Evaluating LLMs in Experiential Context: Insights from a Survey of Recent CHI Publications

Christine Dierk
Adobe Research
San Jose, California, USA
dierk@adobe.com

Jennifer Healey
Adobe Research
San Jose, California, USA
jehealey@adobe.com

Mustafa Doga Dogan
Adobe Research
San Jose, California, USA
doga@adobe.com

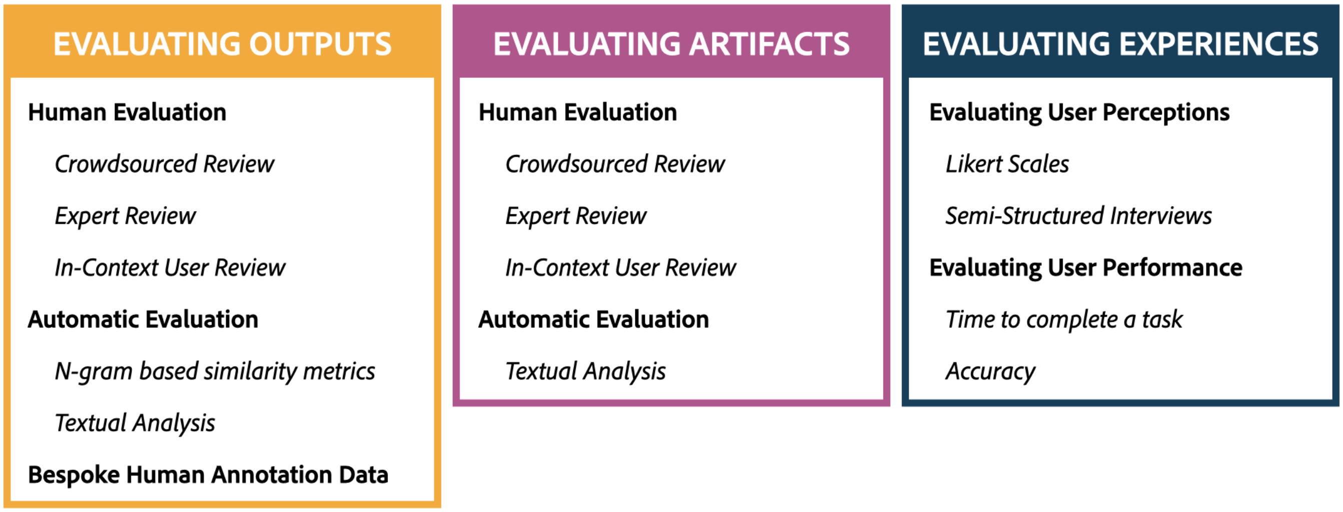


Figure 1: Evaluation strategies observed in a survey of 23 recent CHI publications.

Abstract

The rise of large language models (LLMs) has had far reaching effects across multiple fields, requiring evaluation strategies to assess their impact. In contrast to the framework of quantitative benchmark-based evaluations typically used at AI conferences, evaluating LLMs for human computer interaction requires more nuanced consideration as LLM "performance" in this arena is inherently human-centered and often bespoke to the experiential context. This paper provides a set of insights distilled from a survey of 23 papers recently published at CHI and suggests a lens through which to view HCI LLM evaluation strategies. We discuss the challenges of evaluating LLMs in HCI and provide suggestions to help increase interdisciplinary rigor.

CCS Concepts

• **Human-centered computing → HCI design and evaluation methods**.

Keywords

Evaluation, Human Computer Interaction, Large Language Model, Experience, Artifact, Co-creation, Output

ACM Reference Format:
Christine Dierk, Jennifer Healey, and Mustafa Doga Dogan. 2025. Evaluating LLMs in Experiential Context: Insights from a Survey of Recent CHI Publications. In *Human-centered Evaluation and Auditing of Language Models Workshop at the CHI Conference on Human Factors in Computing Systems (CHI '25), April 26, 2022, Yokohama, Japan.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/35861XX.XXXXX

1 Introduction

Large language models (LLMs) have recently emerged as powerful tools that have transformed multiple fields of research. Human computer interaction is uniquely impacted when LLMs are used as the primary form of "compute" in the interaction. Recently, researchers have been exploring ways to best leverage this powerful new technology to empower novel tools for creativity support [20, 25, 29], UI design [8], literature discovery [17], and beyond [6, 30]. In this paper we conduct a survey to codify how LLMs have been studied at CHI in the past two years and how our methods of evaluation may or may not be suitable for this new type of HCI.

