and levels of expertise [13, 14]. While models were frequently evaluated through benchmarks that provided a set of challenging inputs to the model with clear reference outputs, as LLMs can handle diverse tasks with many being open-ended—different outputs can satisfy the same input—this evaluation approach falls short of adequately assessing LLM performance [12]. Furthermore, human evaluation methods, where human annotators assess generated outputs, are excessively costly to gain a generalizable view of model performance as it would require a large set of annotators to each assess a large set of outputs [6]. Consequently, there has been a move towards employing LLMs as annotators to simulate human evaluations, aiming for more generalizable evaluation results [1, 4, 10, 15, 16, 19]. These approaches use LLMs to annotate the quality of model outputs based on well-known general quality metrics or specific criteria, covering dimensions as objective as "grammar" and as subjective as "helpfulness".

However, LLM-based evaluation methods have their limitations, notably in their failure to encapsulate the perspectives of diverse individuals. Prior studies argue for using LLMs as alternatives to human evaluators by demonstrating how the agreement between LLM evaluations and the aggregated evaluation of a small set of humans is comparable to the agreement between human evaluators [20]. However, having a comparable agreement to the agreement between human evaluators does not indicate that the LLM is reflecting the diverse perspectives of human annotators in its evaluations. For one, while disagreement in human annotations can arise from differences in perspective [7, 8], comparing LLM evaluations to the aggregated evaluations of humans ignores these differences and it is unclear whose perspective the LLM is aligning with. Additionally, as LLMs are frequently tuned on the aggregated preferences of users, prior work has demonstrated that these models can be biased in their perspectives and opinions [3, 5, 11]. As a result, LLM evaluations may be biased towards the perspectives of specific groups and, if LLMs or LLM-based pipelines are optimized according to these evaluations, it could lead to the reinforcement of these biases and the development of systems that only cater to particular user groups.

This position paper proposes a preliminary workflow that can enable researchers to leverage the scalability of LLM-based evaluation methods while reflecting the diverse viewpoints of a diverse populace. The workflow involves a two-step process: (1) verifying that LLM can simulate representative individuals by constructing comprehensive *user models*, and (2) expanding on the set of representatives to sufficiently cover a diverse populace.

2 HOW TO ENCOMPASS MULTIPLE PERSPECTIVES AND LEVERAGE THE SCALABILITY OF LLMS?

To encompass multiple perspectives while preserving the scalability in evaluation that can be provided by LLMs, we propose a two-step preliminary workflow that researchers can follow: (1) verifying that LLM can simulate selected individuals or groups by constructing comprehensive *user models*, and (2) expanding on the set of representatives to sufficiently cover a diverse populace. We also suggested more detailed questions that need to be investigated to realize each sub-step in this preliminary workflow.

1st HEAL Workshop at CHI Conference on Human Factors in Computing Systems, May 12, Honolulu, HI, USA
2024.

2.1 Can LLMs Simulate the Perspectives of Representative Individuals or Groups?

A critical issue in utilizing LLMs as evaluators is that it is unclear whose opinions and perspectives the LLMs are reflected in their evaluations, and whether LLMs can align their evaluations with diverse individuals. We propose that researchers should first systematically investigate this issue by assessing the alignment of LLM evaluations with a few representative individuals or groups, and exploring techniques to steer the LLM to reflect the perspective of these individuals or groups. Instead of immediately investigating whether the LLM can simulate numerous and diverse individuals, we suggest that researchers should first assess whether the LLM can simulate a few but significantly distinct representative individuals in a consistent and scalable manner (i.e., on diverse inputs and outputs). By focusing on representatives, researchers can focus on assessing how to *model* or represent users to an LLM to simulate their evaluations. In particular, various prior methods for user modeling can be tested: surveys [9], personal characteristics [21], profiles [17], and past artifacts/annotations [18]. By assessing and comparing several user modeling methods, researchers can understand how well LLMs can follow an individual's evaluation patterns and how to effectively model individuals with LLMs.

