Evaluation Workflows for Large Language Models (LLMs) that Integrate Domain Expertise for Complex Knowledge Tasks

Annalisa Szymanski
Computer Science and Engineering
University of Notre Dame
Notre Dame, Indiana, USA
aszyman2@nd.edu

Abstract

The growing use of Large Language Models (LLMs) in specialized fields such as healthcare, nutrition, and education has raised critical concerns regarding the accuracy, reliability, and contextual appropriateness of LLM outputs. However, evaluating LLMs is challenging due to the complexity of the information and the need for human input, which is often costly and resource-intensive. My dissertation addresses the challenges in integrating domain expertise into the evaluation of LLM outputs for complex knowledge tasks to build more efficient evaluation workflows. The main objectives of this research are: 1) to investigate when and at what stage domain expertise should be integrated into LLM evaluation 2) to explore the role of domain experts compared to other evaluators such as lay users and LLMs themselves in the evaluation process, and 3) to design evaluation frameworks and tools that guide both the optimal integration of domain experts and leverage the complementary strengths of other evaluation groups. The expected impact of this research includes advancing the design of LLM evaluation tools and workflows that assist developers in identifying where expertise is needed to effectively develop and deploy LLMs in real-world applications.

CCS Concepts

• **Human-centered computing** → **Human computer interaction (HCI)**; • **Computing methodologies** → *Natural language generation.*

Keywords

Large Language Models, Evaluation Methods, LLM-as-a-Judge, Human-AI Interaction

ACM Reference Format:
Annalisa Szymanski. 2025. Evaluation Workflows for Large Language Models (LLMs) that Integrate Domain Expertise for Complex Knowledge Tasks. In *30th International Conference on Intelligent User Interfaces Companion (IUI Companion '25), March 24–27, 2025, Cagliari, Italy*. ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3708557.3716146

IUI Companion '25, Cagliari, Italy

ACM ISBN 979-8-4007-1409-2/25/03
https://doi.org/10.1145/3708557.3716146

1 Motivation and Related Work

Large Language Models (LLMs) have been increasingly utilized across various domains, such as healthcare, nutrition, education, and other specialized fields, to perform complex tasks [1, 4, 5, 16, 17]. Despite their widespread use, significant concerns remain about the accuracy, reliability, and variability in LLM performance, where errors and hallucinations can have serious implications through the perpetuation of biases and potentially harmful misinformation [2, 11, 13, 22]. Addressing these risks requires robust evaluation processes that provide developers with insights on how to select appropriate models and assess their suitability for specific tasks [14].

Several interactive evaluation systems have been proposed to assist developers in testing model performance. Tools such as EvalLM [12] and EvalGen [18] help identify biases, optimize prompts, and select models aligned with a user's objective [6, 10, 12]. In addition, systems like ChainForge [3] or Promptfoo [23] enable users to create custom evaluation criteria for assessing model responses. These systems often rely on the LLMs themselves to act as evaluators or intermediaries in the evaluation process. While the specific goals of these systems may vary, they share a common aim: to improve alignment between a user's preferences and LLM outputs [18], which are valuable in improving models to meet real-world needs.

However, the effectiveness of these evaluation systems for complex tasks remains in question and their utility has been limited to the developer, who focuses on technical aspects that may overlook the context-specific relevance of outputs that domain experts could provide. As highlighted by Pan et al. [15], while LLMs as evaluators can be useful, incorporating diverse human input, such as from domain experts, is essential to ensure alignment with professional standards, including compatibility with ethical concerns and context-specific knowledge [8, 9]. Yet, experts are often excluded from standardized evaluation workflows due to the significant costs, time, and resources required for their involvement [7]. As developers make decisions on how to design and deploy LLM technologies, there is a need to better understand how domain experts can contribute to evaluation workflows, and in what stage their expertise would be most effective.

