Blind Spot in Human-centered AI Evaluation

WILLEM VAN DER MADEN & JICHEN ZHU, IT University Copenhagen, Denmark

Large Language Models (LLMs) are increasingly being used in design practices—such as persona and scenario development, critical reflection, and prototyping—due to their ability to generate insights and support creative processes. However, while the integration of LLMs into human-centered design is expanding, a critical blind spot remains: there is a lack of methodologies for developing evaluation criteria that are truly human-centered and context-sensitive. Existing calls for improved LLM evaluation often focus on the need for better criteria, but they fall short of providing systematic support for creating these criteria based on real user experiences and needs. This position paper argues that to fill this gap, we should draw inspiration from established practices in human-centered design, which offers rich methods for eliciting criteria based on real-world user experiences. By exploring these avenues, we can begin to reframe AI evaluation through a design-oriented lens, guiding the development of LLMs in ways that are more aligned with human-centered design principles. Before AI can be integrated in design, design should be integrated in AI.

Additional Key Words and Phrases: artificial intelligence, human-centered design, evaluation, psychology, human-computer interaction

ACM Reference Format:
Willem van der Maden & Jichen Zhu. 2024. Blind Spot in Human-centered AI Evaluation. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Generative artificial intelligence systems (GenAI) are rapidly becoming part of the designer's toolkit. From generating personas [13] and scenarios [10] to supporting critical reflection [14] and prototyping [9], these AI systems are reshaping how we approach design challenges. Among these, Large Language Models (LLMs) are particularly influential, promising to enhance creativity, streamline processes, and foster innovative design solutions. However, this potential depends on how well these tools align with the specific needs of designers and end-users. Current evaluation methods, which often rely on generalized metrics like accuracy, fail to capture the nuanced and context-specific interactions in real-world design settings [7, 15, 16, 18]. To fully leverage LLMs for human-centered design, we must evaluate them with equally human-centered assessments. Without such methods, we risk deploying AI tools that may not truly serve their intended purpose, despite their technical sophistication. The current situation is akin to using a new design tool without any means to assess its impact on the design process or outcomes.

While there is growing recognition of the need for more human-centered LLM evaluation, existing research efforts often lack systematic frameworks for developing criteria grounded in real user experiences within design contexts—a challenge often referred to as the sociotechnical gap [8]. The unique properties of LLMs complicate the creation of such criteria more than traditional technologies [11]. Most current evaluation methods still focus on metrics like accuracy, which are suited for simpler, task-oriented models but fall short in capturing the nuanced ways designers and users interact with LLMs during creative processes [5, 17]. For instance, assessing an LLM's effectiveness in persona development is fundamentally different from evaluating its utility in narrative content generation. Therefore, there is an urgent need for evaluation approaches that can accommodate the varied applications of LLMs in real-world design scenarios.

This misalignment between evaluation criteria and real-world design applications is a significant bottleneck in developing effective AI tools for design. The quality of these criteria directly impacts the iterative training and refinement of LLMs. Inadequate evaluation metrics lead to poorly developed models, creating a cycle of suboptimal AI performance in design contexts. To bridge this gap, this paper proposes drawing from established HCD practices, which offer robust methods for eliciting criteria based on real-world user experiences. By leveraging these practices, we can reframe AI evaluation through a design-oriented lens, aligning the development of LLMs more closely with human-centered design principles. In essence, this paper argues that before we can effectively integrate AI into design, we must first integrate design principles into AI development. The structure of this paper is as follows: we will begin by addressing the challenges in current LLM evaluation practices, highlighting recent research and the gap in criteria development. Next, I will propose an approach inspired by scale development methodologies to address these challenges. Finally, we will discuss how HCD methods can adapt this approach to the specific needs of LLM evaluation in design contexts. To fully understand the need for this human-centered approach, we must first examine the challenges inherent in current LLM evaluation practices.

2 CHALLENGES IN CURRENT HUMAN-CENTERED EVALUATION PRACTICES

As LLMs are increasingly integrated into daily life, their impacts often emerge from complex interactions. Due to this entanglement, it is difficult to predict capabilities by evaluating them in isolation. The rapid advancement of LLMs creates a moving target for evaluators, challenging their ability to develop reliable assessments that accurately capture real-world performance, while trying to stay in tune with emerging applications[8, 16]. Moreover, the context-sensitivity of LLM outputs and their ability to engage in a wide range of open-ended interactions render traditional, task-specific evaluation metrics inadequate[11].