For this first step, we propose the following process for how researchers can employ LLM-based evaluations that can faithfully simulate individuals' perspectives. First, recruit representative individuals based on the specific task being evaluated. To recruit them, understanding who the target users is key. This involves identifying distinct dimensions to segment the user population such as demographic traits, skill levels, cultural backgrounds, or particular needs the LLM is designed to meet. If there's uncertainty about these dimensions, researchers can conduct pilot studies to help pinpoint user characteristics that significantly impact the perceived experience. For example, if the task is to translate between English and Korean, the researchers need to recruit participants who are capable in English and/or Korean at different levels. Second, ask participants to evaluate sample outputs from the application and collect information from participants to construct user models that can be most suitable to simulate each individual. Then, researchers need to check whether the LLM simulated evaluations have a statistically significant correlation with the annotations from individual human participants. If not, researchers need to enrich their user model of participants by receiving more detailed information before simulating the annotation again with more samples. For example, researchers can obtain more annotated samples and rationales on annotations, or ask clarification or follow-up questions to participants about their context and backgrounds. With this additional information, LLMs can possess more context to learn or infer the users' objectives, experiences, values, and backgrounds that might be necessary to simulate each individual. If the LLM is able to simulate multiple individuals after multiple rounds of iteration, then researchers can employ the LLM and user models to scalably evaluate a larger set of input/output samples while reflecting more perspectives to a certain degree—enabling more generalizable assessments.

2.2 How to Expand the Simulation to Encompass the Diversity of Individuals?

Even if we can sufficiently simulate a few representative users related to the task at hand, it is necessary to validate that the representatives fully represent the diversity among potential users and to identify if there are any blind spots in the represented users. However, since diversity can be defined on an infinite scale, it can be challenging to systematically identify those gaps and missing perspectives. To identify who should be additionally recruited, we suggest that researchers employ two different approaches. First, researchers need to analyze the explicit human characteristics of the current representatives (e.g., demographics, expertise, perspectives) to identify possible gaps. This involves reviewing the range of viewpoints expressed, the variety of backgrounds, and the scope of expertise covered. By mapping these factors against the intended audience or user base, researchers can pinpoint underrepresented areas. Second, researchers can also find gaps by uncovering more implicit characteristics in the users. For example, researchers can ask annotators to provide rationales or explanations for their evaluations, which can be used to uncover the annotator's implicit values, requirements, or intentions. By finding missing perspectives based on these implicit characteristics, researchers can identify differences between individuals that were difficult to notice solely through the explicit characteristics but that may have a more direct impact on users' assessments of output.

Subsequently, researchers can conduct targeted recruitment to focus on filling these gaps by seeking out individuals from these underrepresented user groups, ensuring that a more inclusive range of perspectives is considered. With the additionally recruited participants, researchers can conduct the annotation and simulation steps in Sec.2.1 again. By iterating through this loop and gradually involving more diverse individuals in the evaluation, LLM modelers and application builders can construct a more comprehensive and diverse set of simulated evaluators to understand where a model or system succeeds or fails based on the perspectives of different users, and what aspects need improvements to satisfy a broader range of people.

Though encompassing diversity is an important problem, diversity can be defined on an infinite scale. To expand the pool of simulated evaluators systematically, we suggest regarding diversity within the scope of whether that diversity can change outputs and user experiences.

3 CLOSING THOUGHTS

To conduct evaluations that can encompass the perspectives of diverse people, we propose the following preliminary workflow: (1) employing LLMs to simulate the perspective of representative individuals, and (2) expanding the represented users to run simulations that encompass a more diverse populace. Even if we can simulate and gather the opinions of diverse individuals, researchers need to conduct considerable exploration to determine whose opinions should be given more weight when making decisions about the behavior of the system, and to design novel strategies for governance or aggregation for the world of LLM-as-evaluators.

ACKNOWLEDGMENTS

This work was supported by NAVER-KAIST Hypercreative AI Center.

REFERENCES

[1] Cheng-Han Chiang and Hung-yi Lee. 2023. 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)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 15607–15631. https://doi.org/10.18653/v1/2023.acl-long.870

[2] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question Answering in Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, Brussels, Belgium, 2174–2184. https://doi.org/10.18653/v1/D18-1241

[3] Esin Durmus, Karina Nyugen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388* (2023).