LLMs are becoming increasingly prevalent in HCI research —282 research articles published at CHI in 2024 mention language models[1]. These papers leverage a range of evaluation strategies. Some

CHI '25 HEAL Workshop, Yokohama, Japan

ACM ISBN 979-8-4007-0132-/XX/XX
https://doi.org/10.1145/35861XX.XXXXX

[1] https://dl.acm.org

strategies stem from established techniques in Natural Language Processing (NLP), whereas others take a more human-centered approach. While various methods have been used to evaluate this growing body of work, there has been limited meta-discussion on whether or not these evaluation strategies are appropriate or effective. Since human-centric evaluations necessitate evaluating systems in experiential context, we suspect that there can be no unified framework for how to evaluate LLMs in HCI research. Instead, we argue that there's value in examining existing approaches, understanding how HCI practitioners evaluate LLM-based systems and identifying the strengths and limitations of a human-centered approach. Based on an analysis of 23 representative papers, we contribute an overview of evaluation strategies used in LLM research published at CHI. We identify three types of evaluation strategies: (1) *evaluating outputs*, in other words, evaluating LLM responses directly; (2) *evaluating artifacts*: assessing artifacts co-created by humans and LLMs; and (3) *evaluating experiences*, which includes usability testing, observation studies, and other human-centered evaluation approaches well established in the field of HCI.

While these evaluation strategies are grounded in prior work, we argue that there is a gap in evaluation methodology, constituting an "evaluation crisis" in LLM research. Traditional NLP approaches are effective at quantitatively *evaluating outputs*, but often lack a human-centered perspective. Conversely, evaluation approaches in HCI are effective at *evaluating experiences*, but often lack rigorous technical validation of the LLM outputs. We argue that the most effective evaluations are interdisciplinary, combining strategies informed by both fields. Finally, we provide suggestions to increase the academic rigor of LLM research at CHI, including increased transparency when it comes to model selection and prompt engineering.

2 Methodology

This paper presents a meta-review of evaluation strategies used in 23 papers from CHI 2023 and CHI 2024 that specifically feature LLM or "language model" in the title [2, 3, 5–12, 14, 16–20, 22, 24, 25, 27, 29–31].

2.0.1 Dataset. We gathered 72 research articles published at CHI in 2023 and 2024[2], including "LLM" or "language model" in the title. This inclusion criteria is simultaneously narrow and broad. Narrow—only papers that mention LLMs in the *title* are included, excluding many papers that mention LLMs in the *full text*. Broad—we include *all* papers that mention LLMs in the title, regardless of usage context. We argue that this collection of papers serves as a representative cross-section of LLM research at CHI. In this work, we analyze 23 of the 72 research articles. Most of these papers were randomly selected from the larger sample; a few were chosen by searching for keywords in the full text. For instance, to better understand the use of automatic evaluation strategies at CHI, we selected papers that mention BLEU and ROUGE. This is a work in progress; we are in the process of coding the remaining papers in our larger sample. Results presented in this work are early insights that may be adapted through deeper analysis of the entire corpus.

2.0.2 Analysis. We first conducted open-coding [4] on a randomly selected subset of papers, describing the evaluation methods leveraged in each work. Through this, we identified an initial set of evaluation strategies. We organized these evaluation strategies into three overarching categories: (1) *evaluating outputs*, (2) *evaluating artifacts*, and (3) *evaluating experiences* (See Figure 1). We used the identified evaluation strategies and overarching categories to perform focused coding [4] on additional papers from the sample. As mentioned in the previous section, some of these papers were randomly selected, others were chosen to gain a deeper understanding of particular strategies. Throughout the coding process, we iteratively refined the coding schema.

We code the papers as follows. If the paper evaluates LLM responses directly, this is *evaluating outputs*. These evaluations may be objective (e.g., BLEU, ROUGE) or subjective (e.g., human judgments). If the paper evaluates outcomes co-created by a human and an LLM, this is *evaluating artifacts*. Similar to evaluating outputs, these evaluations may be objective or subjective. However, the distinction lies in human intervention. If a human (e.g., user study participant) modifies the LLM output, engages in additional prompting to achieve a desired outcome, or uses the LLM output in a larger creative process, then evaluations of such outcomes fall under *evaluating artifacts*. If the paper evaluates the experience of interacting with an LLM, this is *evaluating experiences*. These evaluations may be objective (e.g., time to complete a task, number of prompts used) or subjective (e.g., Likert scales to evaluate usability or perceived creativity).

