Detecting Experiential Differences Between LLM Versions Using Psychometric Scales: a Journaling Case Study

Willem van der Maden
IT University of Copenhagen
HCI & Design Section
Denmark
wiva@itu.dk

Pavel Okopnyi
University of Bergen
Department of Information
Science and Media Studies
Norway

Frode Guribye
University of Bergen
Department of Information
Science and Media Studies
Norway

Simone Grassini
University of Bergen
Department of Psychosocial Science
Norway

Jichen Zhu
IT University of Copenhagen
HCI & Design Section
Denmark

Abstract

This short paper investigates the use of psychometric scales to evaluate user experience with Large Language Model (LLM) apps. LLM evaluations have predominantly focused on technical benchmarks measuring capabilities, leaving out the quality of user experience with these apps. Using the domain of the digital journaling, this paper presents a user study (n = 39) with an app powered by two versions of the LLM. The results show that the psychometric scales can detect subtle but consistent differences between the two versions, particularly in interpersonal dimensions like relatedness. This initial investigation suggests that psychological scales may be useful tools for detecting experiential differences between LLM versions in domain-specific apps.

CCS Concepts

• **Human-centered computing** → **Human computer interaction (HCI)**; *Empirical studies in HCI*; User studies; • **Computing methodologies** → *Natural language processing*.

Keywords

large-language models, evaluation, explainability, user experience, psychometrics

ACM Reference Format:
Willem van der Maden, Pavel Okopnyi, Frode Guribye, Simone Grassini, and Jichen Zhu. 2025. Detecting Experiential Differences Between LLM Versions Using Psychometric Scales: a Journaling Case Study. In *Proceedings of ACM CHI2025 Workshop HEAL (CHI'25).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

CHI'25, April 26–May 1, 2025, Yokohama, Japan

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Large Language Models (LLMs) are increasingly integrated into app domains used by the general public, from writing assistants and customer service platforms to healthcare diagnostics and educational tools. As these LLM apps mature technically, there is a growing recognition that their effectiveness also depends on the quality of user experience they provide [39, 62, 66].

Evaluation plays an important role in the development cycle of LLMs as the results are frequently used to guide the iterative development of how these models should be improved. So far, LLM evaluations have predominantly focused on technical benchmarks measuring capabilities such as reasoning, translation, and problem-solving [7, 32]. With few exceptions [e.g., 10, 35], the quality of users' experience with LLMs have not yet been assessed with the specific aim to guide future LLMs development, along with technical benchmarks. As a result, there is an urgent need for methods and empirical results in human-centered AI research to assess user experience with LLMs—including not only interaction-level features (e.g., UI design and usability) but also the emotional, cognitive, and reflective dimensions of user experience—especially in ways that can be used to direct the training of LLMs.

In this paper, we investigate the potential of psychometric scales—validated instruments from psychological research—to evaluate how users experience interacting with LLMs. While established scales like the System Usability Scale (SUS) [3], User Experience Questionnaire (UEQ) [33], and NASA Task Load Index (NASA-TLX) [22] are rigorously validated and measure important aspects of user experience, recent research suggests they are inadequate for capturing the complex dynamics of human-LLM interaction [31, 58]. Psychometric scales offer a way to assess deeper psychological dimensions like mental states and emotional experiences, and have proven valuable in Human-Computer Interaction (HCI) and Explainable AI research for measuring user experience [9, 42]. Recently, they have been increasingly used to assess how well LLM apps[1] can fulfill certain

[1] "LLM apps" refers to apps where LLMs serve as the primary technology powering the core functionality (e.g., ChatGPT, GitHub Copilot, Claude). While our findings may be relevant for apps that incorporate LLMs as supplementary features (e.g., chat support on websites), this paper focuses on evaluating user experiences with primary LLM apps.

purposes [e.g., 19, 20, 29, 38, 43, 48, 55, 59, 65]. These "fit for purpose"[2] evaluations are important to establish baseline effectiveness of LLM apps. However, to meaningfully compare and improve LLM apps, we need psychometric scales sensitive enough to detect experiential differences between versions. Currently, there is a lack of empirical evidence on the feasibility of using such scales to differentiate user experiences, which is essential for advancing our understanding of evaluating and iteratively improving LLM apps in human-facing app domains [13].

