SPHERE: An Evaluation Card for Human-AI Systems

Dora Zhao[1*] , Qianou Ma[2*], Xinran Zhao[2], Chenglei Si[1], Chenyang Yang[2],
Ryan Louie[1], Ehud Reiter[3], Diyi Yang[1†] , Tongshuang Wu[2†]

[1]Stanford University, [2]Carnegie Mellon University, [3]University of Aberdeen
qianoum@andrew.cmu.edu, dorothyz@stanford.edu

Abstract

In the era of Large Language Models (LLMs), establishing effective evaluation methods and standards for diverse human-AI interaction systems is increasingly challenging. To encourage more transparent documentation and facilitate discussion on human-AI system evaluation design options, we present an evaluation card SPHERE, which encompasses five key dimensions: 1) What is being evaluated?; 2) How is the evaluation conducted?; 3) Who is participating in the evaluation?; 4) When is evaluation conducted?; 5) How is evaluation validated? We conduct a review of 39 human-AI systems using SPHERE, outlining current evaluation practices and areas for improvement. We provide three recommendations for improving the validity and rigor of evaluation practices.

1 Introduction

The proliferation of LLMs has changed the way humans interact with AI systems. Compared to previously existing AI models, LLMs can better comprehend and generate human-like text, enabling users to engage with AI through natural language in a more conversational manner (Brown et al., 2020). By leveraging these capabilities, system designers can create human-AI systems[1] that span a range of domains and roles. Despite the rapid advances in designing new human-AI systems, it is still unclear how to best evaluate them. Standard evaluation practices in NLP are better suited for understanding model performance using automated metrics and static benchmarks. While there have been efforts to integrate more human-centered evaluation, there are also concerns about the validity and reproducibility of how these evaluations are conducted (Belz et al., 2023; Howcroft et al., 2020; Gehrmann et al., 2023).

---

* Co-first authors.
† Co-last authors.

Given the wide diversity of human-AI systems, what factors should researchers consider when designing evaluations? How do we ensure that these evaluations are transparent and replicable? To address these questions, we need systematic methods for documenting how these evaluations are conducted. As a step in this direction, we propose the SPHERE evaluation card, which provides a comprehensive template for designing and documenting evaluation protocols used to assess human-AI systems.[1] Although we focus on systems powered by LLMs in this work, the evaluation dimensions we discuss are agnostic to the type of model and can be applied to AI systems more broadly.

SPHERE Evaluation Card for Human-AI Systems
(Subject) **What** is being evaluated?
• **Component**: Model, System
• **Design Goal**: Effectiveness, Efficiency, Satisfaction
(Process) **How** is the evaluation being conducted?
• **Scope**: Intrinsic, Extrinsic
• **Method**: Quantitative, Qualitative
(Handler) **Who** is participating in the evaluation?
• **Automated**: Static, Generative
• **Human**: Expert, User
(Elasped) **When** is evaluation conducted (duration)?
• **Time Scale**: Immediate, Short-term, Long-term
(Robustness) **How** is evaluation validated?
• **Validation**: Reliability, Validity

The purpose of SPHERE is two-fold. First, researchers can use the template when *designing* evaluations (Section 3). We enumerate five high-

---

[1]Following the definitions from the literature (Lee et al., 2023; Amershi et al., 2019), a human-AI system harnesses AI capabilities that are exposed to end users through an interface. These systems consist of an AI model, a user interface, and logic that converts user entry into input for the model. We instantiate the AI as LLM when we refer to human-AI systems and evaluation in this paper.

Table 1: SPHERE covers five dimensions of human-AI system evaluation, with 8 categories and 18 aspects. We provide examples and visualize the distribution of SPHERE aspects from papers published at HCI (left bar) and NLP venues reviewed in our literature survey (details in Section 4).

Category	Aspect	HCI Examples		NLP Examples
What is being evaluated?				
Component	Model	Accuracy of LLM gen. (Lee et al., 2024b)		BERTScore (Glória-Silva et al., 2024)
	System	Knowledge quiz (Shaikh et al., 2024)		Headline quality (Ding et al., 2023)
Design	Effectiveness	System risk & content (Rajashekar et al., 2024)		Label quality & stability (Wei et al., 2024)
Goal	Efficiency	Perceived workload (Lee et al., 2024b)		Time on task (Ding et al., 2023)
	Satisfaction	Likert-scale rating of fun (Wang et al., 2024c)		Likert-scale rating of trust (Ding et al., 2023)
How is an evaluation conducted?				
Scope	Intrinsic	System Usability Scale (Liu et al., 2024a)		Retrieval hit rate (Inan et al., 2024)
	Extrinsic	Engagement & enjoyment (Fan et al., 2024)		Identified concept diversity (Yang et al., 2023)
Method	Quantitative	Interaction logs (Wu et al., 2022)		Micro F1 (Wei et al., 2024)
	Qualitative	Interview & grounded coding (Liu et al., 2023)		Case study (Cai et al.)
Who is participating in the evaluation?				
Human	Expert	Prolific experts (Zavolokina et al., 2024)		ASL expert (Inan et al., 2024)
	User	Students & physicians (Rajashekar et al., 2024)		Crowdworkers (Chakrabarty et al., 2022)
Automated	Static	Perplexity & LIWC scores (Calle et al., 2024)		Precision & recall (Yang et al., 2023)
	Generative	N/A		Consistency by LLaMa2 (Zhao et al., 2024)
When is evaluation conducted (duration)?				
Time	Immediate	# of clicks (Lawley and Maclellan, 2024)		Benchmark (Raheja et al., 2023)
Scale	Short-term	1-hour usability study (Liu et al., 2024a)		10 minutes per poem (Chakrabarty et al., 2022)
	Long-term	3-days session (Fan et al., 2024)		6-months deployment (Inan et al., 2024)
(Meta) How is evaluation validated?				
Validation	Reliability	Krippendroff's α as IRR (Lee et al., 2024b)		Fleiss' κ for annotation (Zhao et al., 2024)
	Validity	Counterbalancing (Wu et al., 2022)		Randomized control (Ding et al., 2023)

level questions that scaffold the dimensions that researchers should think through to ensure that their evaluation aligns with the intended design goal. Within each dimension, we discuss popular practices from both NLP (e.g., benchmarking, LLM-as-a-judge) and from HCI (e.g., experimental user studies, semi-structured interviews) that can be adopted. Second, SPHERE can be used to *document* how evaluation is conducted. To address concerns around the replicability and reproducibility of results, researchers can use the cards to communicate their evaluation design. This practice helps improve transparency in the field and fosters a shared language for researchers to communicate even if the systems being evaluated are different (see examples in Appendix D).[2]

Using SPHERE, we analyze 39 human-LLM systems from NLP and HCI venues (Section 4). From our analysis, we provide three key recommendations for improving evaluation practices: 1) establish evaluations in real-world contexts; 2) strengthen validity and interpretability of results via triangulating various evaluation methods; and 3) rigorously evaluate evaluation practices. These rec-

ommendations bridge the strengths of both communities, for example, HCI's focus on stakeholder relevance and NLP's advancements in automatic quantitative measures, ultimately shaping more robust and actionable evaluation. Finally, we present two case studies (Section 5) showcasing SPHERE for designing and reproducing evaluations.

2 Background

NLP is facing what some scholars have termed an "evaluation crisis" — how to best evaluate the capabilities of generative models remains an open question (Blodgett et al., 2024; Xiao et al., 2024). Established NLP methods, such as static benchmarks, are found to be ill-suited to judge model performance on generative tasks (McIntosh et al., 2024), spurring efforts to make more dynamic and comprehensive benchmarks (Liang et al., 2023). Others have advocated for more *human-centered* evaluations of model capabilities (Liao and Xiao, 2023; Blodgett et al., 2024; Elangovan et al., 2024). These concerns also coincide with alarms around the lack of experimental reproducibility and repeatability in human evaluations of NLP systems (Belz et al., 2023). Taken together,

---

[2]Template for SPHERE is released on sphere-eval.github.io.

there is growing uncertainty about how evaluations should be conducted going forward and how to ensure the quality of evaluation results.

Going beyond model capabilities, evaluating human-AI systems introduces additional challenges. Researchers must consider not only the model performance but also the system's impact on users (Weidinger et al., 2023). Although there are many guidelines about how to design human-AI systems (Amershi et al., 2019; Wright et al., 2020), there is less literature on how we should be evaluating them. Examples of work that tackle system evaluation include Lee et al. (2023)'s framework for human-LLM interaction evaluation that captures the process and user preferences beyond static model outputs quality. Others have also proposed methods focused on assessing the safety of these systems (Ibrahim et al., 2024; Weidinger et al., 2023), or for domain-specific applications (Lee et al., 2024a). However, there is still a gap in articulating a comprehensive overview of how to design evaluations for human-AI systems.

To address this, we present the SPHERE evaluation card: a framework covering five dimensions of human-AI system evaluation that helps researchers *design* and *document* evaluation. As a design tool, SPHERE structures conversations around key evaluation areas. As documentation (Gebru et al., 2021; Mitchell et al., 2019), SPHERE contributes to the transparency and reproducibility of these methods.

3 The SPHERE Evaluation Card

In this section, we present the five dimensions and corresponding aspects of evaluation included in our SPHERE evaluation card (Table 1), and we highlight the challenges and considerations for each dimension. Similar to existing documentation efforts (Mitchell et al., 2019), we note that these dimensions are not intended to be exhaustive. Researchers may want to report additional dimensions of evaluation depending on their system design. See Fig. 1 for an example SPHERE card.

Method   We develop our framework using an expert-based affinity diagramming approach (Hartson, 2012; Lucero, 2015; Harboe and Huang, 2015) across 9 authors.[3] Each author enumerated important aspects and theories of human-AI system evaluation in their domains. Via synchronous conversations, we clustered these aspects into high-level

---

[3] Authors have a median of 4 years experience in human-AI interaction and system evaluation in NLP and HCI.