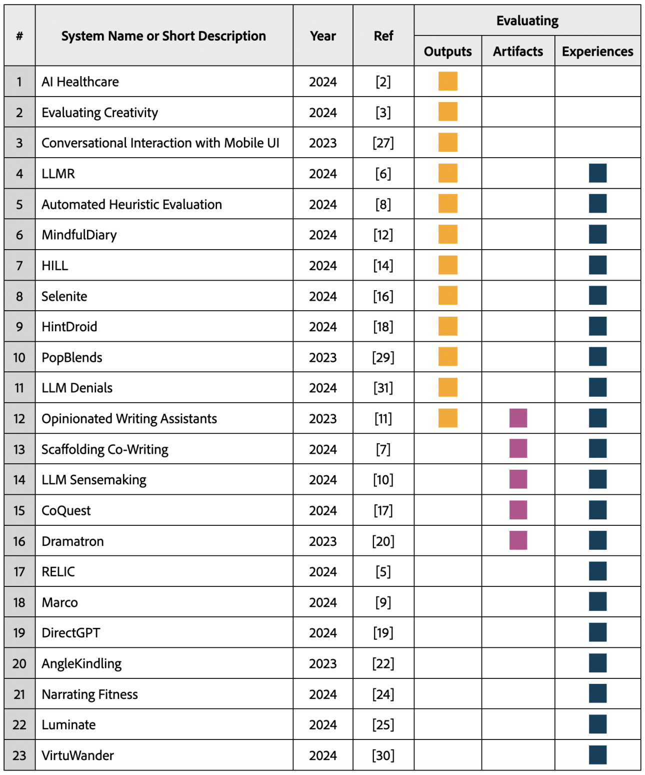
These evaluation types can be used on their own or in combination. As an example, imagine an LLM-powered system designed to support users in crafting short stories. *Evaluating outputs* might entail collecting a dataset of prompts, gathering LLM responses to those prompts, and having crowdworkers evaluate the responses using criteria such as grammar and prompt coherence. Now imagine the authors conducted a user study where participants were asked to use the system to write a short story. *Evaluating artifacts* might entail critiquing the short stories co-written with the system, assuming that study participants modified the LLM outputs and did not blindly accept LLM responses. *Evaluating experiences* might entail measuring system usability with post-test questionnaires (i.e., SUS, NASA-TLX), Likert scales, or semi-structured interviews.

The following sections describe the three evaluation types in detail, highlighting specific examples from the corpus.

3 Evaluating Outputs

One approach for evaluating LLM research is to directly evaluate outputs. Our operational definition of an output is a piece of text generated entirely by an LLM (i.e., the LLM response). This strategy is often used in other fields (e.g., NLP, AI) to evaluate the quality, accuracy, and reliability of LLMs. In HCI research, this category of evaluation is often used for system validation. LLM outputs can be used to enable interactions. Measuring the quality of the outputs is a way to validate whether or not the system supports the proposed interaction. As an example, Liu et al. created an LLM-powered system to support online sensemaking. Authors evaluated LLM outputs (i.e., options and criteria extracted from webpages) to validate that the proposed system could feasibly provide an

[2] We chose these dates as the launch of ChatGPT in November 2022 greatly accelerated the rate of LLM research in HCI and beyond.



#	System Name or Short Description	Year	Ref	Evaluating		
				Outputs	Artifacts	Experiences
1	AI Healthcare	2024	[2]	🟧		
2	Evaluating Creativity	2024	[3]	🟧		
3	Conversational Interaction with Mobile UI	2023	[27]	🟧		
4	LLMR	2024	[6]	🟧		🟦
5	Automated Heuristic Evaluation	2024	[8]	🟧		🟦
6	MindfulDiary	2024	[12]	🟧		🟦
7	HILL	2024	[14]	🟧		🟦
8	Selenite	2024	[16]	🟧		🟦
9	HintDroid	2024	[18]	🟧		🟦
10	PopBlends	2023	[29]	🟧		🟦
11	LLM Denials	2024	[31]	🟧		🟦
12	Opinionated Writing Assistants	2023	[11]	🟧	🟪	🟦
13	Scaffolding Co-Writing	2024	[7]		🟪	🟦
14	LLM Sensemaking	2024	[10]		🟪	🟦
15	CoQuest	2024	[17]		🟪	🟦
16	Dramatron	2023	[20]		🟪	🟦
17	RELIC	2024	[5]			🟦
18	Marco	2024	[9]			🟦
19	DirectGPT	2024	[19]			🟦
20	AngleKindling	2023	[22]			🟦
21	Narrating Fitness	2024	[24]			🟦
22	Luminate	2024	[25]			🟦
23	VirtuWander	2024	[30]			🟦