Specifically, this paper presents our initial results in the domain of LLM apps for journaling. Our research question is: **If, and to what extent can the selected psychometric scales differentiate the user experiences of a journaling app run on the two versions of LLM?** We conducted a user study (n = 39) using a custom-made journaling app (called *Journal Kernel)* powered by two versions of the LLM, *Claude*.[3] Four existing validated scales measured core user experience dimensions (reflection, emotional awareness, motivation, and basic psychological needs). Our findings show that these scales detect subtle but consistent differences between the two LLM versions. This provides preliminary evidence that psychometric measures may effectively capture incremental changes in an LLM app for journaling. Building on these results, we highlight the potential of such instruments to guide the iterative development of LLM apps and outline several directions for future work, including exploring generalizability across domains, improving scale sensitivity, and integrating psychometric evaluation into real-world development cycles.

2 Related Work

LLMs are frequently evaluated using technical metrics like fluency or factual correctness [7, 32], yet recent deployments in creative writing, healthcare, and education have exposed the limitations of such benchmarks for understanding real-world user experiences [18, 42]. Researchers have thus begun integrating human-centered frameworks focused on satisfaction, usability, and trust [10, 15, 30, 35, 47, 60], but largely overlook deeper psychological outcomes like reflection, emotional awareness, or user motivation. Psychometric scales, which have been used primarily to study LLMs' traits (e.g., personality) [12, 41, 49], may fill this gap by assessing how people actually feel when interacting with evolving LLMs [17, 26, 40]. If they reliably capture incremental shifts in user experience introduced by new versions or prompts, product teams gain a practical tool for identifying unanticipated harms and refining human–LLM interactions [13].

One domain where such in-depth psychometric evaluation is especially pertinent is journaling, an introspective practice closely tied to personal wellbeing [25, 36, 67]. As LLM-powered tools provide reflective scaffolding and dynamic prompts, past work (e.g., MindfulDiary [28] and MindScape [46]) has shown that clinical and standardized surveys can confirm whether these systems "work" for mental health or reflection. Yet they have not examined subtle experiential changes that arise when newer LLMs replace older ones. Overlooking such nuances may risk deploying systems that inadvertently harm users [1, 2, 53, 54, 64]. By contrast, our focus is on whether standard psychometric instruments—validated for reflection, motivation, and psychological needs—can differentiate user experiences across LLM versions. This lens positions journaling as an useful testbed for refining evaluation techniques in a truly human-centered manner.

[2]The term *fit-for-purpose* was first used in biomedical research, where it denotes context-specific evaluations of tools or methods [34] and has since been adapted to other fields such as technology design [23], and educational sciences [27].
[3]Available at https://claude.ai

3 Method

We conducted a controlled experimental comparison of two LLM journaling app versions to investigate whether standard psychometric scales can detect meaningful differences in user experience.

3.1 Participants

We recruited 39 participants (19 female, 20 male; age range 20–67, $M = 34.23$, $SD = 12.80$) via Prolific. Most (87%) reported prior AI chatbot experience (e.g., ChatGPT). Participants were randomly assigned to *Claude 2.0* ($n = 17$) or *Claude 3.5 Sonnet* ($n = 22$).

3.2 Journaling App

We developed *Journal Kernel*, a web-based platform that uses a chat-style interface to guide reflective writing (see Figure 1). Both LLM versions had identical Positive Affect Journaling (PAJ) prompts [56], and only anonymized survey data was collected. A pilot study (12 participants) helped refine the interface and prompts.

3.3 Procedure

After providing informed consent, participants chose one of the PAJ prompts (e.g., "Reflect on a meaningful moment that brought you joy recently") and engaged in an open-ended journaling conversation with the assigned LLM. The system responded autonomously based on user input, encouraging deeper reflection. Each participant then summarized their thoughts and completed an online survey. Sessions averaged 9.3 minutes, consistent with other LLM journaling studies [28].

3.4 Measures

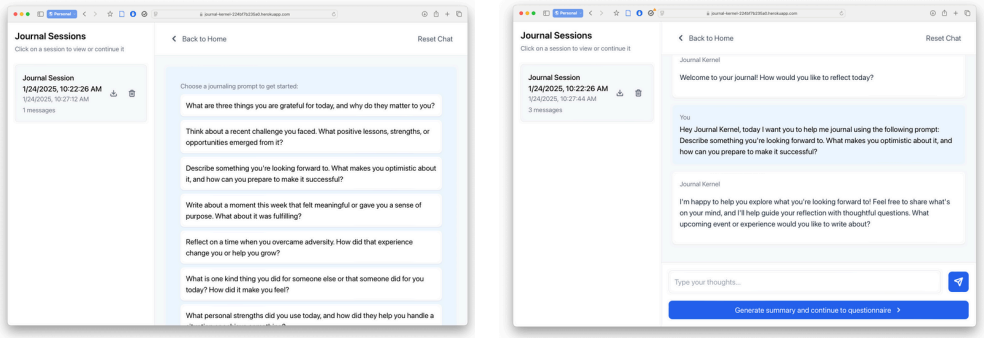
We used 24 items from established scales commonly applied in HCI journaling research [e.g., 25, 28, 36, 46, 67]. Each item was rated on a 5-point Likert scale, with higher scores indicating stronger agreement:

- **Reflection quality** (5 items): Depth of reflection (e.g., "I reconsidered my previous beliefs") [57].
- **Emotional awareness** (4 items): Ability to identify/process emotions (e.g., "I recognized my feelings") [50, 51].
- **Motivation** (4 items): Enjoyment and effort (e.g., "I enjoyed this writing session") [44, 52].
- **Basic psychological needs** (12 items): Autonomy, competence, and relatedness [4].

These short-form scales minimized participant burden while maintaining validity. Data collection conformed to national ethics guidelines and was conducted anonymously.

4 Results

Our analysis revealed consistent differences between the two Claude versions, with Claude 3.5 Sonnet scoring 0.1–0.3 points higher across all measured dimensions (see Figure 2). All scales demonstrated acceptable internal consistency (Cronbach's α = 0.78–0.87). Using independent two-sample t-tests and Cohen's d effect-size measures,



(a) (b)

Figure 1: Screenshots of the *Journal Kernel* interface: (a) prompt selection; (b) interactive journaling session.

we found that while differences did not reach conventional significance thresholds ($p < 0.05$), they consistently favored Claude 3.5 Sonnet. The most pronounced gains appeared in interpersonal aspects, particularly relatedness (+0.15) and autonomy (+0.27), while task-focused measures showed smaller but consistent improvements. Notably, examination of response distributions revealed fewer very low ratings (1 or 2) for Claude 3.5 Sonnet compared to Claude 2.0, suggesting it may reduce the likelihood of negative user experiences. Next, we detail the specific findings for each psychometric scale.

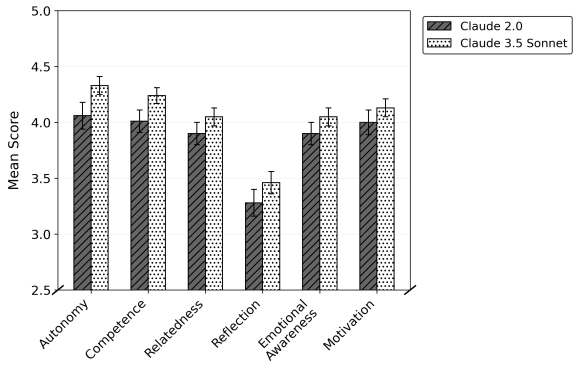


Figure 2: Mean user experience scores (±SE) for Claude 3.5 Sonnet and Claude 2.0 across six psychometric scales. The y-axis begins at 2.5 to emphasize differences in scores and avoid excessive white space, as the average scores across scales for both models are above 3.0—a baseline expected of advanced language models.

Basic Psychological Needs. Participants using Claude 3.5 Sonnet consistently reported higher satisfaction of basic psychological needs compared to those using Claude 2.0. The largest difference was observed in the *relatedness* subscale, with Claude 3.5 Sonnet scoring an average of 4.05 (±0.08 SE) compared to 3.90 (±0.10 SE) for Claude 2.0 ($t(37) = 1.68, p = 0.09, d = 0.34$). Similar trends were observed for *autonomy* ($t(37) = 1.77, p = 0.08, d = 0.35$) and *competence* ($t(37) = 1.69, p = 0.09, d = 0.33$). These results suggest that Claude 3.5 Sonnet provides small but potentially meaningful experiential benefits in interpersonal dimensions.

Reflection Quality. Reflection quality scores showed moderate differences between versions, with Claude 3.5 Sonnet scoring an average of 3.46 (±0.10 SE) compared to 3.28 (±0.12 SE) for Claude 2.0 ($t(37) = 0.59, p = 0.55, d = 0.12$). Notably, both versions received lower absolute ratings on this scale compared to others, indicating room for improvement in supporting deep reflection across both models.

Emotional Awareness. Participants' scores for Emotional Awareness averaged 4.05 (±0.08 SE) for Claude 3.5 Sonnet and 3.90 (±0.10 SE) for Claude 2.0 ($t(37) = 0.64, p = 0.52, d = 0.13$). While the differences are modest, they suggest that Claude 3.5 Sonnet may offer slightly better support for recognizing and processing emotions during journaling. Further research is needed to explore how LLMs can more effectively facilitate deeper emotional engagement.