SPHERE Evaluation Card: AngleKindling
What is being evaluated?
System component's *effectiveness*, *efficiency*, and *satisfaction*: how many pursuable angles were generated, mental demand, and perceived helpfulness.
How is evaluation conducted?
Extrinsic within-subjects study compared with IN-JECT that includes post-task survey (*quantitative* Likert ratings) and semi-structured interview (*qualitative*) on feature preferences and workflow impact.
Who is participating in the evaluation?
12 professional journalists (*domain experts, intended users*). All were English speakers based in the US.
When is evaluation conducted?
Short-term sessions up to 60 minutes with $30.
How is evaluation validated?
Counterbalancing of tool order to reduce learning effects for *validity*.

Figure 1: Example SPHERE card for the system AngleKindling (Petridis et al., 2023).

themes, prioritizing the most salient ones. We then organized the themes using the Who, What, When, and How questions, taking inspiration from how these general dimensions have guided exploration and analysis across domains (Apte et al., 2001). Finally, we refined the dimensions after applying them to two example systems (see Appendix D).

3.1 What is being evaluated?

The first question to answer when designing an evaluation is to determine *what* is being evaluated. To answer this, we discuss two categories of aspects: which part of the system is the focus of the evaluation and what goal the evaluation is testing.

Components   Since human-AI systems consist of multiple components, we must identify what part of the system is being evaluated. A helpful delineation is between evaluating **model** behavior versus the **system** as-a-whole.

We break out the **model** as a separate component in evaluation since it represents a unique design challenge of human-AI systems, driven by uncertainty surrounding model capabilities and the complexity of outputs (Yang et al., 2020). Uncertainty can be compounded by the fact that systems may include multiple models with different functions (e.g., Wang et al. (2024c)).

Yet, models are but one part of a more complex artifact; designers introduce interfaces and inter-

actions that integrate with the model components to form a human-AI **system**. Understanding how these design choices impact users' experiences requires evaluating the system holistically.

Design Goals Evaluations should be formulated to help prove a design goal of the model or system. To taxonomize possible design goals, we use three categories from the ISO standard definition of usability (Bevan et al., 2016):

- **Effectiveness**: accuracy, completeness, and lack of negative consequences with which users achieved specified goals. This category maps closely to the quality criteria used in NLG evaluation (Howcroft et al., 2020; Reiter, 2024).
- **Efficiency**: resources (e.g., time, cognitive effort) required to achieve the model's or system's goals.
- **Satisfaction**: positive attitudes, emotions, and/or comfort resulting from the use.

Considerations: selecting design goals and scenarios We provide two considerations for deciding what to evaluate. First, human-AI systems do not have to aim for all three design goals. In fact, researchers should consider how designing a system for one goal could harm another. For example, systems relying on entirely automated decision-making may be more efficient but can be considered less trustworthy (Hong et al., 2020).

Second, researchers should factor in how goals differ across scenarios. System performance will vary depending on whether we are evaluating the "average" versus worst case. Particularly in high-stakes domains, we must consider this long-tail of system behavior as they may pose immense harm (Bickmore et al., 2018). One popular technique is red-teaming systems, although finding adversarial scenarios may be expensive and hard to identify pre-deployment (Ganguli et al., 2022; Mei et al., 2023). Once deployed, the system should be monitored for failures, requiring corrective action or even removal of the system in extreme cases (Syed, 2015; Wolf et al., 2017).

3.2 How is the evaluation conducted?

Next, we must consider how the evaluation is conducted, including the scope of the evaluation and methods used. Note that human-AI systems may require multiple evaluations employing different scopes and methods. For example, researchers may validate the model or interface design before evaluating the system with users. Each evaluation would be conducted differently.

Scope Evaluating human-AI systems requires assessing both their internal capabilities (**intrinsic evaluation**) and performance in real-world scenarios (**extrinsic evaluation**) (Jones and Galliers, 1995). Prior work has pointed out that NLP systems tend to disproportionately favor intrinsic evaluation, and have pushed for more extrinsic evaluation (Gkatzia and Mahamood, 2015; Gehrmann et al., 2023). While intrinsic evaluations are still important, researchers should be mindful of how well these internal metrics correlate to real-world utility (Reiter and Belz, 2009; Belz and Gatt, 2008).

Method Both **quantitative** and **qualitative** methods can be used to conduct intrinsic and extrinsic evaluations but yield different types of insights and can be complementary to each other. In mixed-method analyses, qualitative methods add more nuance to quantitative results. For example, Petridis et al. (2023)'s quantitative results establish that their proposed system outperforms an existing one, while qualitative responses in a semi-structured interview unearth reasons why their system failed and highlight possible improvements.

Considerations: selecting and implementing evaluation methods Different methods have their own considerations to ensure rigorous execution. Quantitative analysis provides generalizable insights and facilitates comparison across groups. Therefore, when using quantitative methods, researchers should consider selecting metrics and datasets that are representative of the system's task (e.g., when benchmarking) or sampling participants to ensure experiments have sufficient statistical power (Charness et al., 2012). On the other hand, qualitative analysis aims to provide a deeper understanding or "thick description" (Geertz, 2008) of behavior. Generalizability may *not* be a priority (Donmoyer et al., 2000; Soden et al., 2024). Researchers may want to consider extant theory, sampling strategies, and coding techniques (e.g., open coding, selective coding) (Cairns and Cox, 2008; Cole and Gillies, 2022).

3.3 Who is participating in the evaluation?

When designing evaluation methods, researchers must consider who (or what) is participating in this evaluation, including human-centered evaluations and automated methods spanning from standard benchmarking techniques to LLM judges.

3.3.1 Human Evaluators

When selecting human evaluators, we must account for how participants' identities or backgrounds may influence how they interact with human-AI systems. HCI work has provided frameworks for thinking about how demographic background (Bardzell and Bardzell, 2011; Ogbonnaya-Ogburu et al., 2020; Schlesinger et al., 2017) influences how individuals may interact with technology. In addition to these factors, in our SPHERE evaluation cards, we highlight two aspects: whether participants are the intended design targets and their level of expertise.

Intended Users When selecting human evaluators, we must first decide, whether the system is evaluated by the user (i.e., the intended design target) or another stakeholder. While responses from intended users will more closely mirror how people will interact with the deployed system, there are other stakeholders who may be affected by the system and whose input should be considered. For example, Ma et al. (2024) had teachers evaluate an AI-infused tutoring system, for which students are the intended users; other potential evaluators could have been parents or administrators. Working across stakeholder groups can lead to tensions when communities may have differing priorities or standards for evaluation (Reiter and Belz, 2009).

Experts Another point to consider is the expertise of the evaluators. Expertise can refer to the domain expertise, such as having trained physicians evaluate a clinical decision support system (Rajashekar et al., 2024) and having teachers evaluating a tutoring system (Ma et al., 2024). Here, expertise is not to be confused with experts from crowdworking platforms, which can also refer to workers who have passed qualification studies (Chakrabarty et al., 2022). Domain expert evaluations can produce more reliable judgments than non-experts (Yesilada et al., 2009) and bring field-specific insights, but their perceptions may not align with those of the intended system users. Working with expert evaluators can also be more expensive and time-intensive.

3.3.2 Automated Evaluators

There are many ways to automatically evaluate systems. We point out two main categories: evaluations where outputs are compared to an established reference and evaluations using generative models to make judgments (Zheng et al., 2024).

Static Evaluators We refer to methods that compare model or system behaviors to some existing ground-truth behavior as static evaluation. For example, benchmarking with perplexity metrics or rule-based evaluations (Blagec et al., 2022; Van Miltenburg et al., 2020) fall in this category. Prior work has raised concerns about the usefulness and validity of existing benchmarks for assessing generative tasks (McIntosh et al., 2024). Benchmarking may not be necessary when systems involve less model implementation, such as when researchers prompt an off-the-shelf model.

Generative Evaluators Researchers can also use LLMs to judge outputs (typically originating from another LLM).[4] For example, Zhao et al. (2024) use LLaMA-2-70B to rate the system's responses along the dimensions of consistency, relevance, empathy, and commonsense. Generative evaluators allow researchers to run more ablations under controlled conditions and iterate on system design quickly (Zheng et al., 2024). Generative evaluators operate under the premise that the models' choices are similar to that of humans; however, there remain concerns about biases inherent in dataset construction methods (Wang et al., 2024b) or preferences for outputs generated from certain models (Li et al., 2023; Yin et al., 2023).

Considerations: specifying relevant evaluators When deciding who to include as evaluators, researchers should aim for specificity. Some human-AI systems are presented as "general-purpose" or without a defined intended user group. Works like Chatbot Arena (Chiang et al., 2024) open up system evaluation to an indiscriminate potential end-user. Nonetheless, there is still value in recruiting specific user groups, such as participants who may be historically excluded from the development of such technologies (Ogbonnaya-Ogburu et al., 2020) or users who might interact with the system in more high-risk domains (Rauh et al., 2024).

Similar concerns apply when using automated methods. Researchers should reflect on whose perspectives are excluded from the evaluation. Generative models may not adequately simulate diverse personas or capture the nuances of different identity groups (Wang et al., 2024a; Cheng et al., 2023).

---

[4]We would not consider an embedding-based evaluation method such as BERTScore (Zhang et al., 2020) as a generative evaluation. The method compares BERT embeddings between the output and a reference rather than using the model for generation.

Benchmarks are also not immune; they are shaped by the design biases and positionality of the dataset creator (Santy et al., 2023).