[4] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. GPTScore: Evaluate as You Desire. *ArXiv* abs/2302.04166 (2023). https://api.semanticscholar.org/CorpusID:256662188

[5] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2023. Bias and Fairness in Large Language Models: A Survey. arXiv:2309.00770 [cs.CL]

[6] Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2022. Repairing the Cracked Foundation: A Survey of Obstacles in Evaluation Practices for Generated Text. *ArXiv* abs/2202.06935 (2022). https://api.semanticscholar.org/CorpusID:246822399

[7] Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. 2022. Jury Learning: Integrating Dissenting Voices into Machine Learning Models. In *CHI Conference on Human Factors in Computing Systems (CHI '22)*. ACM. https://doi.org/10.1145/3491102.3502004

[8] Mitchell L. Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S. Bernstein. 2021. The Disagreement Deconvolution: Bringing Machine Learning Performance Metrics In Line With Reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Yokohama</city>, <country>Japan</country>, </conf-loc>) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 388, 14 pages. https://doi.org/10.1145/3411764.3445423

[9] Chenyan Jia, Michelle S. Lam, Minh Chau Mai, Jeffrey T. Hancock, and Michael S. Bernstein. 2023. Embedding Democratic Values into Social Media AIs via Societal Objective Functions. *ArXiv* abs/2307.13912 (2023). https://api.semanticscholar.org/CorpusID:260164566

[10] Tae Soo Kim, Yoonjoo Lee, Jamin Shin, Young-Ho Kim, and Juho Kim. 2023. EvalLM: Interactive Evaluation of Large Language Model Prompts on User-Defined Criteria. *ArXiv* abs/2309.13633 (2023). https://api.semanticscholar.org/CorpusID:262459331

[11] Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in Large Language Models. In *Proceedings of The ACM Collective Intelligence Conference* (<conf-loc>, <city>Delft</city>, <country>Netherlands</country>, </conf-loc>) *(CI '23)*. Association for Computing Machinery, New York, NY, USA, 12–24. https://doi.org/10.1145/3582269.3615599

[12] Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to Progress in Long-form Question Answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (Eds.). Association for Computational Linguistics, Online, 4940–4957. https://doi.org/10.18653/v1/2021.naacl-main.393

[13] Jakub L'ala, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G. Rodriques, and Andrew D. White. 2023. PaperQA: Retrieval-Augmented Generative Agent for Scientific Research. *ArXiv* abs/2312.07559 (2023). https://api.semanticscholar.org/CorpusID:266191420

[14] Yoonjoo Lee, Tae Soo Kim, Sungdong Kim, Yohan Yun, and Juho Kim. 2023. DAPIE: Interactive Step-by-Step Explanatory Dialogues to Answer Children's Why and How Questions. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023). https://api.semanticscholar.org/CorpusID:258216670

[15] Ruosen Li, Teerth Patel, and Xinya Du. 2023. PRD: Peer Rank and Discussion Improve Large Language Model based Evaluations. *ArXiv* abs/2307.02762 (2023). https://api.semanticscholar.org/CorpusID:259360619

[16] Yang Liu, Dan Iter, Yichong Xu, Shuo Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. In *Conference on Empirical Methods in Natural Language Processing*. https://api.semanticscholar.org/CorpusID:257804696

[17] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. arXiv:2304.03442 [cs.HC]

[18] Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2023. LaMP: When Large Language Models Meet Personalization. *arXiv preprint arXiv:2304.11406* (2023).

[19] Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2024. FLASK: Fine-grained Language Model Evaluation based on Alignment Skill Sets. arXiv:2307.10928 [cs.CL]

[20] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Haotong Zhang, Joseph Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *ArXiv* abs/2306.05685 (2023). https://api.semanticscholar.org/CorpusID:259129398

[21] Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024. SOTOPIA: Interactive Evaluation for Social Intelligence in Language Agents. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=mM7VurbA4r