The aim of my dissertation is to develop methods and frameworks to incorporate domain experts into the evaluation process of LLMs, and to focus on when and how to efficiently include their expertise while addressing the challenge of resource scalability. Through this exploration, my research will examine the critical tasks where domain expertise is most valuable and compare these evaluations with the roles of lay users and LLMs themselves. My research suggests that we can systematically explore how expert

input and non-expert feedback complement one another in optimizing outcomes across different contexts and tasks. The end goal is to create frameworks and tools that can guide AI developers in determining when each type of evaluation by domain experts, lay users, or the LLM themselves is most appropriate and for which specific tasks. These frameworks will be aimed to structure evaluation processes across various domains to ensure the right level of expertise is applied at appropriate stages. My research has the potential to transform how LLMs are deployed across critical fields such that they align with real-world needs while maintaining trust and accountability in their outputs.

2 Goals and Research Questions

The objectives of my doctoral research aim to meet the goal of building efficient evaluation workflows that integrate domain expertise most effectively. The primary research questions guiding my dissertation work are as follows:

(1) When is domain expertise most crucial, and what are the optimal stages for integrating expertise into the evaluation process?
(2) What are the similarities and differences among domain experts, lay users, or LLM themselves in the evaluation of complex tasks?
(3) How can we leverage the complementary strengths of domain experts, lay users, and LLMs to create effective evaluation frameworks and tools to guide developers in determining where each group's input is most critical?

3 PRIOR RESEARCH, METHODS, AND RESULTS

The following studies have already been completed and serve as a foundation for my future dissertation work. The studies are framed through case studies within complex knowledge task domains where LLMs are applied.

Study 1: Integrating Domain Expertise in LLM Evaluations. To explore the integration of domain expertise into LLMs, we conducted interviews and focus groups within the nutrition domain with registered dietitians (RDs). We focused on refining GPT-4 outputs to create a customized nutrition assistant tailored to provide accurate and reliable food product explanations to end users. Through semi-structured interviews, RDs assessed the strengths and weaknesses of LLM outputs at varying levels of prompt specificity. Based on their feedback, we developed a set of design guidelines used for the developer to prompt the LLM in providing accurate and personalized nutrition information. These guidelines were integrated into a custom GPT-4 model, and focus groups with dietitians led to further refinements of the prompt instructions. This study demonstrated that incorporating expert input into LLM evaluation and prompt instruction refinement improves LLM design guidelines and ensures that the generated nutrition information aligns with professional standards to offer more personalized and trustworthy outputs for end users. This study was presented at the CHI Conference on Human Factors in Computing Systems in May 2024 [20].

Study 2: Assessing LLM-as-a-Judge versus Domain Expert Evaluation. As a follow up to our work with domain experts, we evaluated how well the LLMs-as-a-Judge performs in comparison to domain experts when evaluating LLM outputs. Both domain experts and LLMs were tasked with performing pairwise comparisons of model outputs in the specialized fields of nutrition and mental health, and the results revealed that LLMs agreed with experts only 64%-68% of the time, respectively. In addition, results indicated variations in agreement depending on the domain specific task and aspect questions, indicating the need for domain-specific evaluation frameworks in the future. The study discusses the limitations of the LLM-as-a-Judge approach in complex tasks and supports integrating domain expert evaluators alongside LLMs in the evaluation process to improve overall reliability. This paper is accepted into the ACM Conference on Intelligent User Interfaces (IUI) 2025 [21].

Study 3: Different Perspectives of Evaluation Criteria Development. A common approach to evaluating LLMs is to use metrics, or evaluation criteria, which are assertions used to assess performance that help ensure output alignment with domain-specific standards. While current approaches effectively create metrics tailored to developer needs, there is a gap in understanding evaluation criteria generated by domain experts compared to those generated by lay users and LLMs themselves. This study explores the alignment across the types of evaluation criteria created by different sources. We further investigate how the criteria-setting process evolves, analyzing the evaluation criteria when considering only the prompt *(a priori)* and after reading the output *(a posteriori)*. Our findings reveal complementary strengths: domain experts create instructional, fact-based criteria with long-term impact; lay users prioritize usability and clarity with short-term impact; and LLMs address immediate task requirements. We propose a staged evaluation workflow that incorporates the strengths of these sources to enhance trust, reliability, and alignment of LLM outputs with end-user needs and preferences. This work contributes to the literature by advancing the understanding of how different evaluators bring unique perspectives to the evaluation process and proposing strategies to optimize collaboration between domain experts, users, and LLMs in the evaluation of complex tasks. This paper is submitted to a conference awaiting review [19].