Motivation. Motivation scores were 4.13 (±0.08 SE) for Claude 3.5 Sonnet and 4.00 (±0.11 SE) for Claude 2.0 ($t(37) = 0.53, p = 0.59, d = 0.11$). These results indicate small but consistent improvements in participants' enjoyment and willingness to engage with the journaling task when using Claude 3.5 Sonnet. While encouraging, this effect was less pronounced than for other measures, suggesting opportunities for further optimization of user engagement.

5 Discussion

Through a controlled experiment comparing two versions of an LLM journaling app, we observed initial evidence suggesting that standard psychological scales might detect subtle differences in how users experience an LLM-supported journaling conversation. This preliminary investigation explored the sensitivity of psychometric instruments in human-LLM evaluation contexts, examining different dimensions of user experience.

The results indicate consistent but modest differences across all measured dimensions, with Claude 3.5 Sonnet scoring slightly higher than Claude 2.0. The largest differences appeared in interpersonal aspects, such as relatedness (+0.15), and basic psychological needs like autonomy (+0.27) and competence (+0.21). Task-focused measures, including reflection quality and motivation, showed smaller changes, suggesting potential variation in how different scales capture experiential shifts. While the limited sample size and statistical power prevent us from drawing definitive conclusions, the overall pattern of higher scores for Claude 3.5 Sonnet suggests that psychometric

scales hold promise as tools for detecting nuanced differences in user experiences between LLM versions.

5.1 Implications

Our study offers several preliminary insights into the use of psychometric scales for evaluating human-LLM interaction. Below are the key implications for LLM evaluation, product development, and HCI research.

For LLM Evaluation. The consistent pattern of differences between versions, while modest, suggests that psychological scales commonly used in HCI journaling studies can likely detect experiential changes in LLM-supported journaling conversations. This finding extends current approaches to LLM evaluation, which face ongoing challenges in transparency, reproducibility, and domain-specificity [5, 32, 37, 45]. Addressing these challenges is particularly important as existing frameworks often rely on simplified satisfaction metrics or unreliable self-assessments [7, 68]. In contrast, psychometric scales offer potential advantages through their standardized measurement protocols, established construct validity, and transparent interpretive criteria. While existing frameworks often rely on simplified metrics or unreliable self-assessments [16], psychometric scales offer potential advantages through their standardized measurement protocols and established construct validity. Considering that LLMs may require entirely new evaluation approaches [63], we start with existing psychological scales, and depending on future research results, there may be a need for new scales that are more sensitive to human-LLM interaction.

For Product Development. Our findings suggest that established HCI measurement scales can potentially be adopted 'off-the-shelf' for evaluating LLM apps—that is, using existing scales without modifying them for LLM-specific interactions. While our results are specific to journaling with Claude models and cannot be generalized across domains or other LLMs, the fact that these unmodified scales detected differences suggests that this might be worth exploring further. The ability to use existing scales provides a practical solution for practitioners who need accessible tools for developing and evaluating domain-specific LLM apps [10]. In other words, while standardized frameworks for human-LLM interaction are still in development [e.g., 10, 24, 35], practitioners can draw upon existing HCI research in similar app domains to inform their evaluations. These established measures can complement emerging LLM-specific evaluation approaches. Most psychological scales assess multiple dimensions of a phenomenon to achieve construct validity [6, 8]. This multidimensional approach enables developers to more precisely attribute user experience changes to specific aspects of their system, offering more granular insights than general satisfaction metrics [14, 61]. For instance, while a prompt engineering change might improve overall satisfaction, these scales could reveal which specific experiential dimensions were affected.

For HCI Research. Our study extends prior work suggesting HCI measures could inform LLM evaluation [42] by providing initial empirical evidence for their utility. Just as HCI evolved from technical evaluation to studying situated experiences [21], LLM evaluation needs a similar transformation. The CHI community's expertise in empirical studies of technology use and impact is particularly valuable here. Rather than focusing solely on technical capabilities, we need more research examining how specific LLM apps affect users in real-world contexts. This study demonstrates one approach to such empirical work, but many opportunities remain for HCI researchers to shape how we evaluate and understand human-LLM interaction.