Figure 2: Overview of papers evaluated in this survey.

accurate high-quality overview to users. Strategies for evaluating outputs include *Human Evaluation* and *Automatic Evaluation*. We also discuss how human-centered approaches to evaluating outputs often necessitate the creation of *Bespoke Human Annotation Data*.

3.1 Human Evaluation

A common strategy for evaluating outputs entails having humans manually review them; for example, showing multiple outputs to humans and asking them to assign scores. The scores can either be absolute or relative. In the relative case, human annotators are asked to compare outputs between systems, or asked to compare outputs to those manually created by humans. Some evaluations in this category include context (i.e., the prompt), whereas others do not. Some evaluations use crowdworkers to review LLM outputs, whereas others use domain experts. Prior work suggests that expert review can increase the validity of LLM evaluations, especially for contexts that require specialized expertise [2].

In NLP, human evaluation tends to be crowdsourced and conducted at large-scale. While relatively inexpensive and easily scalable, this style of evaluation faces issues related to quality and bias [20]. Alternatively, human evaluation at CHI tends to engage a smaller number of domain experts [3, 8]. In a "high-stakes" healthcare context, Calle et al. invited certified tobacco treatment specialists to conduct an expert review of LLM-generated intervention messages for smoking cessation [2]. Using expert-written messages as a benchmark, experts evaluated the LLM-generated messages on quality, accuracy, credibility, and persuasiveness. Another distinction between evaluation strategies is context. In NLP evaluations, crowdworkers may consider the input prompt while evaluating LLM outputs. In HCI evaluations, participants in a user study may be asked to consider their individual usage contexts while evaluating LLM outputs, and may themselves have written the input prompt. To investigate LLM denial styles, Wester et al. conducted two user studies in which an LLM denied user-drafted requests [31]. Participants rated the LLM responses on measures of usefulness, appropriateness, and relevance. Unlike crowdworkers evaluating outputs in isolation, participants evaluated the outputs within the experiential context.

While human evaluation of LLM outputs typically involves assigning scores to well-defined criteria, some works take a more qualitative approach. As an example, Kim et al. conducted thematic analysis and open coding on LLM-generated messages in a healthcare context [12].

3.2 Automatic Evaluation

Evaluation of LLM outputs can additionally include an objective quantitative analysis of the text itself, leveraging automatic evaluation methods. Only three papers in our survey use such automatic evaluation methods, borrowing strategies from NLP and computational linguistics. Two papers used summarization and machine translation metrics (i.e., BLEU [21], ROUGE [15], CIDEr [26], and METEOR [1]) to assess the accuracy of their method versus expert ground truth, one for mobile UI design summarization[27] and one for the generation of text hints for form fields[18]. A third paper used these methods to ensure that LLM-generated motivational messages were sufficiently unlike prior messages, additionally using automated textual analysis (i.e., Linguistic Inquiry and Word Count[3]) to check that messages were of sufficient linguistic quality[2].

While efficient and scalable, automatic evaluation methods in NLP were designed for specific tasks (e.g., abstractive summarization) and may not generalize to broader applications in HCI. Furthermore, many of these metrics have been shown to have poor correlation with human judgments, are often uninterpretable, and have certain biases [23]. Finally, these methods rely on "gold standard" references (i.e., ground truth datasets) that do not exist for many bespoke HCI use-cases. All three papers leveraging automatic evaluation methods additionally conducted a subjective evaluation. Two papers used human evaluation to further assess the LLM outputs [2, 27]. The remaining paper conducted a between-subjects user study to assess usefulness of the proposed system [18].