These characteristics of LLMs have contributed to what researchers term an "evaluation crisis" in AI [18]. Current techno-centric evaluation methods, which rely on generic benchmarks and automated assessments, fall short in capturing the real-world complexity of LLM use[11]. While current work is ongoing to develop more human-centered evaluation processes [e.g., 3, 7, 12], a clear gap remains: the current state-of-the-art is focused developing methods that better align with real-world scenarios and user intents; however, the criteria and metrics used to evaluate these more granular scenarios remain superficial (e.g., Enjoyment: did you like this interaction) and lacking rigor and scientific grounding (e.g., developed without proper investigation/whatever).

Taking this gap together with calls from recent literature to foster transparency [6], standardization [2], and construct validation [1], we may look to established methods from psychology to support the development of criteria and metrics for human-centered AI evaluation. These challenges highlight the need for a more systematic approach to developing evaluation criteria that can capture the nuanced interactions between users and LLMs. To address this, we can draw inspiration from established practices in scale development, while adapting them to the unique demands of AI evaluation.

3 CRITERIA ELICITATION AND OPERATIONALIZATION

Scale development in psychology typically involves several key steps: construct definition, item generation, content validation, scale administration, item analysis, and reliability and validity assessment [4]. Each step ensures the resulting scale accurately measures the intended construct. The item generation phase is crucial and involves eliciting and conceptualizing experiences within a specific context. For instance, when developing a scale for work-related stress, researchers might conduct interviews with employees to understand their experiences of stress in the workplace. This process helps identify relevant dimensions of the construct. Following elicitation, researchers operationalize these criteria into observable metrics. This step translates conceptual understanding into measurable items, allowing for quantitative assessment. For example, a work stress item might ask respondents to rate how often they feel overwhelmed by their workload on a Likert-scale.

However, the dynamic nature of LLMs presents unique challenges to this process. Traditional scale development is relatively static, with iterations occurring between versions but not after finalization. This approach does not align with the emergent nature of LLMs, whose capabilities can change rapidly and unpredictably. Moreover, scales typically require contextual consistency, which is difficult to achieve with LLMs due to their high context-dependency and versatility across different applications. **To address the unique characteristics of LLMs**, Liao & Xiao [8], propose learning from HCI evaluation methods as this field has grappled with similar challenges in assessing complex technologies. HCI offers various methods (e.g., field studies) for understanding *human-computer interactions* by examining how users experience them, and then converts these insights into quantifiable standards and measurements. Next, we will discuss a subset of HCI methods that we often see in human-centered design (HCD) practices. Namely, HCD is uniquely positioned to support these two critical stages, particularly for LLMs, due to its focus on understanding user experiences and operationalizing them into design requirements. In the next section, we will explore how HCD methods can contribute to developing more effective, context-sensitive, and human-centered evaluation criteria for LLMs, potentially addressing many of the challenges outlined here.

4 HCD CAN HELP

Leveraging user experiences is crucial for effectively eliciting and operationalizing evaluation criteria, as it grounds the process in real-world contexts and allows for the detection of subtle changes and nuances over time. By understanding how people experience interactions with LLMs, we move beyond evaluating these systems in isolation and instead assess them in the complex, dynamic environments where they are actually used. To explore this further, we will examine several qualitative methods commonly used in HCD and discuss how they can support the development of more nuanced, context-sensitive evaluation criteria for LLMs in various ways. For elicitation, several HCD methods can be employed:

- Cultural Probes: This method involves giving users kits with open-ended tasks (e.g., photo diaries, postcards) to capture their experiences with LLMs in their natural environments. For instance, designers could document their interactions with an LLM-powered design tool over a week, providing rich, contextual insights into the tool's impact on their creative process.
- Contextual Inquiry: Researchers can observe and interview users as they interact with LLMs in their typical work environment. This method could reveal nuanced aspects of LLM use in design tasks, such as how designers leverage LLM suggestions during ideation or how they negotiate between AI-generated and human-created content.
- Experience Sampling Method (ESM): This technique involves prompting users to provide brief reports on their experiences at random intervals. In the context of LLM evaluation, designers could be prompted to rate their satisfaction with LLM outputs or describe their emotional state during LLM interactions throughout their workday, capturing real-time, in-situ experiences.