5.2 Limitations

This study used a simple between-subjects design, which limited our control over individual differences and exposure to only one model version per participant. In future work, we plan to implement a more robust mixed methods approach combining between and within-subjects comparisons using a Latin square design with appropriate control groups to better isolate effects. Our randomization procedure resulted in uneven group sizes (n=17 and n=22), partly due to participant attrition and reintroduction, which represents a methodological limitation. We tested only two specific versions of Claude, which constrains generalizability to other LLM architectures or release cycles. Additionally, our experimental design contained too many degrees of freedom, potentially diluting our ability to detect significant effects. In future studies, we will exercise tighter experimental control and compare models with proven experiential differences, such as those identified through community-driven leaderboards. Each participant only completed a single journaling session, so we could not assess how user perceptions evolve with prolonged usage. Lastly, our findings reflect the timeframe and recruitment choices (e.g., using Prolific), meaning we could not conduct planned follow-up interviews due to anonymity requirements. As a result, we cannot definitively attribute observed differences to any one factor, such as model improvements versus interface updates.

6 Conclusion & Future Work

This short paper discussed a pilot study examining whether standard psychometric scales, commonly used in HCI research, could detect experiential differences between LLM versions. While our results show modest effects as expected in an initial exploration, they suggest potential value in using these scales not just for baseline effectiveness evaluation, but for comparative assessment that could guide iterative development. This preliminary work provides a foundation for refining our approach in future studies with larger samples and more controlled designs.

Our future work focuses on three key directions: (1) investigating generalizability across different LLM architectures (e.g., variations in model scale, such as parameter size differences between LLaMA 3.1 8B and 405B [11]), prompting strategies (e.g., task-specific prompts designed to achieve concrete objectives versus vibe-specific prompts aimed at setting a particular tone, style, or emotional resonance), domains (e.g., critical thinking, educational tutoring), and user populations (e.g., across diverse cultural backgrounds and demographics), (2) integrating these measurement approaches into actual LLM development cycles to understand their practical impact using a local case study, and (3) examining how this psychometric evaluation approach relates to broader LLM evaluation challenges, particularly its limitations and complementarity with issues like scalability. Through this work, we aim to support more effective evaluation of LLM-human interaction as these systems become increasingly integrated into daily life, affecting real user experiences and wellbeing.

References

[1] A. N. Author and B. Researcher. 2023. How and why psychologists should respond to the harms associated with generative AI. *Nature Communications Psychology* 1, 1 (2023), 1–15. https://doi.org/10.1038/s44271-023-00001-2

[2] C. D. Author and E. F. Scholar. 2023. The mental health implications of artificial intelligence adoption. *Nature Human Behaviour* 7, 8 (2023), 1203–1215. https://doi.org/10.1038/s41562-023-01651-4

[3] John Brooke. 1996. SUS: A quick and dirty usability scale. *Usability Evaluation in Industry* 189, 194 (1996), 4–7.

[4] Ryan Burnell, Dorian Peters, Richard M Ryan, and Rafael A Calvo. 2023. Technology evaluations are associated with psychological need satisfaction across different spheres of experience: an application of the METUX scales. *Frontiers in Psychology* 14 (2023), 1092288. https://doi.org/10.3389/fpsyg.2023.1092288

[5] 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.

[6] Donald T Campbell and Donald W Fiske. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin* 56, 2 (1959), 81.

[7] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A Survey on Evaluation of Large Language Models. *ACM Transactions on Intelligent Systems and Technology* 15, 3 (March 2024), 39:1–39:45. https://doi.org/10.1145/3641289

[8] Lee Anna Clark and David Watson. 2016. Constructing validity: Basic issues in objective scale development. (2016).

[9] Teresa Datta and John P. Dickerson. 2023. Who's Thinking? A Push for Human-Centered Evaluation of LLMs using the XAI Playbook. http://arxiv.org/abs/2303.06223 arXiv:2303.06223 [cs].

[10] 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

[11] Hugging Face. 2024. Llama 3.1 - 405B, 70B & 8B with multilinguality and long context. (2024). https://huggingface.co/blog/llama31 Accessed: 2025-01-23.

[12] Qixiang Fang, Daniel L. Oberski, and Dong Nguyen. 2024. PATCH! Psychometrics-AssisTed benCHmarking of Large Language Models: A Case Study of Proficiency in 8th Grade Mathematics. https://doi.org/10.48550/arXiv.2404.01799 arXiv:2404.01799.