3.3 Bespoke Human Annotation Data

Many of the approaches described thus far rely on ground truth datasets. However, "gold standard" data does not exist for many real-world applications. Thus, human-centered approaches often necessitate the creation of bespoke human annotation data. This strategy tasks human experts with generating reference data that is then used as a baseline with which to evaluate outputs, leveraging human evaluation or automatic evaluation techniques. This strategy is distinct from traditional NLP ground truth datasets (e.g., reference

[3] https://www.liwc.app

summaries for automatic summarization) in terms of generalizability of the data. Traditional ground truth datasets are generalizable but may not accurately represent real-world tasks. Conversely, bespoke human annotation data is specific to the proposed system and usage context, meaning that the data is not generalizable but more closely resembles intended usage.

We illustrate this strategy with two examples of bespoke human annotation data generated to evaluate LLM outputs. To evaluate using an LLM for automated heuristic evaluation, Duan et al. conducted a traditional heuristic evaluation study with design experts who manually identified guideline violations in a set of UIs [8]. The violations identified by these human experts were then used to contextualize LLM outputs. In another example, Chakrabarty et al. designed a test to objectively evaluate the creativity of a piece of writing [3]. Creative writers used the test to assess 48 short stories that were either written by experts or LLM-generated, creating a ground truth dataset. This dataset was later used to evaluate an LLM's ability to assess creative writing.

4 Evaluating Artifacts

Another approach is to evaluate the artifacts created with the system. Rather than evaluating the LLM outputs directly, this type of evaluation examines the artifacts co-created by the human user and the AI agent. Our operational definition of an artifact is anything created by humans with the assistance of LLMs. These artifacts are often text (e.g., scripts [20], emails [10], and research questions [17]), but could take other forms —for example, images co-created by users and LLM-powered design tools. Artifacts are evaluated the same way *outputs* are evaluated: using human judgments and automatic metrics. However, artifacts differ in that they always exist in experiential context. For this reason, co-created artifacts are often evaluated by the study participants who created them [10, 17, 20]; however, artifacts can also be evaluated by external experts [20], or crowdworkers [11]. Co-created artifacts can also be evaluated quantitatively; for example, leveraging automated tools for text analysis [7] or tracking writer modifications to LLM-generated sentences [11, 20]. In our survey, all works evaluating *artifacts* also evaluated *experiences* (See Figure 2).

We now discuss several exemplar papers to further communicate strategies for *evaluating artifacts*. To investigate the impact of opinionated language models in co-writing tasks, Jakesch et al. conducted a large scale online experiment where participants (N=1506) co-wrote short statements with biased writing assistants [11]. Authors employed crowdworkers to evaluate these co-written short statements, analyzing opinions at the sentence-level. Authors also evaluated the co-written artifacts quantitatively, using comprehensive interaction logs at the key-stroke level to determine how much text was written by the participant versus suggested by the model. As another example, Dhillon et al. evaluated the quality of co-written text using automated tools for text analysis (i.e., TAACO[4] and TAALES[5]), as well as human evaluation [7].

5 Evaluating Experiences

The final strategy is to measure user experience of the LLM-powered tool or system. This is the most common evaluation strategy observed in our survey; all but three papers evaluate experience. Many works in this category leverage traditional HCI approaches such as surveys, usability studies, observation studies, and comparison to baseline interfaces. For LLM-powered technologies designed to facilitate a process, usability studies can help identify issues, measure users' performance (e.g., accuracy, time to complete a task), and provide qualitative feedback. Alternatively, observation studies can reveal how participants use the technology for more open-ended tasks, such as creative writing [20], validating LLM generations [5], or research question ideation[17]. Another approach is to compare LLM-powered technology to a baseline [10, 16]. This baseline is often a similar interface without LLM functionality and AI-powered features. These types of evaluation can be within-subjects [7, 16, 22] or between-subjects [11, 18, 24]. Experience evaluations can assess *user perceptions* of the LLM-powered technology or *user performance* while using the LLM-powered technology. Echoing prior work [13], we found Likert scale questionnaires and open-ended interviews to be effective techniques for eliciting user feedback. Papers in our corpus evaluated the experience of LLM-powered technologies by assessing perceived usefulness of the system [5, 22, 25, 30], self-reported frustration [7, 31], and user trust [17], among other dimensions. Complementary quantitative metrics include time to complete a task [10, 16, 19], accuracy [18], number of prompts [19], and number of generated artifacts [17, 29]. Several works in the corpus conducted case studies to further evaluate the user experience of the LLM-powered technology [10, 16].