For operationalization, HCD methods can facilitate the translation of elicited experiences into measurable criteria:

- Affinity Diagramming Workshops: Collaborative sessions where stakeholders categorize and prioritize insights from elicitation methods, helping to identify key dimensions for evaluation.
- Journey Mapping: Creating visual representations of user experiences with LLMs can highlight critical moments that should be captured in evaluation criteria.
- Participatory Design Sessions: Involving users in crafting evaluation questions or metrics ensures that the operationalized criteria resonate with real-world experiences.
- HCD methods also support ongoing iteration and refinement of evaluation criteria:
- Co-design Workshops: Regular sessions with users can help interpret evaluation data, uncovering new aspects of LLM interaction that weren't captured in earlier iterations.
- Retrospective Interviews: Periodic in-depth interviews with users about their evolving experiences with LLMs can reveal shifts in usage patterns or expectations, informing updates to evaluation criteria.
- Community Feedback Panels: Establishing ongoing dialogue with a diverse group of LLM users allows for continuous input on the relevance and effectiveness of evaluation criteria.

These HCD methods, when integrated into the evaluation development process, enable a more dynamic and responsive approach to LLM assessment. They facilitate the continuous refinement of criteria in response to evolving LLM capabilities and changing user needs, ensuring that evaluation methods remain relevant and effective over time.

5 CONCLUSIONS AND FUTURE WORK

The systematic integration of HCD methods into a scale development pipeline for AI evaluation offers multiple advantages that address core challenges in current practices. This approach enhances the relevance and validity of evaluations by aligning criteria with specific contexts of LLM use, ensuring more meaningful and actionable insights. It democratizes access to sophisticated evaluation techniques, supporting practitioners and non-expert researchers across various sectors in developing context-sensitive criteria. **In short, to integrate AI into HCD practices, we must first integrate HCD into AI evaluation practices.** In conclusion, we see several key areas warrant further investigation:

(1) **Methodology Refinement**: Future work should focus on developing and testing specific methodologies that combine HCD techniques with scale development processes. This could involve creating step-by-step guides or frameworks that researchers and practitioners can follow.
(2) **Cross-Domain Applicability**: Research is needed to explore how this integrated approach can be adapted for different domains beyond design, such as healthcare, education, or finance, where LLMs are increasingly being deployed.
(3) **Longitudinal Studies**: Long-term studies should be conducted to assess the effectiveness of this approach in capturing the evolving nature of LLM capabilities and user experiences over time.
(4) **Ethical Considerations**: Further investigation is required into how this approach can be used to develop evaluation criteria that specifically address ethical concerns in AI, such as bias, fairness, and transparency.
(5) **Scalability and Efficiency**: Research should explore ways to streamline and potentially automate parts of this process to make it more accessible and efficient for widespread adoption.
(6) **Comparative Studies**: Future work could compare the effectiveness of this integrated approach with traditional evaluation methods to quantify its benefits and identify areas for improvement.
(7) **Tool Development**: There's potential for developing software tools or platforms that facilitate the implementation of this integrated approach, making it easier for non-experts to apply these methods.

REFERENCES

[1] John Burden. 2024. Evaluating AI Evaluation: Perils and Prospects. https://doi.org/10.48550/arXiv.2407.09221 arXiv:2407.09221 [cs].

[2] Ryan Burnell, Wout Schellaert, John Burden, Tomer D. Ullman, Fernando Martinez-Plumed, Joshua B. Tenenbaum, Danaja Rutar, Lucy G. Cheke, Jascha Sohl-Dickstein, Melanie Mitchell, Douwe Kiela, Murray Shanahan, Ellen M. Voorhees, Anthony G. Cohn, Joel Z. Leibo, and Jose Hernandez-Orallo. 2023. Rethink reporting of evaluation results in AI. *Science* 380, 6641 (April 2023), 136–138. https://doi.org/10.1126/science.adf6369 Publisher: American Association for the Advancement of Science.