4 Proposed Future Work and Next Steps

The three studies discussed above have established that there are complementary strengths that can be leveraged to determine the most optimal solution for integrating different evaluation perspectives. Below are two proposed studies for my dissertation research.

Study 4: Comprehensive Evaluation Framework for LLMs. Building on my previous research that highlights the complementary strengths of domain experts, lay users, and LLMs, I propose a study that focuses on developing a comprehensive evaluation framework for LLM outputs in complex tasks. My prior studies have shown that different evaluator groups are most effective at specific stages of the LLM evaluation workflow, yet no established framework currently exists to guide the optimal integration of these groups. To effectively collect the necessary information, we will build on prior research to identify the methods for gathering key evaluation criteria specific to the task and domain at hand. This will be accomplished through structured interviews and evaluations, focusing on

the relevant stages of the process, including the pre-development phase, prompt engineering phase, and output evaluation phase. This approach will help developers determine what type of information is required from each evaluator group at each stage of the process and when their input is most valuable. Once the framework is developed, I will evaluate its effectiveness through case studies in multiple domains (such as healthcare, education, and mental health) and tasks (e.g., factual accuracy, user experience, ethical considerations).

Study 5: Development of Automated Evaluation Tool for LLMs. Following the development of a comprehensive evaluation framework, this study will explore the creation of an automated evaluation tool to streamline and optimize the LLM evaluation process. The aim is to design a system that uses the framework developed in the previous study to perform routine evaluation tasks, with LLMs handling preliminary assessments and identifying areas requiring further human review. By integrating expertise from domain experts and lay users at critical stages, the system can efficiently enhance the evaluation process. This tool will be used to benefit a developer who is looking to evaluate an LLM before deployment. As I am considering this work, I imagine that the tools will have to provide a "score" or alert to indicate when certain outputs require expert review. This research will contribute to the field by demonstrating how automated systems can work in tandem with human evaluators to enhance LLM performance, and ultimately improve the scalability and effectiveness of LLM evaluations for complex tasks.

References

[1] Muhammad Aurangzeb Ahmad, Ilker Yaramis, and Taposh Dutta Roy. 2023. Creating trustworthy llms: Dealing with hallucinations in healthcare ai. *arXiv preprint arXiv:2311.01463* (2023).

[2] Konstantinos Andriopoulos and Johan Pouwelse. 2023. Augmenting LLMs with Knowledge: A survey on hallucination prevention. *arXiv preprint arXiv:2309.16459* (2023).

[3] Ian Arawjo, Chelse Swoopes, Priyan Vaithilingam, Martin Wattenberg, and Elena L Glassman. 2024. ChainForge: A Visual Toolkit for Prompt Engineering and LLM Hypothesis Testing. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18.

[4] John W Ayers, Adam Poliak, Mark Dredze, Eric C Leas, Zechariah Zhu, Jessica B Kelley, Dennis J Faix, Aaron M Goodman, Christopher A Longhurst, Michael Hogarth, et al. 2023. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine* 183, 6 (2023), 589–596.

[5] Angeline Chatelan, Aurélien Clerc, and Pierre-Alexandre Fonta. 2023. ChatGPT and future artificial intelligence chatbots: what may be the influence on credentialed nutrition and dietetics practitioners? *Journal of the Academy of Nutrition and Dietetics* 123, 11 (2023), 1525–1531.