6 Discussion

In a survey of 23 papers recently published at CHI, we identified three categories of evaluation strategies. In this section, we discuss how evaluation strategies can be combined for comprehensive system evaluation. We also discuss several challenges of evaluating LLMs at CHI, before detailing limitations & future work.

6.1 Combined Approaches

More than half of the papers in our sample (13/23) used a combination of approaches. We describe three of these papers to detail how combined approaches can enable a more holistic system evaluation and uncover hidden insights.

6.1.1 AI-Powered Heuristic Evaluation. Duan et al. conducted three studies to evaluate using an LLM for automated heuristic evaluation [8]. The first study leveraged expert evaluation of LLM outputs to validate that the proposed system could feasibly provide accurate and helpful heuristic evaluation feedback (*evaluating outputs*). The second study further validated system performance by quantitatively comparing LLM outputs to bespoke human annotation data generated by experts manually conducting heuristic evaluations (*evaluating outputs*). Finally, the authors conducted a usability study in which they tasked design experts with using the LLM tool to iteratively refine a set of user interfaces, probing user perceptions of the system (*evaluating experiences*).

[4] https://www.linguisticanalysistools.org/taaco.html
[5] https://www.linguisticanalysistools.org/taales.html

6.1.2 Co-Writing Screenplays with Dramatron. As another example, Mirowski et al. recruited 15 experts to co-write scripts and screenplays with an LLM-based support tool [20]. Participants provided feedback on the interactive co-authorship process via Likert scales and open-ended interviews (*evaluating experiences*). Participants also analyzed and provided an artistic opinion of the scripts co-written with the system (*evaluating artifacts*). As an additional evaluation of creative outcomes, authors staged performances of several of the scripts, which were critiqued by independent reviewers (*evaluating artifacts*).

6.1.3 Research Question Ideation with CoQuest. Finally, Liu et al. proposed a novel LLM-based agent supporting research question ideation. In a user study, participants' individually rated each generated research question on dimensions of novelty, value, surprise, and relevance (*evaluating artifacts*). Participants also evaluated the LLM-powered system on dimensions of control, creativity, meta-creativity, cognitive load, and trust (*evaluating experiences*). By evaluating both artifacts and experiences, authors uncovered an interesting finding — one interaction design made users "feel" more creative; however, research questions generated from a different interaction design were rated as more creative.

These exemplar works demonstrate how HCI practitioners can combine evaluation strategies for LLM research.

6.2 Challenges of Evaluating LLMs at CHI

While HCI provides much needed perspective to LLM discourse, human-centered evaluations of LLMs have several limitations. HCI practitioners tend to evaluate a single model. Where authors discuss model selection, insights are sometimes anecdotal and not always derived from technical evaluation. In addition, prompt engineering efforts are not always communicated. Not all papers in our sample directly share the prompts used. This limits replicability, and it's unclear how robust the presented systems are to changes in prompts. Evaluating LLM-powered applications in experiential contexts can more closely approximate real-world usage; however, these studies are not able to be conducted at the same scale as crowdsourced evaluations. While there are numerous HCI methods for *evaluating experiences* both quantitatively and qualitatively, HCI methods for *evaluating outputs* tend to be qualitative only (i.e., subjective ratings of output quality). Only four papers in our sample quantitatively evaluated LLM responses. Three papers used automatic evaluation methods [2, 18, 27][6] while a fourth quantified compilation errors in LLM-generated code [6]. These papers were intentionally selected from the larger corpus of 72 papers using keywords related to automatic evaluation (i.e., BLEU, ROUGE); thus, under representation of this evaluation type is intrinsic and not due to sampling bias.

HCI research could increase academic rigor and reproducibility by increasing transparency around model selection and prompt engineering. Furthermore, automatic evaluation strategies in NLP have evolved beyond N-gram based similarity metrics (e.g., BLEU, ROUGE). Future evaluations in HCI could leverage more advanced evaluation metrics from the field of NLP [23].

6.3 Limitations and Future Work

This is a work in progress. We are in the process of coding the remaining papers in our dataset. Results presented in this work are early insights. As we examine more works, we expect to iteratively refine our coding schema and surface additional insights.

This survey includes research articles published at CHI in 2023 and 2024 with LLM or "language model" in the title. This excludes works published at other HCI conferences (e.g., UIST, DIS, C&C, etc), as well as numerous works that leverage LLMs but do not mention them in the title. This work is part of a larger effort characterizing evaluation strategies for LLM research in HCI more broadly. In the future, we are interested in surveying evaluation strategies employed at other HCI conferences (e.g., UIST, DIS, C&C) to assess if there are any significant differences between HCI sub-communities. We are also interested in examining CHI 2025 proceedings to assess how evaluation strategies evolve over time.