[3] Michael Desmond, Zahra Ashktorab, Qian Pan, Casey Dugan, and James M. Johnson. 2024. EvaluLLM: LLM assisted evaluation of generative outputs. In *Companion Proceedings of the 29th International Conference on Intelligent User Interfaces*. ACM, Greenville SC USA, 30–32. https://doi.org/10.1145/3640544.3645216

[4] Robert F. DeVellis and Carolyn T. Thorpe. 2021. *Scale Development: Theory and Applications*. SAGE Publications. Google-Books-ID: QddDEAAAQBAJ.

[5] Oliver Farnsworth and George Attwell. 2024. Enhanced Problem Solving with Large Language Models: Integrating Trial-and-Error and Chain of Thought Methods. https://doi.org/10.36227/techrxiv.172055569.99689115/v1

[6] Michael Feffer, Anusha Sinha, Wesley Hanwen Deng, Zachary C. Lipton, and Hoda Heidari. 2024. Red-Teaming for Generative AI: Silver Bullet or Security Theater? http://arxiv.org/abs/2401.15897 arXiv:2401.15897 [cs].

[7] 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. http://arxiv.org/abs/2212.09746 arXiv:2212.09746 [cs].

[8] Q. Vera Liao and Ziang Xiao. 2023. Rethinking Model Evaluation as Narrowing the Socio-Technical Gap. http://arxiv.org/abs/2306.03100 arXiv:2306.03100 [cs].

[9] Yimeng Liu and Misha Sra. 2024. DanceGen: Supporting Choreography Ideation and Prototyping with Generative AI. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*. 920–938.

[10] Viktor Malakuczi, Mariya Ershova, Andrea Gentile, Cristina Gironi, Michele Saviano, and Luca Imbesi. 2024. Design in dialogue: AI as an aid of imagination for future scenarios. In *DRS2024: Boston, 23–28 June*, Caroline Gray, Eleni Ciliotta Chehade, Paul Hekkert, Laura Forlano, Paolo Ciuccarelli, and Peter Lloyd (Eds.). Boston, USA. https://doi.org/10.21606/drs.2024.1171

[11] Timothy R. McIntosh, Teo Susnjak, Tong Liu, Paul Watters, and Malka N. Halgamuge. 2024. Inadequacies of Large Language Model Benchmarks in the Era of Generative Artificial Intelligence. http://arxiv.org/abs/2402.09880 arXiv:2402.09880 [cs].

[12] 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. http://arxiv.org/abs/2407.03479 arXiv:2407.03479 [cs].

[13] Joongi Shin, Michael A. Hedderich, Bartłomiej Jakub Rey, Andrés Lucero, and Antti Oulasvirta. 2024. Understanding Human-AI Workflows for Generating Personas. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference* (Copenhagen, Denmark) *(DIS '24)*. Association for Computing Machinery, New York, NY, USA, 757–781. https://doi.org/10.1145/3643834.3660729

[14] Aswathy Sreenivasan and M Suresh. 2024. Design Thinking and Artificial Intelligence: A Systematic Literature Review Exploring Synergies. *International Journal of Innovation Studies* (2024).

[15] Jiayin Wang, Fengran Mo, Weizhi Ma, Peijie Sun, Min Zhang, and Jian-Yun Nie. 2024. A User-Centric Benchmark for Evaluating Large Language Models. http://arxiv.org/abs/2404.13940 arXiv:2404.13940 [cs].

[16] Laura Weidinger, Joslyn Barnhart, Jenny Brennan, Christina Butterfield, Susie Young, Will Hawkins, Lisa Anne Hendricks, Ramona Comanescu, Oscar Chang, Mikel Rodriguez, Jennifer Beroshi, Dawn Bloxwich, Lev Proleev, Jilin Chen, Sebastian Farquhar, Lewis Ho, Iason Gabriel, Allan Dafoe, and William Isaac. 2024. Holistic Safety and Responsibility Evaluations of Advanced AI Models. http://arxiv.org/abs/2404.14068 arXiv:2404.14068 [cs].

[17] Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, and William Isaac. 2023. Sociotechnical Safety Evaluation of Generative AI Systems. http://arxiv.org/abs/2310.11986 arXiv:2310.11986 [cs].

[18] Ziang Xiao, Wesley Hanwen Deng, Michelle S. Lam, Motahhare Eslami, Juho Kim, Mina Lee, and Q. Vera Liao. 2024. Human-Centered Evaluation and Auditing of Language Models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–6. https://doi.org/10.1145/3613905.3636302