3.4 When is evaluation conducted (duration)?

While many standard benchmark evaluations for AI systems can be run in seconds, evaluating human-AI systems requires us to consider time factors. Drawing from Newell and Card (1985)'s time scales of human action, we discuss evaluations that can occur at three different time scales:

- **Immediate**: When evaluation occurs at the time-scale of milliseconds and seconds, rational thought processes are not yet at play (Newell and Card, 1985). For instance, telemetry or log data can be used to analyze real-time interactions (Liu et al., 2024d; Kim et al., 2024b). Similarly, automated benchmarking approaches measure performance at a fixed point in time.
- **Short-term**: Evaluating over the course of minutes or hours sheds light on human behaviors and thoughts in a bounded context. Short-term evaluation is crucial for measuring the benefits of interacting with a human-AI system. However, they may be biased by known psychological phenomena, such as the novelty effect (Elston, 2021), and fail to capture longer-term impacts of usage.
- **Long-term**: Studies operating on longer time scales (days, months, years, and more) capture behavioral changes and effects from social interaction that may not appear in isolated laboratory experiments. Over time, users also may form different mental models of the systems, changing their interaction patterns. For example, Bansal et al. (2019) found that users' trust in AI systems evolved with prolonged exposure and interaction.

Considerations: balancing desired outcomes with practicalities of evaluation duration When deciding the duration of evaluation, researchers must weigh trade-offs between desired outcomes and practical factors. Immediate evaluations, such as automated methods, are cheap and efficient, but they may fail to capture the consequences of interacting with the system. Alternatively, longitudinal studies are crucial to understanding sustained impacts and broader implications of AI adoption in real-world workflows (e.g., privacy expectations (Khowaja et al., 2024), workforce impact (Butler et al., 2023)). With new technologies, the novelty effect can bias short-term evaluation (Long et al., 2024). Long-term evalua-

tions, however, are time-consuming and financially expensive (Caruana et al., 2015). Researchers also need to manage high attrition rates and possibly intervene if drop-out follows systematic patterns (Hogan et al., 2004).

3.5 How is evaluation validated?

Finally, researchers must ensure that their evaluations are sound and replicable. Drawing on concepts from the social sciences (Bandalos, 2018; Drost, 2011), we present two qualities — reliability and validity — that should be assessed when evaluating the evaluation design. We refer to this step as "meta-evaluation."

Reliability Researchers must consider reliability, or whether the evaluation produces consistent results. Three important dimensions of reliability are as follows: stability over time ("Do results remain the same across time points?"); equivalence ("Are results consistent across different versions of the evaluation?"); and internal consistency ("Do the components of my evaluation method measure the same concept?") (Drost, 2011). For example, Taeb et al. (2024) included three Likert-scale questions to evaluate system usefulness; to check internal consistency, we want to validate that the questions all measure the same concept.

Validity Measures can be reliable but still not be valid. Researchers must also consider whether their evaluation techniques meaningfully capture the intended construct (Adcock and Collier, 2001). Assessing validity is particularly important when evaluating properties that are socially constructed (e.g., system's impact on creativity (He et al., 2023; Fan et al., 2024)) compared to more objective measures (e.g., task completion time (Liu et al., 2023)).

Considerations: carefully executing valid, reliable, and replicable evaluations For reliability, a popular measure of internal consistency is Cronbach's α, which intuitively captures correlations between test items (Tavakol and Dennick, 2011). When humans or models are used to rate system outputs, researchers must account for the reliability of their judgments. For example, LLM judges can produce different responses even when given the same prompt (Stureborg et al., 2024; Shi et al., 2024). We can measure annotation consistency or inter-rater reliability (IRR) across different raters (Artstein and Poesio, 2008; Gwet, 2001). Common metrics for IRR include Fleiss's κ and

Cohen's κ for categorical labels, correlation metrics for continuous scores, and rank-based correlations, including Spearman's ρ or Kendall's τ for rank labels. Methods such as test-retest and parallel measures of the same concept are also useful for assessing other aspects of reliability (Drost, 2011).

Depending on the evaluation, there are different considerations for validity. For behavioral experiments, researchers must mitigate systematic biases (e.g., confirmation bias (Hart et al., 2009), anchoring bias (Block and Harper, 1991)) that can impact experimental results. For automated methods, particularly generative evaluators, model biases or errors may differ from those of humans (Bavaresco et al., 2024). Thus, researchers need to check the consistency between the automated results and human perception (Gehrmann et al., 2023; Shankar et al., 2024). Overall, evaluation experiments must be carefully executed in order to be meaningful and provide real-world utility (Reiter, 2024).

4 Applying the SPHERE Evaluation Card

SPHERE provides a structured framework for analyzing and improving how human-AI systems are evaluated. In this section, we apply SPHERE to works published in NLP and HCI venues. Through this analysis, we uncover trends, gaps, and best practices that might otherwise be overlooked, such as underrepresented evaluation aspects or core evaluation designs. We distill actionable lessons to guide researchers in designing more robust evaluation paradigms that align with the real-world contexts in which these systems operate.

4.1 Method

To identify human-LLM systems, we searched for papers published in human-computer interaction (CHI) or natural language processing (*CL) venues between Jan. 2022 and Sept. 2024. We kept papers that mentioned ["human*" OR "user*"] AND "system*" AND ["large language model*" OR "LLM*"] in the abstract or title and then manually inspected to ensure that a human-LLM system was introduced.[1] We reviewed 39 papers — with 21 papers from HCI venues and 18 from NLP. Six of the authors analyzed the papers. We first independently coded two of the 39 papers (Krippendorf's $\alpha = 0.69$) before discussing disagreements synchronously. The remaining papers were then divided among

the authors. We visualize the distribution of codes across papers from HCI and NLP venues in Table 1. See Appendix A for details.

4.2 Recommendations

Building on insights learned in our analysis with SPHERE, we propose three recommendations to improve the quality of human-AI system evaluations.

4.2.1 Evaluate to reflect real-world use

A core challenge of evaluating human-AI systems is bridging the gap between model benchmark performance and real-world usage. Extrinsic evaluation is critical for understanding how systems will perform in the real world. Our analysis found that only 10 of the 18 NLP papers included extrinsic evaluation compared to 20 of the 21 HCI papers.

Test systems in the real world The most common method across the 30 papers is running within- or between-subjects experiments to quantitatively compare systems for a pre-determined task. One issue with this paradigm is that it is limited to controlled laboratory settings, which may lack ecological validity. Results may not always translate to real-world scenarios. Only two of the reviewed papers deployed the system to the real world and studied user behavior in situ. Fan et al. (2024) had participants use the ContextCam system in their day-to-day lives over three days, and Inan et al. (2024) had public users use their multimodal dialogue system for over six months. Going beyond general public deployment, researchers can consider evaluating their systems in users' real-world workflows, providing deeper insight into the system's utility, particularly for professional contexts (e.g., Knoll et al. (2022)). Real-world deployment gives insight into how actual users perceive the system in a realistic setting and across contexts — the ultimate test for human-AI systems.

Recruit evaluators from relevant stakeholder groups Selecting evaluators is crucial to ensure population validity. Researchers should recruit relevant stakeholders when running human evaluations. Only three of the 13 NLP papers reviewed with user evaluations recruited domain experts. Others did not provide details on user background or relied on crowdworkers and students who may not represent the target user population. Using crowdworkers or convenience sampling can be more time and cost-efficient. However, downsides include higher variance in responses and a potential lack

of relevant expertise for domain-specific evaluation (Karpinska et al., 2021).

Finding the right evaluators is challenging. One way to recruit more representative users is through collaborating with relevant organizations. For example, when evaluating their mental health counselor training system, Hsu et al. (2025) worked with 7 Cups of Tea, an existing platform for online mental health support, to recruit participants. Examining evaluation practices has implications upstream regarding how systems should be designed. Working more closely with users during the design process, such as conducting formative studies or adopting co-design practices, better motivates the system and forms relationships for finding evaluators (Burkett, 2012). There has also been growing interest in using LLMs to simulate human raters or even using human-LLM collaboration for annotation (Li et al., 2023; Gao et al., 2025). As addressed in Section 3.3 and Section 3.5, if adopting these methods, we recommend researchers first evaluate whether the generated responses align with users.

4.2.2 Cross-verify results across different evaluation methods

With the vast space of applications and interactions that human-AI systems afford, we argue that it is important to adopt different methods to increase the evaluation robustness and cross-verify the findings. However, less than half of the papers in our corpus utilized all intrinsic, extrinsic, quantitative, and qualitative methods to cross-validate their evaluation results. When using the SPHERE card to design evaluations, we advocate researchers pay attention to whether a multifaceted evaluation is adopted to help mitigate biases inherent in any single evaluation method (Section 3.2).

Triangulate results To improve confidence in evaluation results, methodological triangulation — or using more than one method for evaluation — has become a popular approach for tackling complex and nuanced questions across disciplines (Heale and Forbes, 2013; Tashakkori and Creswell, 2007). For example, as discussed in Section 3.2, using mixed methods allows researchers to blend qualitative and quantitative insights.

Triangulation is also important for establishing credibility, as no method is without its limitations. For example, user preferences and benchmark results may not correlate with user task performance, and different methods will help reveal a richer set of insights in combination (Mozannar et al., 2025).

In the current landscape of evaluation, extrinsic evaluations may overly rely on self-reported measures. Specifically, eight of the 25 papers that included a quantitative extrinsic evaluation only reported users' Likert scale ratings. While self-perceived ratings provide useful information about the system, they are also subject to cognitive biases and can be unreliable (Elangovan et al., 2024; Leung, 2011; Bishop and Herron, 2015). Self-reported results can be supplemented with other measurements. For example, Kim et al. (2024b) and Liu et al. (2024d) analyze logs (e.g., click behavior, length of input) to understand how users interact with the system at a more granular level.

Ground methods in existing practices As a product of triangulation, researchers may draw on methods outside their field. In this case, it is important to base techniques in existing practices. For example, seven papers from NLP venues include qualitative results (four offer case studies and three present select results as examples). One concern is that they do not justify how case studies were created or how qualitative examples were selected, raising concerns about methodological rigor.