[6] Yida Chen, Aoyu Wu, Trevor DePodesta, Catherine Yeh, Kenneth Li, Nicholas Castillo Marin, Oam Patel, Jan Riecke, Shivam Raval, Olivia Seow, et al. 2024. Designing a Dashboard for Transparency and Control of Conversational AI. *arXiv preprint arXiv:2406.07882* (2024).

[7] Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *Journal of Artificial Intelligence Research* 77 (2023), 103–166.

[8] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964* (2020).

[9] Junfeng Jiao, Saleh Afroogh, Yiming Xu, and Connor Phillips. 2024. Navigating llm ethics: Advancements, challenges, and future directions. *arXiv preprint arXiv:2406.18841* (2024).

[10] Minsuk Kahng, Ian Tenney, Mahima Pushkarna, Michael Xieyang Liu, James Wexler, Emily Reif, Krystal Kallarackal, Minsuk Chang, Michael Terry, and Lucas Dixon. 2024. Llm comparator: Visual analytics for side-by-side evaluation of large language models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–7.

[11] Adam Tauman Kalai and Santosh S Vempala. 2024. Calibrated language models must hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*. 160–171.

[12] Tae Soo Kim, Yoonjoo Lee, Jamin Shin, Young-Ho Kim, and Juho Kim. 2024. Evallm: Interactive evaluation of large language model prompts on user-defined criteria. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–21.

[13] Fang Liu, Yang Liu, Lin Shi, Houkun Huang, Ruifeng Wang, Zhen Yang, and Li Zhang. 2024. Exploring and evaluating hallucinations in llm-powered code generation. *arXiv preprint arXiv:2404.00971* (2024).

[14] Adian Liusie, Vatsal Raina, Yassir Fathullah, and Mark Gales. 2024. Efficient LLM Comparative Assessment: a Product of Experts Framework for Pairwise Comparisons. *arXiv preprint arXiv:2405.05894* (2024).

[15] Qian Pan, Zahra Ashktorab, Michael Desmond, Martin Santillan Cooper, James Johnson, Rahul Nair, Elizabeth Daly, and Werner Geyer. 2024. Human-Centered Design Recommendations for LLM-as-a-Judge. *arXiv preprint arXiv:2407.03479* (2024).

[16] Malik Sallam. 2023. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. In *Healthcare*, Vol. 11. MDPI, 887.

[17] Souvika Sarkar, Mohammad Fakhruddin Babar, Monowar Hasan, and Shubhra Kanti Karmaker. 2024. LLMs as On-demand Customizable Service. *arXiv preprint arXiv:2401.16577* (2024).

[18] Shreya Shankar, JD Zamfirescu-Pereira, Björn Hartmann, Aditya G Parameswaran, and Ian Arawjo. 2024. Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences. *arXiv preprint arXiv:2404.12272* (2024).

[19] Annalisa Szymanski, Simret Araya Gebreegziabher, Oghenemaro Anuyah, Ronald A Metoyer, and Toby Jia-Jun Li. 2024. Comparing Criteria Development Across Domain Experts, Lay Users, and Models in Large Language Model Evaluation. *arXiv preprint arXiv:2410.02054* (2024).

[20] Annalisa Szymanski, Brianna L Wimer, Oghenemaro Anuyah, Heather A Eicher-Miller, and Ronald A Metoyer. 2024. Integrating Expertise in LLMs: Crafting a Customized Nutrition Assistant with Refined Template Instructions. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–22.

[21] Annalisa Szymanski, Noah Ziems, Heather A Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A Metoyer. 2024. Limitations of the LLM-as-a-Judge Approach for Evaluating LLM Outputs in Expert Knowledge Tasks. *arXiv preprint arXiv:2410.20266* (2024).

[22] Weixuan Wang, Barry Haddow, Alexandra Birch, and Wei Peng. 2023. Assessing the reliability of large language model knowledge. *arXiv preprint arXiv:2310.09820* (2023).

[23] Ian Webster. 2023. promptfoo: Test your prompts. https://www.promptfoo.dev/.