7 Conclusion

We examined 23 papers published at CHI in 2023 and 2024 as a first attempt at categorizing evaluation methods for LLM research in HCI. Through this survey, we surfaced three overarching categories of evaluation strategies: (1) *evaluating outputs*, (2) *evaluating artifacts*, and (3) *evaluating experiences.* We presented these categories using examples from our corpus to demonstrate specific strategies. We discussed combined approaches and several challenges of evaluating LLMs from a human-centered perspective. We hope this lens enables deeper reflection on evaluation strategies for LLM research at CHI and beyond.

[6] One of these papers was a continuation of work previously published at UIST [28], which may explain its unique evaluation strategy in comparison with other papers published at CHI.

References

[1] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss (Eds.). Association for Computational Linguistics, Ann Arbor, Michigan, 65–72. https://aclanthology.org/W05-0909/

[2] Paul Calle, Ruosi Shao, Yunlong Liu, Emily T Hébert, Darla Kendzor, Jordan Neil, Michael Businelle, and Chongle Pan. 2024. Towards AI-Driven Healthcare: Systematic Optimization, Linguistic Analysis, and Clinicians' Evaluation of Large Language Models for Smoking Cessation Interventions. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 436, 16 pages. doi:10.1145/3613904.3641965

[3] Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. Art or Artifice? Large Language Models and the False Promise of Creativity. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 30, 34 pages. doi:10.1145/3613904.3642731

[4] Kathy Charmaz. 2006. *Constructing grounded theory: A practical guide through qualitative analysis.* sage.

[5] Furui Cheng, Vilém Zouhar, Simran Arora, Mrinmaya Sachan, Hendrik Strobelt, and Mennatallah El-Assady. 2024. RELIC: Investigating Large Language Model Responses using Self-Consistency. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 647, 18 pages. doi:10.1145/3613904.3641904

[6] Fernanda De La Torre, Cathy Mengying Fang, Han Huang, Andrzej Banburski-Fahey, Judith Amores Fernandez, and Jaron Lanier. 2024. LLMR: Real-time Prompting of Interactive Worlds using Large Language Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 600, 22 pages. doi:10.1145/3613904.3642579

[7] Paramveer S. Dhillon, Somayeh Molaei, Jiaqi Li, Maximilian Golub, Shaochun Zheng, and Lionel Peter Robert. 2024. Shaping Human-AI Collaboration: Varied Scaffolding Levels in Co-writing with Language Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA)

(CHI '24). Association for Computing Machinery, New York, NY, USA, Article 1044, 18 pages. doi:10.1145/3613904.3642134

[8] Peitong Duan, Jeremy Warner, Yang Li, and Bjoern Hartmann. 2024. Generating Automatic Feedback on UI Mockups with Large Language Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 6, 20 pages. doi:10.1145/3613904.3642782

[9] Raymond Fok, Nedim Lipka, Tong Sun, and Alexa F Siu. 2024. Marco: Supporting Business Document Workflows via Collection-Centric Information Foraging with Large Language Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 842, 20 pages. doi:10.1145/3613904.3641969

[10] Katy Ilonka Gero, Chelse Swoopes, Ziwei Gu, Jonathan K. Kummerfeld, and Elena L. Glassman. 2024. Supporting Sensemaking of Large Language Model Outputs at Scale. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 838, 21 pages. doi:10.1145/3613904.3642139

[11] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-Writing with Opinionated Language Models Affects Users' Views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 111, 15 pages. doi:10.1145/3544548.3581196

[12] 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 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 701, 20 pages. doi:10.1145/3613904.3642937

[13] David Ledo, Steven Houben, Jo Vermeulen, Nicolai Marquardt, Lora Oehlberg, and Saul Greenberg. 2018. Evaluation Strategies for HCI Toolkit Research *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–17. doi:10.1145/3173574.3173610

[14] Florian Leiser, Sven Eckhardt, Valentin Leuthe, Merlin Knaeble, Alexander Mädche, Gerhard Schwabe, and Ali Sunyaev. 2024. HILL: A Hallucination Identifier for Large Language Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 482, 13 pages. doi:10.1145/3613904.3642428