To demonstrate how we can ground our methods in other fields, we refer to HCI papers for examples on how to conduct qualitative analyses. For inductive analyses, papers often adopt a grounded-theory approach and discuss how they developed codes (Thornberg et al., 2014). For example, Arawjo et al. (2024) analyzed interview data "through a combination of inductive thematic analysis through affinity diagramming." For deductive analysis, Lee et al. (2024b) introduced a rubric with five dimensions. They provided details on dimensions that were created, defined, and then applied across annotators. See Appendix C.2 for examples of how qualitative methods are described.

4.2.3 Rigorously evaluate evaluation

It is essential to critically assess the evaluation methodologies themselves to understand the quality of findings and support replication in the future. Consistent with prior work (Thomson et al., 2024; Card et al., 2020), we find there is a lack of rigorous meta-evaluations or validation of evaluation methodologies across our corpus. For example, the number of users participating in the evaluation tends to be small — only five of the 39 papers we surveyed included more than 30 participants

in their sample size. This can lead to limited generalizability, introduce additional biases, and risk underpowering statistical tests (Christley, 2010).

Expand meta-evaluation methods   Current practices for measuring reliability and validity are also constrained in scope. For reliability, papers focused almost exclusively on inter-rater reliability. Only one study reported on internal reliability (measured using Cronbach's α). Other methods, such as test-retest or split-half, are not employed. For validity measures, papers mentioned using randomized controlled experiment designs, employing counterbalancing, and drawing from pre-validated surveys (e.g., System Usability Scale (Bangor et al., 2008), NASA Task Load Index (Hart, 2006)). We suggest researchers consider other practices to ensure validity when applicable to customized evaluations, such as factor analysis, which is commonly used to check the validity of evaluation items in educational tests and surveys (Knekta et al., 2019).

Document evaluation practices for replication   Meta-evaluation plays a critical part in stewarding scientific best practices. In theory, evaluation methods should be clearly documented and reproducible by others in the community; however, in practice, problems with replication have plagued both the NLP and HCI communities (Belz et al., 2023; Echtler and Häußler, 2018). Our SPHERE card aims to facilitate the design and documentation of evaluation to support replication.

5   Case Study

Finally, we present two case studies demonstrating how SPHERE can be used both for designing and reproducing evaluations.

5.1   Using SPHERE to design evaluations

We recruited two first authors (A1, A2) from the sample of 39 papers surveyed in Section 4. The authors were first asked to reflect on how they would improve the evaluation design from their paper. Then, they were asked to create a SPHERE card for their paper before repeating the reflection process.

The authors reported that using SPHERE encouraged them to engage in deeper considerations about the extrinsic implications of their systems and the validity of their evaluations. For example, after creating the SPHERE card, both authors discussed integrating a long-term deployment study as part of their evaluation plan to *"understand how the system can help users in their daily workflow"* (A1)

and *"to increase ecological validity"* (A2). A1 also discussed integrating statistical testing for repeated model evaluation, arising from the section on meta-evaluation in SPHERE's framework.

5.2   Using SPHERE to reproduce evaluations

SPHERE serves not only as a design tool but also as documentation. In our second case study, we recruited two PhD students — one who focuses on NLP (P1) and the other on HCI (P2). The participants were asked to write a reproduction plan for a human-AI system from the other domain using only the paper as reference. After writing the initial plan, they were given the SPHERE card for the paper and asked to revise their plans. Then, we presented the reproduction plans written before seeing the SPHERE card and after to the original authors of the paper to evaluate.

Participants reported improved understanding and confidence in reproducing the system's evaluation when given the SPHERE card. As P1 stated, before having access to SPHERE they found it difficult to understand the overarching design of the evaluation. In contrast, SPHERE *"allow[ed] for quick understanding"* (P2) of the evaluation plan. Moreover, plans created when participants had access to SPHERE were more detailed and accurate. Before having access to SPHERE, P2's reproduction plan only included the user study from the paper, overlooking the automated evaluation and case study. However, these details were included in the reproduction plan after having access to SPHERE.

6   Conclusion

We introduce the SPHERE evaluation card that can be used a priori to help researchers design more robust evaluations and post-hoc to standardize documentation on evaluation protocols. SPHERE includes five key dimensions: what is being evaluated; how is evaluation conducted; who is participating in the evaluation; when is evaluation conducted; and how is the evaluation validated. Using SPHERE, we survey 39 papers presenting new human-LLM systems published in HCI and NLP venues and present recommendations charting how evaluation practices should improve going forward. Through the adoption of SPHERE evaluation card, we hope to facilitate new evaluation practices that are more realistic, rigorous, and reproducible.

Limitations

This work has several limitations stemming from the scope and methodology of our review and framework application. First of all, we applied the SPHERE card on a set of 39 *CL or CHI papers with human-AI systems. There is a vast space of venues that might publish on human-AI systems, such as domain-specific applications (e.g., AIED, medical journals) and trustworthy AI (e.g., FAccT). By selecting papers from only *CL and CHI, we may not fully capture the breadth of evaluation practices in diverse fields and miss some domain-specific insights. Additionally, our focus on human-AI systems that require an explicit interface may have excluded more NLP-centered studies that do not meet this criterion.

Furthermore, we limited our inquiry on human-AI systems to human-LLM systems. While this focus is in response to the wide adoption of LLMs in current human-AI system research, it excludes other AI modalities, such as vision-centric systems, which may present unique evaluation challenges and opportunities. SPHERE has the potential to be applied as a model-agnostic evaluation framework; nonetheless, it should be viewed as a starting point for broader inquiries into human-AI system evaluation. Future work could aim to include a more diverse and representative set of studies and explore evaluation practices across a wider range of AI capabilities and application areas.

Finally, we evaluate SPHERE using a case study that qualitatively examines how practitioners may apply the card across two settings. In future work, we encourage more comprehensive evaluation methods, such as quantitative analyses on how effective SPHERE is, or qualitative studies with think-aloud protocols, similar to Boyd (2021)'s evaluation of Datasheets for Datasets (Gebru et al., 2021), to better understand practitioners' thought processes when using SPHERE.

Potential Risks

A potential risk of this work is having documentation serve solely as an additional burden upon researchers, instead of fostering any positive change for human-AI system evaluation. For example, Heger et al. (2022) interviewed machine learning practitioners on using datasheets (Gebru et al., 2021) for documenting datasets. They found that many of the practitioners they interviewed prioritized efficiency, and viewed documentation as taking time away from more important tasks. Some participants would complete the datasheets with the minimal amount of information required. In this case, SPHERE runs the risk of being completed performatively, which deviates from our intentions of having the framework foster discussion about how to design evaluations and improve transparency for others in the community.

Acknowledgment

References

Robert Adcock and David Collier. 2001. Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review*, 95(3):529–546.

Angus Addlesee, Neeraj Cherakara, Nivan Nelson, Daniel Hernandez Garcia, Nancie Gunson, Weronika Sieińska, Christian Dondrup, and Oliver Lemon. 2024. Multi-party Multimodal Conversations Between Patients, Their Companions, and a Social Robot in a Hospital Memory Clinic. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL): System Demonstrations*, pages 62–70.

Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for human-AI interaction. In *ACM CHI Conference on Human Factors in Computing Systems*.

Prakash R Apte, Harish Shah, and Darrell Mann. 2001. "5W's and an H of Innovation: TRIZ. *TRIZ Journal*.

Ian Arawjo, Chelse Swoopes, Priyan Vaithilingam, Martin Wattenberg, and Elena L. Glassman. 2024. ChainForge: A Visual Toolkit for Prompt Engineering and LLM Hypothesis Testing. In *ACM CHI Conference on Human Factors in Computing Systems*.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596.

DL Bandalos. 2018. *Measurement theory and applications for the social sciences*. Guilford Publications.

Aaron Bangor, Philip T Kortum, and James T Miller. 2008. An empirical evaluation of the system usability

scale. *Intl. Journal of Human–Computer Interaction*, 24(6):574–594.

Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-AI team performance. In *AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*.

Shaowen Bardzell and Jeffrey Bardzell. 2011. Towards a feminist HCI methodology: social science, feminism, and HCI. In *ACM CHI Conference on Human Factors in Computing Systems*.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, and 1 others. 2024. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. *CoRR*.

Anja Belz and Albert Gatt. 2008. Intrinsic vs. extrinsic evaluation measures for referring expression generation. In *Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

Anya Belz, Craig Thomson, Ehud Reiter, Gavin Abercrombie, Jose M. Alonso-Moral, Mohammad Arvan, Anouck Braggaar, Mark Cieliebak, Elizabeth Clark, Kees van Deemter, Tanvi Dinkar, Ondřej Dušek, Steffen Eger, Qixiang Fang, Mingqi Gao, Albert Gatt, Dimitra Gkatzia, Javier González-Corbelle, Dirk Hovy, and 23 others. 2023. Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous human evaluations in NLP. In *Workshop on Insights from Negative Results in NLP*. Association for Computational Linguistics.

Nigel Bevan, Jim Carter, Jonathan Earthy, Thomas Geis, and Susan Harker. 2016. New ISO standards for usability, usability reports and usability measures. In *International Conference on Human-Computer Interaction*.

Timothy W Bickmore, Ha Trinh, Stefan Olafsson, Teresa K O'Leary, Reza Asadi, Nathaniel M Rickles, and Ricardo Cruz. 2018. Patient and consumer safety risks when using conversational assistants for medical information: An observational study of siri, alexa, and google assistant. *J Med Internet Res*, 20(9):e11510.

Phillip A Bishop and Robert L Herron. 2015. Use and misuse of the likert item responses and other ordinal measures. *International Journal of Exercise Science*, 8(3):297.

Kathrin Blagec, Georg Dorffner, Milad Moradi, Simon Ott, and Matthias Samwald. 2022. A global analysis of metrics used for measuring performance in natural language processing. In *NLP Power! The First Workshop on Efficient Benchmarking in NLP*.