[13] K. J. Kevin Feng, Q. Vera Liao, Ziang Xiao, Jennifer Wortman Vaughan, Amy X. Zhang, and David W. McDonald. 2024. Canvil: Designerly Adaptation for LLM-Powered User Experiences. arXiv:2401.09051 [cs.HC] https://arxiv.org/abs/2401.09051

[14] Steven Fokkinga, Pieter Desmet, and Paul Hekkert. 2020. Impact-centered design: Introducing an integrated framework of the psychological and behavioral effects of design. *International Journal of Design* 14, 3 (2020), 97.

[15] Anushay Furqan, Chelsea Myers, and Jichen Zhu. 2017. Learnability through adaptive discovery tools in voice user interfaces. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 1617–1623.

[16] R. Michael Furr. 2011. *Scale Construction and Psychometrics for Social and Personality Psychology*. https://api.semanticscholar.org/CorpusID:14128547

[17] Bofei Gao, Feifan Song, Yibo Miao, Zefan Cai, Zhe Yang, Liang Chen, Helan Hu, Runxin Xu, Qingxiu Dong, Ce Zheng, Shanghaoran Quan, Wen Xiao, Ge Zhang, Daoguang Zan, Keming Lu, Bowen Yu, Dayiheng Liu, Zeyu Cui, Jian Yang, Lei Sha, Houfeng Wang, Zhifang Sui, Peiyi Wang, Tianyu Liu, and Baobao Chang. 2024. Towards a Unified View of Preference Learning for Large Language Models: A Survey. arXiv:2409.02795 [cs.CL] https://arxiv.org/abs/2409.02795

[18] Declan Grabb. 2023. The impact of prompt engineering in large language model performance: a psychiatric example. In *Journal of Medical Artificial Intelligence*, Vol. 6. 20. https://doi.org/10.21037/jmai-23-71

[19] Michael Guevarra et al. 2025. An LLM-Guided Tutoring System for Social Skills Training. *arXiv preprint arXiv:2501.09870* (2025). https://arxiv.org/abs/2501.09870

[20] Jieun Han, Haneul Yoo, Junho Myung, Minsun Kim, Hyunseung Lim, Yoonsu Kim, Tak Yeon Lee, Hwajung Hong, Juho Kim, So-Yeon Ahn, and Alice Oh. 2024. LLM-as-a-tutor in EFL Writing Education: Focusing on Evaluation of Student-LLM Interaction. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*. Association for Computational Linguistics, 284–293. https://doi.org/10.18653/v1/2024.customnlp4u-1.21

[21] Steve Harrison, Deborah Tatar, and Phoebe Sengers. 2007. The three paradigms of HCI. In *Alt. Chi. Session at the SIGCHI Conference on human factors in computing systems San Jose, California, USA*. 1–18.

[22] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in Psychology* 52 (1988), 139–183.

[23] A. et al. Hasselgren. 2019. Blockchain in healthcare: A fit-for-purpose assessment. *Journal of Medical Internet Research* 21, 12 (2019), e13679. https://doi.org/10.2196/13679

[24] Lujain Ibrahim, Saffron Huang, Lama Ahmad, and Markus Anderljung. 2024. Beyond static AI evaluations: advancing human interaction evaluations for LLM harms and risks. https://doi.org/10.48550/arXiv.2405.10632 arXiv:2405.10632 [cs].

[25] Pelin Karaturhan, Ecem Arikan, Pelin Durak, Asım Evren Yantaç, and Kemal Kuşçu. 2022. Combining Momentary and Retrospective Self-Reflection in a Mobile Photo-Based Journaling Application. In *Proceedings of the Nordic Conference on Human-Computer Interaction*. ACM, 1–12.

[26] Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. 2023. A survey of reinforcement learning from human feedback. *arXiv preprint arXiv:2312.14925* (2023).

[27] C. Maria Keet. 2023. Fit For Purpose. In *The What and How of Modelling Information and Knowledge*. Springer, Cham, 127–156. https://doi.org/10.1007/978-3-031-39695-3_7

[28] Taewan Kim, Seolyeong Bae, Hyun Ah Kim, Su-woo Lee, Hwajung Hong, Chanmo Yang, and Young-Ho Kim. 2024. MindfulDiary: Harnessing Large Language Model to Support Psychiatric Patients' Journaling. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–20.

[29] Taewan Kim, Donghoon Shin, Young-Ho Kim, and Hwajung Hong. 2024. DiaryMate: Understanding User Perceptions and Experience in Human-AI Collaboration for Personal Journaling. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–15.

[30] 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 (CHI '24)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3613904.3642216 event-place: <conf-loc>, <city>Honolulu</city>, <state>HI</state>, <country>USA</country>, </conf-loc>.

