

“I Want to Be Pushed, I Want to Grow”: Enabling Social Workers to Design Evaluations of LLM Augmentation in Their Work (Extended Abstract)

Anna Kawakami¹, Chloe Qianhui Zhao¹, Renee Shelby²,
Fernando Diaz¹, Haiyi Zhu¹, Kenneth Holstein³

¹Carnegie Mellon University

²Google Research

³EPFL

akawakam@andrew.cmu.edu, qianhuiz@andrew.cmu.edu, reneeshelby@google.com,
diazf@cmu.edu, haiyiz@andrew.cmu.edu, ken.holstein@epfl.ch

Extended version – <https://drive.google.com/file/d/14afU70spagIF9Wbi-v3phVjqBs-4Hfi-/view?usp=sharing>

Extended Abstract

Frontline workers across occupations—from teaching and social work, to healthcare and finance—are increasingly asked to adopt AI systems to assist their work (Brynjolfsson, Li, and Raymond 2025). Yet, workers themselves rarely have a voice in defining what it would look like for AI to positively augment their day-to-day work, and how to design and evaluate for it (Kawakami et al. 2026). Today, decisions about which parts of workers’ work to augment, how to augment them, and how to measure what “good performance” looks like are often left to technical experts or organizational leaders. Without worker involvement in these decisions, AI evaluations often reflect impoverished understandings of workers’ labor and the expertise involved (Yang, Steinfeld, and Zimmerman 2019; Yin et al. 2026).

A growing literature focuses on developing AI evaluations for work (Budler et al. 2025), including through new benchmarks and metrics to measure the automatability of specific work tasks (e.g., (Vishwakarma et al. 2025; Bedi et al. 2025; Wang et al. 2025)). More recently, ML and NLP researchers have begun to recognize the need to involve workers themselves in shaping these AI evaluations (Bedi et al. 2025; Arora et al. 2025; Gebreegziabher et al. 2025), and have explored methods to solicit worker feedback to do so. However, this work has typically constrained worker involvement to narrow sub-activities such as developing test cases or providing expert labels for benchmarks, without engaging workers in larger decisions (e.g., what concepts are worth measuring in the first place?) (Kawakami, Zhu, and Holstein 2026). By contrast, HCI research has solicited workers’ perspectives on how new LLM-based systems should augment their professions in the future (e.g., Brachman et al. 2025; Woodruff et al. 2024), but has not directly engaged them in designing concrete measurement instruments. **Taken together, it remains an open question how to meaningfully support workers in the end-to-end, collaborative design of AI measurement instruments—from defining what pos-**

itive AI augmentation, to *operationalizing* practical measurement instruments that can capture these notions.

This paper presents a case study with a local school social work organization that is actively exploring how to integrate LLMs into their work. In partnership with workers, we explore an approach to *worker-driven AI measurement design*—where workers collaboratively shape decisions around *which tasks* AI should augment, *what* “successful” augmentation looks like for those tasks, and *how* it should be measured. Through an iterative series of eight workshops with 19 workers, we explore how to engage workers in (1) identifying realistic LLM use cases for which workers actually want LLM-based support, (2) developing a shared understanding of what “good” or “bad” LLM performance looks like for a use case, and (3) operationalizing this understanding through the design of a benchmark (including an LLM-as-a-judge system and a set of test cases).

Overall, workers collaboratively developed an LLM-as-a-judge system to evaluate model performance on a specific use case central to their work: developing useful reflection questions to guide future meetings with teachers, based on workers’ classroom observations. Workers first collaboratively designed 16 realistic test cases that they believed would be challenging for an LLM, based on their past experience using LLMs. In early workshops, workers identified a measurement goal that is largely absent in existing industry AI evaluations: measuring how effectively LLMs can function as “**people challengers**” (a play on current LLM tendencies to be “people pleasers”). In the context of their work, this meant challenging workers to “*reflect on their own possible assumptions and biases [and helping] them expand their perspectives, providing opportunities for learning and growth.*” Workers iteratively operationalized this goal into an LLM-as-a-judge rubric. Our validation shows that the resulting LLM-as-a-judge system has strong agreement with worker judgments, and the benchmark can differentiate the performance of different state-of-the-art LLMs.

This paper contributes (1) **worker-driven AI measurement**—a bottom-up, collaborative design approach that complements existing evaluation practices for AI in work, (2) an early **case study** exploring the approach with a local social work organization, and (3) **opportunities for future work** on supporting worker-driven AI measurement design.

References

- Arora, R. K.; Wei, J.; Hicks, R. S.; Bowman, P.; Quiñonero-Candela, J.; Tsimpourlas, F.; Sharman, M.; Shah, M.; Val-lone, A.; Beutel, A.; et al. 2025. HealthBench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*.
- Bedi, S.; Cui, H.; Fuentes, M.; Unell, A.; Wornow, M.; Banda, J. M.; Kotecha, N.; Keyes, T.; Mai, Y.; Oez, M.; et al. 2025. MedHELM: Holistic evaluation of large language models for medical tasks. *arXiv preprint arXiv:2505.23802*.
- Brachman, M.; El-Ashry, A.; Dugan, C.; and Geyer, W. 2025. Current and future use of large language models for knowledge work. *Proceedings of the ACM on Human-Computer Interaction*, 9(7): 1–24.
- Brynjolfsson, E.; Li, D.; and Raymond, L. 2025. Generative AI at work. *The Quarterly Journal of Economics*, 140(2): 889–942.
- Budler, L. C.; Chen, H.; Chen, A.; Topaz, M.; Tam, W.; Bian, J.; and Stiglic, G. 2025. A brief review on benchmarking for large language models evaluation in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 15(2): e70010.
- Gebreegziabher, S. A.; Chiang, C.; Wang, Z.; Ashktorab, Z.; Brachman, M.; Geyer, W.; Li, T. J.-J.; and Gómez-Zarà, D. 2025. MetricMate: An interactive tool for generating evaluation criteria for LLM-as-a-judge workflow. In *Proceedings of the 4th Annual Symposium on Human-Computer Interaction for Work*, 1–18.
- Kawakami, A.; Taylor, J.; Fox, S.; Zhu, H.; and Holstein, K. 2026. AI failure loops in devalued work: The confluence of overconfidence in AI and underconfidence in worker expertise. *Big Data & Society*, 13(1): 20539517261424164.
- Kawakami, A.; Zhu, H.; and Holstein, K. 2026. AI measurement as a collaborative design practice: A multidisciplinary synthesis and future directions. *Available at SSRN 6276543*.
- Vishwakarma, H.; Agarwal, A.; Patil, O.; Devaguptapu, C.; and Chandran, M. 2025. Can LLMs help you at work? A Sandbox for evaluating LLM agents in enterprise environments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 9178–9212.
- Wang, Z. Z.; Shao, Y.; Shaikh, O.; Fried, D.; Neubig, G.; and Yang, D. 2025. How do AI agents do human work? comparing AI and human workflows across diverse occupations. *arXiv preprint arXiv:2510.22780*.
- Woodruff, A.; Shelby, R.; Kelley, P. G.; Rousso-Schindler, S.; Smith-Loud, J.; and Wilcox, L. 2024. How knowledge workers think generative AI will (not) transform their industries. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–26.
- Yang, Q.; Steinfeld, A.; and Zimmerman, J. 2019. Unremarkable AI: Fitting intelligent decision support into critical, clinical decision-making processes. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, 1–11.
- Yin, Y.; Karumbaiah, S.; Kim, Y.; and Saxena, D. 2026. When AI breaks, teachers repair: Pedagogical repair work in situated classroom practice. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency*, 2206–2224.