Richard A Block and David R Harper. 1991. Overconfidence in estimation: Testing the anchoring-and-adjustment hypothesis. *Organizational Behavior and Human Decision Processes*, 49(2):188–207.

Su Lin Blodgett, Jackie Chi Kit Cheung, Vera Liao, and Ziang Xiao. 2024. Human-centered evaluation of language technologies. In *Conference on Empirical Methods in Natural Language Processing (EMNLP): Tutorial Abstracts*.

Karen L Boyd. 2021. Datasheets for datasets help ml engineers notice and understand ethical issues in training data. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–27.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*.

Ingrid Burkett. 2012. An introduction to co-design. *Sydney: Knode*, 12.

Jenna Butler, Sonia Jaffe, Nancy Baym, Mary Czerwinski, Shamsi Iqbal, Kate Nowak, Sean Rintel, Mihaela Vorvoreanu Abigail Sellen (VP Distinguished Scientist), Brent Hecht, and Jaime Teevan. 2023. Microsoft new future of work report 2023. Microsoft Research Tech Report MSR-TR-2023-34.

Yuzhe Cai, Shaoguang Mao, Wenshan Wu, Zehua Wang, Yaobo Liang, Tao Ge, Chenfei Wu, WangYou WangYou, Ting Song, Yan Xia, Nan Duan, and Furu Wei. Low-code LLM: Graphical User Interface over Large Language Models. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations)*.

Paul Cairns and Anna L Cox. 2008. *Research methods for human-computer interaction*, volume 10. Cambridge University Press Cambridge.

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 *ACM CHI Conference on Human Factors in Computing Systems*.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Edward Joseph Caruana, Marius Roman, Jules Hernández-Sánchez, and Piergiorgio Solli. 2015. Longitudinal studies. *Journal of thoracic disease*, 7(11):E537.

Tuhin Chakrabarty, Vishakh Padmakumar, and He He. 2022. Help me write a Poem - Instruction Tuning as a Vehicle for Collaborative Poetry Writing. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Gary Charness, Uri Gneezy, and Michael A Kuhn. 2012. Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization*, 81(1):1–8.

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 *ACM CHI Conference on Human Factors in Computing Systems*. ACM.

Myra Cheng, Tiziano Piccardi, and Diyi Yang. 2023. CoMPosT: Characterizing and evaluating caricature in llm simulations. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: an open platform for evaluating llms by human preference. In *International Conference on Machine Learning (ICML)*.

Robert M Christley. 2010. Power and error: increased risk of false positive results in underpowered studies. *The Open Epidemiology Journal*, 3(1).

Tom Cole and Marco Gillies. 2022. More than a bit of coding:(un-) grounded (non-) theory in HCI. In *ACM CHI Conference on Human Factors in Computing Systems: Extended Abstracts (CHI EA)*.

Zijian Ding, Alison Smith-Renner, Wenjuan Zhang, Joel Tetreault, and Alejandro Jaimes. 2023. Harnessing the power of LLMs: Evaluating human-AI text co-creation through the lens of news headline generation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP): Findings*.

Robert Donmoyer and 1 others. 2000. Generalizability and the single-case study. *Case study method: Key issues, key texts*, pages 45–68.

Ellen A Drost. 2011. Validity and reliability in social science research. *Education Research and perspectives*, 38(1):105–123.

Florian Echtler and Maximilian Häußler. 2018. Open source, open science, and the replication crisis in HCI. In *ACM CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI EA)*.

Aparna Elangovan, Ling Liu, Lei Xu, Sravan Babu Bodapati, and Dan Roth. 2024. ConSiDERS-the-human evaluation framework: Rethinking human evaluation for generative large language models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Dirk M Elston. 2021. The novelty effect. *Journal of the American Academy of Dermatology*, 85(3):565–566.

Xianzhe Fan, Zihan Wu, Chun Yu, Fenggui Rao, Weinan Shi, and Teng Tu. 2024. ContextCam: Bridging Context Awareness with Creative Human-AI Image Co-Creation. In *ACM CHI Conference on Human Factors in Computing Systems*.

Hao Fei, Han Zhang, Bin Wang, Lizi Liao, Qian Liu, and Erik Cambria. 2024. EmpathyEar: An Open-source Avatar Multimodal Empathetic Chatbot. In *Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, and 17 others. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *Preprint*, arXiv:2209.07858.

Mingqi Gao, Xinyu Hu, Xunjian Yin, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2025. Llm-based nlg evaluation: Current status and challenges. *Computational Linguistics*.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Clifford Geertz. 2008. Thick description: Toward an interpretive theory of culture. In *The cultural geography reader*, pages 41–51. Routledge.

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

Dimitra Gkatzia and Saad Mahamood. 2015. A snapshot of NLG evaluation practices 2005-2014. In *European Workshop on Natural Language Generation (ENLG)*.

Diogo Glória-Silva, Rafael Ferreira, Diogo Tavares, David Semedo, and Joao Magalhaes. 2024. Plan-Grounded Large Language Models for Dual Goal Conversational Settings. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1271–1292, St. Julian's, Malta. Association for Computational Linguistics.

Kilem Gwet. 2001. Handbook of inter-rater reliability. *Gaithersburg, MD: STATAXIS Publishing Company*, pages 223–246.

Gunnar Harboe and Elaine M Huang. 2015. Real-world affinity diagramming practices: Bridging the paper-digital gap. In *ACM CHI Conference on Human Factors in Computing Systems*.

Sandra G Hart. 2006. NASA-task load index (nasa-tlx); 20 years later. In *Human Factors and Ergonomics Society Annual Meeting*, volume 50.

William Hart, Dolores Albarracin, Alice Eagly, Inge Brechan, Matthew Lindberg, and Lisa Merrill. 2009. Feeling validated versus being correct: A meta-analysis of selective exposure to information. *Psychological bulletin*, 135:555–88.

Rex Hartson. 2012. *The UX Book: Process and Guidelines for Ensuring a Quality User Experience*. Elsevier.

Jun-Yan He, Zhi-Qi Cheng, Chenyang Li, Jingdong Sun, Wangmeng Xiang, Xianhui Lin, Xiaoyang Kang, Zengke Jin, Yusen Hu, Bin Luo, and 1 others. 2023. Wordart designer: User-driven artistic typography synthesis using large language models. In *Conference on Empirical Methods in Natural Language Processing: Industry Track (EMNLP)*.

Roberta Heale and Dorothy Forbes. 2013. Understanding triangulation in research. *Evidence-based nursing*, 16(4):98–98.

Amy K Heger, Liz B Marquis, Mihaela Vorvoreanu, Hanna Wallach, and Jennifer Wortman Vaughan. 2022. Understanding machine learning practitioners' data documentation perceptions, needs, challenges, and desiderata. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–29.

Joseph W Hogan, Jason Roy, and Christina Korkontzelou. 2004. Handling drop-out in longitudinal studies. *Statistics in medicine*, 23(9):1455–1497.

Sungsoo Ray Hong, Jessica Hullman, and Enrico Bertini. 2020. Human factors in model interpretability: Industry practices, challenges, and needs. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–26.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *International Conference on Natural Language Generation*.

Shang-Ling Hsu, Raj Sanjay Shah, Prathik Senthil, Zahra Ashktorab, Casey Dugan, Werner Geyer, and Diyi Yang. 2025. Helping the helper: Supporting peer counselors via ai-empowered practice and feedback. *Proc. ACM Hum.-Comput. Interact. 9, 2, Article CSCW095 (April 2025),*.

Songbo Hu, Xiaobin Wang, Moy Yuan, Anna Korhonen, and Ivan Vulić. 2024. DIALIGHT: Lightweight Multilingual Development and Evaluation of Task-Oriented Dialogue Systems with Large Language Models. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations (NAACL)*.

Lujain Ibrahim, Saffron Huang, Lama Ahmad, and Markus Anderljung. 2024. Beyond static ai evaluations: advancing human interaction evaluations for llm harms and risks. *arXiv preprint arXiv:2405.10632*.

Mert Inan, Katherine Atwell, Anthony Sicilia, Lorna Quandt, and Malihe Alikhani. 2024. Generating Signed Language Instructions in Large-Scale Dialogue Systems. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track (NAACL)*.

Karen Sparck Jones and Julia R Galliers. 1995. *Evaluating natural language processing systems: An analysis and review*. Springer Science & Business Media.

Irina Jurenka, Markus Kunesch, Kevin R McKee, Daniel Gillick, Shaojian Zhu, Sara Wiltberger, Shubham Milind Phal, Katherine Hermann, Daniel Kasenberg, Avishkar Bhoopchand, and 1 others. 2024. Towards responsible development of generative AI for education: An evaluation-driven approach. *arXiv preprint arXiv:2407.12687*.

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. The perils of using mechanical turk to evaluate open-ended text generation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Sunder Ali Khowaja, Parus Khuwaja, Kapal Dev, Weizheng Wang, and Lewis Nkenyereye. 2024. ChatGPT needs spade (sustainability, privacy, digital divide, and ethics) evaluation: A review. *Cognitive Computation*.

Hannah Kim, Kushan Mitra, Rafael Li Chen, Sajjadur Rahman, and Dan Zhang. 2024a. MEGAnno+: A Human-LLM Collaborative Annotation System. In *Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (EACL)*.

Tae Soo Kim, Yoonjoo Lee, Jamin Shin, Young-Ho Kim, and Juho Kim. 2024b. EvalLM: Interactive Evaluation of Large Language Model Prompts on User-Defined Criteria. In *ACM CHI Conference on Human Factors in Computing Systems*. ACM.

Eva Knekta, Christopher Runyon, and Sarah Eddy. 2019. One size doesn't fit all: Using factor analysis to gather validity evidence when using surveys in your research. *CBE—Life Sciences Education*, 18(1):rm1.