[31] Emily Kuang, Minghao Li, Mingming Fan, and Kristen Shinohara. 2024. Enhancing UX Evaluation Through Collaboration with Conversational AI Assistants: Effects of Proactive Dialogue and Timing. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3613904.3642168 event-place: <conf-loc>, <city>Honolulu</city>, <state>HI</state>, <country>USA</country>, </conf-loc>.

[32] Md Tahmid Rahman Laskar, Sawsan Alqahtani, M. Saiful Bari, Mizanur Rahman, Mohammad Abdullah Matin Khan, Haidar Khan, Israt Jahan, Amran Bhuiyan, Chee Wei Tan, Md Rizwan Parvez, Enamul Hoque, Shafiq Joty, and Jimmy Huang. 2024. A Systematic Survey and Critical Review on Evaluating Large Language Models: Challenges, Limitations, and Recommendations. https://doi.org/10.48550/arXiv.2407.04069 arXiv:2407.04069 [cs].

[33] Bettina Laugwitz, Theo Held, and Martin Schrepp. 2008. Construction and evaluation of a user experience questionnaire. In *Symposium of the Austrian HCI and Usability Engineering Group*. Springer, 63–76.

[34] J. W. Lee, R. S. Weiner, and J. M. et al. Sailstad. 2005. Method validation and measurement of biomarkers in nonclinical and clinical samples in drug development. *Pharmaceutical Research* 22 (2005), 499–511. https://doi.org/10.1007/s11095-005-2495-9

[35] 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. 2023. Evaluating Human-Language Model Interaction. http://arxiv.org/abs/2212.09746 arXiv:2212.09746 [cs].

[36] Yi-Chieh Lee, Naomi Yamashita, and Yun Huang. 2021. Exploring the Effects of Incorporating Human Experts to Deliver Journaling Guidance through a Chatbot. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–27.

[37] Alyssa Lees, Danny Driess, Michael Moor, Jonathan Baan, Anna Klimášková, and Mustafa Suleyman. 2023. QUEST: A Framework for Systematic Human Evaluation of Large Language Models in Healthcare. *arXiv preprint arXiv:2310.14373* (2023).

[38] Chen Li et al. 2024. E-EVAL: A Comprehensive Chinese K-12 Education Evaluation Benchmark for Large Language Models. *arXiv preprint arXiv:2401.15927* (2024). https://arxiv.org/abs/2401.15927

[39] Haitao Li et al. 2024. Understanding User Experience in Large Language Model Interactions. *arXiv preprint arXiv:2401.08329* (2024). https://arxiv.org/abs/2401.08329

[40] Junlong Li, Fan Zhou, Shichao Sun, Yikai Zhang, Hai Zhao, and Pengfei Liu. 2024. Dissecting Human and LLM Preferences. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Bangkok, Thailand, 1790–1811.

[41] Yuan Li, Yue Huang, Hongyi Wang, Xiangliang Zhang, James Zou, and Lichao Sun. 2024. Quantifying AI Psychology: A Psychometrics Benchmark for Large Language Models. *arXiv preprint arXiv:2406.17675* (2024).

[42] 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].

[43] James Derek Lomas, Willem van der Maden, Sohhom Bandyopadhyay, Giovanni Lion, Nirmal Patel, Gyanesh Jain, Yanna Litowsky, Haian Xue, and Pieter Desmet. 2024. Improved Emotional Alignment of AI and Humans: Human Ratings of Emotions Expressed by Stable Diffusion v1, DALL-E 2, and DALL-E