[15] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013/

[16] Michael Xieyang Liu, Tongshuang Wu, Tianying Chen, Franklin Mingzhe Li, Aniket Kittur, and Brad A Myers. 2024. Selenite: Scaffolding Online Sensemaking with Comprehensive Overviews Elicited from Large Language Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 837, 26 pages. doi:10.1145/3613904.3642149

[17] Yiren Liu, Si Chen, Haocong Cheng, Mengxia Yu, Xiao Ran, Andrew Mo, Yiliu Tang, and Yun Huang. 2024. How AI Processing Delays Foster Creativity: Exploring Research Question Co-Creation with an LLM-based Agent *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 17, 25 pages. doi:10.1145/3613904.3642698

[18] Zhe Liu, Chunyang Chen, Junjie Wang, Mengzhuo Chen, Boyu Wu, Yuekai Huang, Jun Hu, and Qing Wang. 2024. Unblind Text Inputs: Predicting Hint-text of Text Input in Mobile Apps via LLM. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 51, 20 pages. doi:10.1145/3613904.3642939

[19] Damien Masson, Sylvain Malacria, Géry Casiez, and Daniel Vogel. 2024. DirectGPT: A Direct Manipulation Interface to Interact with Large Language Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 975, 16 pages. doi:10.1145/3613904.3642462

[20] Piotr Mirowski, Kory W. Mathewson, Jaylen Pittman, and Richard Evans. 2023. Co-Writing Screenplays and Theatre Scripts with Language Models: Evaluation by Industry Professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 355, 34 pages. doi:10.1145/3544548.3581225

[21] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. In *BLEU: a method for automatic evaluation of machine translation* (Philadelphia, Pennsylvania) *(ACL '02)*. Association for Computational Linguistics, USA, 311–318. doi:10.3115/1073083.1073135

[22] Savvas Petridis, Nicholas Diakopoulos, Kevin Crowston, Mark Hansen, Keren Henderson, Stan Jastrzebski, Jeffrey V Nickerson, and Lydia B Chilton. 2023. AngleKindling: Supporting Journalistic Angle Ideation with Large Language Models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 225, 16 pages. doi:10.1145/3544548.3580907

[23] Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2022. A Survey of Evaluation Metrics Used for NLG Systems. *ACM Comput. Surv.* 55, 2, Article 26 (Jan. 2022), 39 pages. doi:10.1145/3485766

[24] Konstantin R. Strömel, Stanislas Henry, Tim Johansson, Jasmin Niess, and Paweł W. Woźniak. 2024. Narrating Fitness: Leveraging Large Language Models for Reflective Fitness Tracker Data Interpretation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 646, 16 pages. doi:10.1145/3613904.3642032

[25] Sangho Suh, Meng Chen, Bryan Min, Toby Jia-Jun Li, and Haijun Xia. 2024. Luminate: Structured Generation and Exploration of Design Space with Large Language Models for Human-AI Co-Creation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 644, 26 pages. doi:10.1145/3613904.3642400

[26] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4566–4575. doi:10.1109/CVPR.2015.7299087

[27] Bryan Wang, Gang Li, and Yang Li. 2023. Enabling Conversational Interaction with Mobile UI using Large Language Models *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 432, 17 pages. doi:10.1145/3544548.3580895

[28] Bryan Wang, Gang Li, Xin Zhou, Zhourong Chen, Tovi Grossman, and Yang Li. 2021. Screen2Words: Automatic Mobile UI Summarization with Multimodal Learning. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '21)*. Association for Computing Machinery, New York, NY, USA, 498–510. doi:10.1145/3472749.3474765

[29] Sitong Wang, Savvas Petridis, Taeahn Kwon, Xiaojuan Ma, and Lydia B Chilton. 2023. PopBlends: Strategies for Conceptual Blending with Large Language Models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 435, 19 pages. doi:10.1145/3544548.3580948

[30] Zhan Wang, Lin-Ping Yuan, Liangwei Wang, Bingchuan Jiang, and Wei Zeng. 2024. VirtuWander: Enhancing Multi-modal Interaction for Virtual Tour Guidance through Large Language Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 612, 20 pages. doi:10.1145/3613904.3642235

[31] Joel Wester, Tim Schrills, Henning Pohl, and Niels van Berkel. 2024. "As an AI language model, I cannot": Investigating LLM Denials of User Requests. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 979, 14 pages. doi:10.1145/3613904.3642135