Tom Knoll, Francesco Moramarco, Alex Papadopoulos Korfiatis, Rachel Young, Claudia Ruffini, Mark Perera, Christian Perstl, Ehud Reiter, Anja Belz, and Aleksandar Savkov. 2022. User-driven research of medical note generation software. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*.

Lane Lawley and Christopher Maclellan. 2024. VAL: Interactive Task Learning with GPT Dialog Parsing. In *ACM CHI Conference on Human Factors in Computing Systems*. ACM.

Mina Lee, Katy Ilonka Gero, John Joon Young Chung, Simon Buckingham Shum, Vipul Raheja, Hua Shen, Subhashini Venugopalan, Thiemo Wambsganss, David Zhou, Emad A Alghamdi, and 1 others. 2024a. A design space for intelligent and interactive writing assistants. In *ACM CHI Conference on Human Factors in Computing Systems*.

Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, and 1 others. 2023. Evaluating human-language model interaction. *Transactions on Machine Learning Research*.

Yoonjoo Lee, Hyeonsu B Kang, Matt Latzke, Juho Kim, Jonathan Bragg, Joseph Chee Chang, and Pao Siangliulue. 2024b. PaperWeaver: Enriching Topical Paper Alerts by Contextualizing Recommended Papers with User-collected Papers. In *ACM CHI Conference on Human Factors in Computing Systems*.

Shing-On Leung. 2011. A comparison of psychometric properties and normality in 4-, 5-, 6-, and 11-point likert scales. *Journal of social service research*, 37(4):412–421.

Qintong Li, Leyang Cui, Lingpeng Kong, and Wei Bi. 2023. Collaborative evaluation: Exploring the synergy of large language models and humans for open-ended generation evaluation. *arXiv preprint arXiv:2310.19740*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, and 31 others. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research (TMLR)*.

Q Vera Liao and Ziang Xiao. 2023. Rethinking model evaluation as narrowing the socio-technical gap. *arXiv preprint arXiv:2306.03100*.

Michael Xieyang Liu, Advait Sarkar, Carina Negreanu, Benjamin Zorn, Jack Williams, Neil Toronto, and Andrew D. Gordon. 2023. "What It Wants Me To Say": Bridging the Abstraction Gap Between End-User Programmers and Code-Generating Large Language Models. In *ACM CHI Conference on Human Factors in Computing Systems*.

Michael Xieyang Liu, Tongshuang Wu, Tianying Chen, Franklin Mingzhe Li, Aniket Kittur, and Brad A Myers. 2024a. Selenite: Scaffolding Online Sensemaking with Comprehensive Overviews Elicited from Large Language Models. In *ACM CHI Conference on Human Factors in Computing Systems*. ACM.

Renjie Liu, Yanxiang Zhang, Yun Zhu, Haicheng Sun, Yuanbo Zhang, Michael Huang, Shanqing Cai, Lei Meng, and Shumin Zhai. 2024b. Proofread: Fixes all errors with one tap. In *Annual Meeting of the Association for Computational Linguistics: System Demonstrations (ACL)*.

Xingyu Bruce Liu, Jiahao Nick Li, David Kim, Xiang 'Anthony' Chen, and Ruofei Du. 2024c. Human I/O: Towards a Unified Approach to Detecting Situational Impairments. In *ACM CHI Conference on Human Factors in Computing Systems*.

Yiren Liu, Si Chen, Haocong Cheng, Mengxia Yu, Xiao Ran, Andrew Mo, Yiliu Tang, and Yun Huang. 2024d. How AI Processing Delays Foster Creativity: Exploring Research Question Co-Creation with an LLM-based Agent. In *ACM CHI Conference on Human Factors in Computing Systems*.

Tao Long, Katy Ilonka Gero, and Lydia B Chilton. 2024. Not just novelty: A longitudinal study on utility and customization of an AI workflow. In *ACM Designing Interactive Systems Conference (DIS)*.

Andrés Lucero. 2015. Using affinity diagrams to evaluate interactive prototypes. In *Human-Computer Interaction (INTERACT)*, pages 231–248. Springer.

Xiang Luo, Zhiwen Tang, Jin Wang, and Xuejie Zhang. 2024. DuetSim: Building User Simulator with Dual Large Language Models for Task-Oriented Dialogues. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5414–5424, Torino, Italia. ELRA and ICCL.

Pingchuan Ma, Rui Ding, Shuai Wang, Shi Han, and Dongmei Zhang. 2023. InsightPilot: An LLM-Empowered Automated Data Exploration System. In *Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP)*.

Qianou Ma, Hua Shen, Kenneth Koedinger, and Tongshuang Wu. 2024. How to teach programming in the AI era? using llms as a teachable agent for debugging. *International Conference on Artificial Intelligence in Education (AIED)*.

Timothy R McIntosh, Teo Susnjak, Nalin Arachchilage, Tong Liu, Paul Watters, and Malka N Halgamuge. 2024. Inadequacies of large language model benchmarks in the era of generative artificial intelligence. *arXiv preprint arXiv:2402.09880*.

Alex Mei, Sharon Levy, and William Wang. 2023. AS-SERT: Automated safety scenario red teaming for evaluating the robustness of large language models. In *Findings of the Conference on Empirical Methods in Natural Language Processin (EMNLP)*.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.

Hussein Mozannar, Valerie Chen, Mohammed Alsobay, Subhro Das, Sebastian Zhao, Dennis Wei, Manish Nagireddy, Prasanna Sattigeri, Ameet Talwalkar, and David Sontag. 2025. The RealHumanEval: Evaluating large language models' abilities to support programmers. *Transactions on Machine Learning Research (TMLR)*.

Angel Navarro and Francisco Casacuberta. 2023. Exploring Multilingual Pretrained Machine Translation Models for Interactive Translation. In *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track*.

Allen Newell and Stuart K Card. 1985. The prospects for psychological science in human-computer interaction. *Human-computer interaction*, 1(3):209–242.

Jingwei Ni, Julia Bingler, Chiara Colesanti-Senni, Mathias Kraus, Glen Gostlow, Tobias Schimanski, Dominik Stammbach, Saeid Ashraf Vaghefi, Qian Wang, Nicolas Webersinke, Tobias Wekhof, Tingyu Yu, and Markus Leippold. 2023. CHATREPORT: Democratizing Sustainability Disclosure Analysis through LLM-based Tools. In *Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP)*.

Ihudiya Finda Ogbonnaya-Ogburu, Angela DR Smith, Alexandra To, and Kentaro Toyama. 2020. Critical race theory for HCI. In *ACM CHI Conference on Human Factors in Computing Systems*.

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 *ACM CHI Conference on Human Factors in Computing Systems*.

Vipul Raheja, Dhruv Kumar, Ryan Koo, and Dongyeop Kang. 2023. CoEdIT: Text Editing by Task-Specific Instruction Tuning. In *Conference on Empirical Methods in Natural Language Processing: Findings (EMNLP)*.

Niroop Channa Rajashekar, Yeo Eun Shin, Yuan Pu, Sunny Chung, Kisung You, Mauro Giuffre, Colleen E Chan, Theo Saarinen, Allen Hsiao, Jasjeet Sekhon, Ambrose H Wong, Leigh V Evans, Rene F. Kizilcec, Loren Laine, Terika Mccall, and Dennis Shung. 2024. Human-Algorithmic Interaction Using a Large Language Model-Augmented Artificial Intelligence Clinical Decision Support System. In *ACM CHI Conference on Human Factors in Computing Systems*.

Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Ramona Comanescu, Canfer Akbulut, Tom Stepleton, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, and 1 others. 2024. Gaps in the safety evaluation of generative AI. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*.

Ehud Reiter. 2024. *Natural Language Generation*. Springer.

Ehud Reiter and Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558.

Sebastin Santy, Jenny T Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. NLPositionality: Characterizing design biases of datasets and models. In *Annual Meeting Of The Association For Computational Linguistics (ACL)*.

Ari Schlesinger, W Keith Edwards, and Rebecca E Grinter. 2017. Intersectional HCI: Engaging identity through gender, race, and class. In *ACM CHI Conference on Human Factors in Computing Systems*.

Omar Shaikh, Valentino Emil Chai, Michele Gelfand, Diyi Yang, and Michael S. Bernstein. 2024. Rehearsal: Simulating Conflict to Teach Conflict Resolution. In *ACM CHI Conference on Human Factors in Computing Systems*.

Shreya Shankar, JD Zamfirescu-Pereira, Björn Hartmann, Aditya Parameswaran, and Ian Arawjo. 2024. Who validates the validators? aligning llm-assisted evaluation of llm outputs with human preferences. In *ACM Symposium on User Interface Software and Technology (UIST)*.

Lin Shi, Chiyu Ma, Wenhua Liang, Weicheng Ma, and Soroush Vosoughi. 2024. Judging the judges: A systematic investigation of position bias in pairwise comparative assessments by llms. *arXiv preprint arXiv:2406.07791*.

Robert Soden, Austin Toombs, and Michaelanne Thomas. 2024. Evaluating interpretive research in hci. *Interactions*, 31(1):38–42.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. Large language models are inconsistent and biased evaluators. *arXiv preprint arXiv:2405.01724*.

Matthew Syed. 2015. *Black box thinking: why most people never learn from their mistakes–but some do*. Penguin.

Maryam Taeb, Amanda Swearngin, Eldon Schoop, Ruijia Cheng, Yue Jiang, and Jeffrey Nichols. 2024. AXNav: Replaying Accessibility Tests from Natural Language. In *ACM CHI Conference on Human Factors in Computing Systems*.

Abbas Tashakkori and John W Creswell. 2007. The new era of mixed methods.