3. arXiv:2405.18510 [cs.AI] https://arxiv.org/abs/2405.18510
[44] Angela MacIsaac, Aislin R Mushquash, and Christine Wekerle. 2022. Writing Yourself Well: Dispositional Self-Reflection Moderates the Effect of a Smartphone App-Based Journaling Intervention on Psychological Wellbeing across Time. *Behaviour Research and Therapy* (11 2022). Published online: 29 November 2022.
[45] 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].
[46] Subigya Nepal, Arvind Pillai, William Campbell, Talie Massachi, Michael V Heinz, Ashmita Kunwar, Eunsol Soul Choi, Xuhai Xu, Joanna Kuc, Jeremy F Huckins, et al. 2024. MindScape Study: Integrating LLM and Behavioral Sensing for Personalized AI-Driven Journaling Experiences. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 8, 4 (2024), 1–44.
[47] 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].
[48] Zachary A. Pardos et al. 2024. Leveraging LLM-Respondents for Item Evaluation: a Psychometric Analysis. *arXiv preprint* (2024). https://arxiv.org/html/2407.10899v1
[49] Max Pellert, Clemens M. Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. 2024. AI Psychometrics: Assessing the Psychological Profiles of Large Language Models Through Psychometric Inventories. *Perspectives on Psychological Science* 19 (2024), 808–826. https://doi.org/10.1177/17456916231214460
[50] Carolien Rieffe, Paul Oosterveld, Anne C. Miers, Mark Meerum Terwogt, and Verena Ly. 2008. Emotion awareness and internalising symptoms in children and adolescents: The Emotion Awareness Questionnaire revised. *Personality and Individual Differences* 45, 8 (2008), 756–761. https://doi.org/10.1016/j.paid.2008.08.001
[51] Christoph Rühlemann. 2022. How is emotional resonance achieved in storytellings of sadness/distress? *Frontiers in Psychology* 13 (2022), 952119. https://doi.org/10.3389/fpsyg.2022.952119
[52] Richard M. Ryan and Edward L. Deci. 2000. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist* 55, 1 (2000), 68–78. https://doi.org/10.1037/0003-066X.55.1.68
[53] Saqib Shah. 2023. Snapchat's My AI chatbot is making people paranoid as it 'knows your current location'. *Standard UK* (2023).
[54] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, et al. 2023. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 723–741.
[55] John Smith et al. 2024. Designing LLM-Agents with Personalities: A Psychometric Approach. *arXiv preprint arXiv:2410.19238* (2024). https://arxiv.org/abs/2410.19238
[56] Joshua M Smyth, Jillian A Johnson, Brandon J Auer, Erik Lehman, Giampaolo Talamo, and Christopher N Sciamanna. 2018. Online positive affect journaling in the improvement of mental distress and well-being in general medical patients with elevated anxiety symptoms: A preliminary randomized controlled trial. *JMIR mental health* 5, 4 (2018), e11290.
[57] Suzanne Ho-wai So, James Bennett-Levy, Helen Perry, Debbie Helen Wood, and Chee-wing Wong. 2018. The Self-Reflective Writing Scale (SRWS): A new measure to assess self-reflection following self-experiential cognitive behaviour therapy training. *Reflective Practice* 19, 4 (2018), 505–521.
[58] Åsne Stige, Efpraxia D Zamani, Patrick Mikalef, and Yuzhen Zhu. 2023. Artificial intelligence (AI) for user experience (UX) design: a systematic literature review and future research agenda. *Information Technology & People* (2023).
[59] Rossella Suriano, Alessio Plebe, Alessandro Acciai, and Rosa Angela Fabio. 2025. Student interaction with ChatGPT can promote complex critical thinking skills. *Learning and Instruction* 95 (Feb. 2025), 102011. https://doi.org/10.1016/j.learninstruc.2024.102011
[60] Thomas Yu Chow Tam, Sonish Sivarajkumar, Sumit Kapoor, Alisa V Stolyar, Katelyn Polanska, Karleigh R McCarthy, Hunter Osterhoudt, Xizhi Wu, Shyam Visweswaran, Sunyang Fu, et al. 2024. A framework for human evaluation of large language models in healthcare derived from literature review. *NPJ Digital Medicine* 7, 1 (2024), 258.
[61] Willem Van Der Maden, Derek Lomas, and Paul Hekkert. 2023. A framework for designing AI systems that support community wellbeing. *Frontiers in Psychology* 13 (Jan. 2023), 1011883. https://doi.org/10.3389/fpsyg.2022.1011883
[62] Jiayin Wang, Weizhi Ma, Peijie Sun, Min Zhang, and Jian-Yun Nie. 2024. Understanding User Experience in Large Language Model Interactions. http://arxiv.org/abs/2401.08329 arXiv:2401.08329 [cs].
[63] 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].
[64] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 214–229.
[65] Xuhai Xu et al. 2023. Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data. *arXiv preprint arXiv:2307.14385* (2023). https://arxiv.org/abs/2307.14385
[66] Yuanyuan Xu, Weiting Gao, et al. 2024. Enhancing User Experience and Trust in Advanced LLM-Based Conversational Agents. *Computing and Artificial Intelligence* 4, 1 (2024), 1–20. https://link.springer.com/article/10.1007/s44230-024-00012-3
[67] Xiaoyi Zhang, Laura R Pina, and James Fogarty. 2016. Examining Unlock Journaling with Diaries and Reminders for In Situ Self-Report in Health and Wellness. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 5658–5664.
[68] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Lin, Eric P Li, Joseph Xing, Joseph E Gonzalez, et al. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685* (2023).