Mohsen Tavakol and Reg Dennick. 2011. Making sense of cronbach's alpha. *International journal of medical education*, 2:53.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Craig Thomson, Ehud Reiter, and Anya Belz. 2024. Common flaws in running human evaluation experiments in NLP. *Computational Linguistics*, pages 1–11.

Robert Thornberg, Kathy Charmaz, and 1 others. 2014. Grounded theory and theoretical coding. *The SAGE handbook of qualitative data analysis*, 5(2014):153–69.

Emiel Van Miltenburg, Chris van der Lee, Thiago Castro Ferreira, and Emiel Krahmer. 2020. Evaluation rules! on the use of grammars and rule-based systems for nlg evaluation. In *Workshop on Evaluating NLG Evaluation*.

Angelina Wang, Jamie Morgenstern, and John P Dickerson. 2024a. Large language models cannot replace human participants because they cannot portray identity groups. *arXiv preprint arXiv:2402.01908*.

Sitong Wang, Savvas Petridis, Taeahn Kwon, Xiaojuan Ma, and Lydia B Chilton. 2023. PopBlends: Strategies for Conceptual Blending with Large Language Models. In *ACM CHI Conference on Human Factors in Computing Systems*.

Yidong Wang, Zhuohao Yu, Wenjin Yao, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. 2024b. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. In *International Conference on Learning Representations (ICLR)*.

Zhan Wang, Lin-Ping Yuan, Liangwei Wang, Bingchuan Jiang, and Wei Zeng. 2024c. VirtuWander: Enhancing Multi-modal Interaction for Virtual Tour Guidance through Large Language Models. In *ACM CHI Conference on Human Factors in Computing Systems*. ACM.

Xiang Wei, Yufeng Chen, Ning Cheng, Xingyu Cui, Jinan Xu, and Wenjuan Han. 2024. CollabKG: A Learnable Human-Machine-Cooperative Information Extraction Toolkit for (Event) Knowledge Graph Construction. In *Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*.

Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, and 1 others. 2023. Sociotechnical safety evaluation of generative ai systems. *arXiv preprint arXiv:2310.11986*.

Marty J Wolf, K Miller, and Frances S Grodzinsky. 2017. Why we should have seen that coming: comments on microsoft's tay" experiment," and wider implications. *ACM SIGCAS Computers and Society*, 47(3):54–64.

Austin P Wright, Zijie J Wang, Haekyu Park, Grace Guo, Fabian Sperrle, Mennatallah El-Assady, Alex Endert, Daniel Keim, and Duen Horng Chau. 2020. A comparative analysis of industry human-AI interaction guidelines. *arXiv preprint arXiv:2010.11761*.

Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. In *ACM CHI Conference on Human Factors in Computing Systems*.

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 *ACM CHI Conference on Human Factors in Computing Systems: Extended Abstracts (CHI EA)*.

Chenyang Yang, Rishabh Rustogi, Rachel Brower-Sinning, Grace Lewis, Christian Kaestner, and Tongshuang Wu. 2023. Beyond Testers' Biases: Guiding Model Testing with Knowledge Bases using LLMs. In *Conference on Empirical Methods in Natural Language Processing: Findings (EMNLP)*.

Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining whether, why, and how human-AI interaction is uniquely difficult to design. In *ACM CHI Conference on Human Factors in Computing Systems*.

Yeliz Yesilada, Giorgio Brajnik, and Simon Harper. 2009. How much does expertise matter? a barrier walkthrough study with experts and non-experts. In *International ACM SIGACCESS Conference on Computers and Accessibility*.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuan-Jing Huang. 2023. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics (ACL)*.

Ja Eun Yu and Debaleena Chattopadhyay. 2024. Reducing the Search Space on demand helps Older Adults find Mobile UI Features quickly, on par with Younger Adults. In *ACM CHI Conference on Human Factors in Computing Systems*.

Liudmila Zavolokina, Kilian Sprenkamp, Zoya Katashinskaya, Daniel Gordon Jones, and Gerhard Schwabe. 2024. Think Fast, Think Slow, Think Critical: Designing an Automated Propaganda Detection

Tool. In *ACM CHI Conference on Human Factors in Computing Systems*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yu Zhang, Jingwei Sun, Li Feng, Cen Yao, Mingming Fan, Liuxin Zhang, Qianying Wang, Xin Geng, and Yong Rui. 2024. See Widely, Think Wisely: Toward Designing a Generative Multi-agent System to Burst Filter Bubbles. In *ACM CHI Conference on Human Factors in Computing Systems*.

Runcong Zhao, Wenjia Zhang, Jiazheng Li, Lixing Zhu, Yanran Li, Yulan He, and Lin Gui. 2024. NarrativePlay: Interactive Narrative Understanding. In *Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (EACL)*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2024. Judging LLM-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems (NeurIPS)*.

Wazeer Deen Zulfikar, Samantha Chan, and Pattie Maes. 2024. Memoro: Using Large Language Models to Realize a Concise Interface for Real-Time Memory Augmentation. In *ACM CHI Conference on Human Factors in Computing Systems*.

A Methodology Details

A.1 Filtering Criteria

We surveyed 39 papers introducing human-LLM systems which were identified using the process outlined in Fig. 2. First, we collected papers from venues focused on human-computer interactions (CHI) and natural language processing (*CL released on ACL Anthology). We filtered using regular expressions to find papers that contained keywords ["human*" OR "user*"] AND "system*" AND ["large language model*" OR "LLM*"]. We also used a large language model to filter whether the abstract and title discussed a new human-LLM system using the following prompt:

```
"You are a helpful literature
review assistant whose job is
to read the below paper and help
me decide if it satisfies all my
criteria.  The criteria are:
(1) The paper presents a
human-LLM interaction system and
```

```
evaluate the system in some ways.
(2) The system described in the
paper must have human interact
with large language model in some
ways.
The paper is:
Title:  $TITLE
Abstract:  $ABSTRACT

Please give me a binary answer
(yes/no) on whether this paper
satisfies all the criteria
(you should say no as long as
any one of the criteria is not
met).  Do not include any other
explanation, the output should be
either yes or no."
```

A.2 Papers Reviewed

The outcome of the annotation using the taxonomy in Table 1 and the coding guide in Appendix B is presented in Fig. 3.

A.2.1 Papers published at an HCI Venue

Lee et al. (2024b); Lawley and Maclellan (2024); Liu et al. (2024a); Zavolokina et al. (2024); Cheng et al. (2024); Calle et al. (2024); Yu and Chattopadhyay (2024); Zulfikar et al. (2024); Liu et al. (2024c); Taeb et al. (2024); Wu et al. (2022); Wang et al. (2024c); Rajashekar et al. (2024); Arawjo et al. (2024); Wang et al. (2023); Zhang et al. (2024); Kim et al. (2024b); Fan et al. (2024); Liu et al. (2024d); Liu et al. (2023)

A.2.2 Papers published at an NLP Venue

Zhao et al. (2024); Ding et al. (2023); Ma et al. (2023); Cai et al.; Chakrabarty et al. (2022); Glória-Silva et al. (2024); Fei et al. (2024); Raheja et al. (2023); Inan et al. (2024); Wei et al. (2024); Addlesee et al. (2024); Liu et al. (2024b); Yang et al. (2023); Kim et al. (2024a); Hu et al. (2024); Luo et al. (2024); Ni et al. (2023); Navarro and Casacuberta (2023)

B Annotation Guide

We provide the codebook used to label the evaluations conducted in each paper below. This codebook is also available in our GitHub repository.
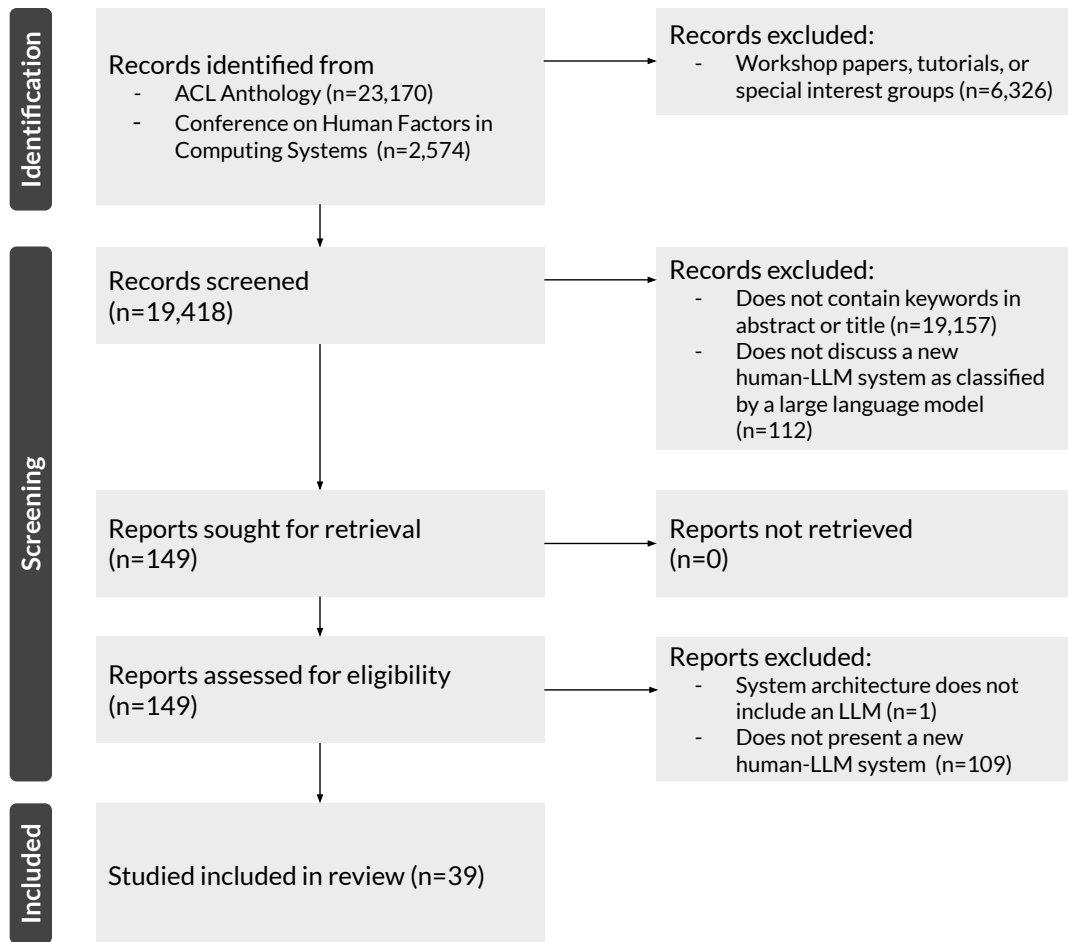
Figure 2: PRISMA diagram depicting the search strategy used to identify human-LLM systems for inclusion in our literature review.
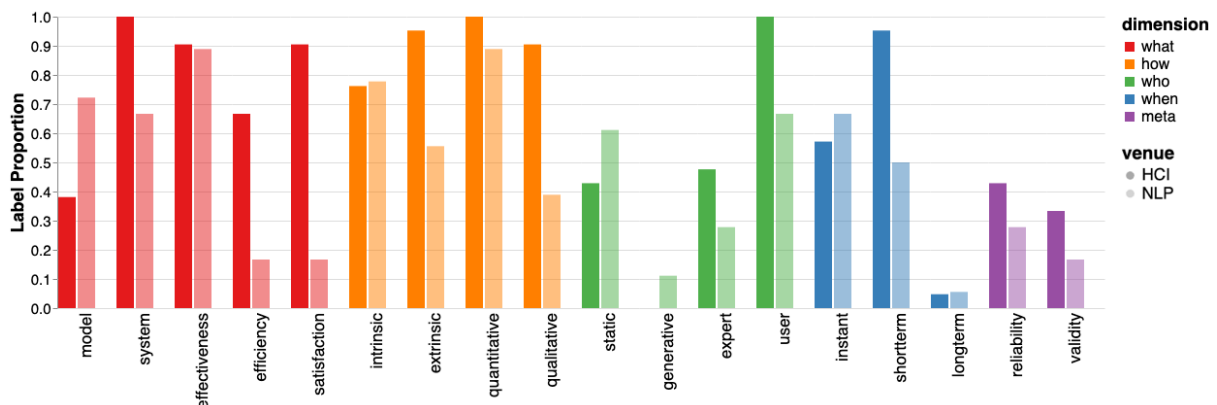

Figure 3: Distribution of evaluation annotations on the 39 papers by HCI or NLP venues using SPHERE.

B.1 What is being evaluated?

B.1.1 Component types

1. **what_model**: The evaluation focuses on only the model capabilities, including the model's performance pre-deployment in traditional benchmarking settings and performance in-situ as users continue to interact with the model. Mark 1 if the

authors evaluate the model and 0 if not.

- For example: The authors benchmark the performance of a fine-tuned model used in their system.

2. **what_system**: The evaluation covers the system as a whole to understand how these design choices may impact users' experiences, includ-

1357

ing different layers of the system, such as interfaces or interactions. Mark 1 if the authors evaluate the system and 0 if not.

B.1.2 Design goals

For each design goal, mark 1 if the authors include evaluation that covers the concept and 0 otherwise.

1. **what_effectiveness**: Evaluating the accuracy, completeness, and lack of negative consequences with which users achieved specified goals.

   - For example: Are users able to successfully complete the task the system is designed for? What is the performance of the model?

2. **what_efficiency**: Evaluating resources (such as time or effort) needed by users to achieve their goals.

   - For example: How long did it take for a user to complete the task? What was the cognitive burden or mental load required to complete the task?

3. **what_satisfaction**: Evaluating positive attitudes, emotions, and/or comfort resulting from use of a system, product, or service.

   - For example: What was the user's satisfaction with the overall system or parts of the system? How much did the user trust the system?

B.2 How is the evaluation being conducted?

B.2.1 Scope

1. **how_intrinsic**: Assessing the system or the internal model components on specific tasks that they are designed to perform, by evaluating how well they achieve these tasks according to some predefined criteria or benchmarks. Mark 1 if they include intrinsic evaluation and 0 if not.

   - For example, in a writing assistant system, intrinsic evaluation might involve assessing the system's accuracy on grammatical correctness with automatic metrics or how users might rate the functionality of different system features.

2. **how_extrinsic**: Measuring the effectiveness of the system in the context of its application in real-world scenarios, when interacting with users. Mark 1 if they include extrinsic evaluation and 0 if not.

   - For example, in a writing assistant system, extrinsic evaluation might involve recruiting users to co-write with the system and seeing how this impacts their writing style, productivity, and satisfaction.

B.2.2 Method

1. **how_quantitative**: Measuring and analyzing numerical data to assess system performance and impact. Examples include measuring Likert scale ratings or running benchmark evaluations on the model. Mark 1 if the authors included any quantitative methods and 0 if not.

2. **how_qualitative**: Analyzing non-numerical data to gain deeper insights into user experiences, perceptions, and the contextual factors influencing system performance. Examples include conducting semi-structured interviews with users. Mark 1 if the authors included any qualitative methods and 0 if not.

   - For example, in NarrativePlay, the authors include some brief analysis of responses. However, we do *not* count listing examples as qualitative analysis.

B.3 Who is participating in the evaluation process?

B.3.1 Automated Evaluators

Use these tags *only* when evaluation not involving human participants is used.

1. **who_static**: Mark 1 if any static evaluation not directly performed by a human or LLM is included and 0 if not. For example, benchmarking a model's capability is an example of static evaluation.

2. **who_generative**: Mark 1 if the authors use a language model in a generative capacity and 0 if not for evaluation. Examples include simulating participants with LLMs, using LLM to annotate and rate text, or using LLM-as-a-judge. Using a technique like BERTScore is *not* generative since it is embedding-based.

B.3.2 Human Evaluators

Use these tags *only* when human evaluation is used.

1. **who_expert**: If the human evaluator is a domain expert or has equivalent expertise in the area that the system is designed for. Mark 1 if expert evaluators are included and 0 otherwise.

- For example, if the system is a tutoring system, a teacher would be considered an expert.
- In AngleKindling, which helps journalists come up with framings for papers, the evaluation is conducted with NYC journalists, who are domain experts in this field.

2. **who_user**: If the evaluator is a direct user or target audience of the system. Mark 1 if the intended user is included as a human evaluator and 0 otherwise.

- For example, for a student-facing tutoring system, student evaluators will be the design target, but a teacher will not.
- In AngleKindling, the journalists are also the intended users for the system, so we would mark 1.
- If you have a general-purpose system, any user (including crowdworkers, PhD students) would be a user.

B.4 When is evaluation conducted?

The time-scale over which the evaluation occurs.

1. **when_immediate**: Evaluating real-time or immediate interactions. Mark 1 if any immediate evaluation is conducted and 0 otherwise.

2. **when_shortterm**: Evaluation is deployed for a short duration of time to measure short-term benefits of using the system. Mark 1 if any short-term evaluation is conducted and 0 otherwise.

3. **when_longterm**: Evaluation is conducted over a longer duration of time. This is typically done to understand long-term behavioral changes, practical feasibility, etc. Mark 1 if any long-term evaluation is conducted and 0 otherwise.

B.5 How is evaluation validated?

The methods for ensuring the reliability and validity of the evaluation methods. Mark 1 only if the authors *explicitly* mention any techniques for reliability and validity.

1. **reliability**: Mark 1 if the authors include techniques or measures to ensure that the evaluation judgments are consistent.

- For example, looking at internal consistency (e.g., inter-rater reliability, Cronbach alpha, split-half reliability), consistency over time (test-retest reliability), and reproducibility of results.

2. **validity**: Mark 1 if the authors include techniques or measures to ensure that the evaluation method measures the correct constructs.

- For example, methods include removing human biases using experiment designs, using statistical methods like factor analysis, mentioning ecological validity in their experiment setup, etc.

C Additional Results

C.1 Paper Annotations

We provide the annotations for the 39 papers we reviewed using the SPHERE framework in Table 2 and Table 3. The annotations are also available on our website and GitHub repository.

C.2 Quotations

We provide quotations from papers in our literature review describing their qualitative methods in Table 4.

D Case Studies

We present examples of evaluation cards for two human-AI systems: LearnLM (Jurenka et al., 2024) (Fig. 4) and AngleKindling (Petridis et al., 2023) (Fig. 5). We selected these two systems as examples since they span different application domains (education and journalism) and different modes of model development commonly observed in current human-AI systems, ranging from fine-tuning and aligning LLM (LearnLM) to prompt engineering (AngleKindling).

D.1 Takeaways

Using SPHERE, in addition to describing existing system evaluations, we can also identify areas of improvement in evaluation practices for our case studies: extrinsic measures, long-term evaluation, and validations.

First, both cases provide both intrinsic and extrinsic evaluations, but there is still a lack of *quantitative extrinsic* evaluation for the overall *effectiveness* and *efficiency* of the *system*. The current quantitative extrinsic evaluation focuses on self-perceived ratings of the systems' performance. While capturing users' perceptions is important, self-perceived ratings may be unreliable and subject to human bias. Including objective, quantitative measures can provide a more holistic picture of system performance.

System	Model	System	Effective	Efficiency	Satisfac.	Intrinsic	Extrinsic	Quant.	Qual.
Addlesee et al. (2024)	✓	✓	✓			✓		✓	
Calle et al. (2024)		✓	✓	✓		✓		✓	
Ding et al. (2023)		✓	✓	✓	✓	✓	✓	✓	✓
Inan et al. (2024)	✓	✓	✓		✓	✓	✓	✓	
Liu et al. (2023)		✓	✓	✓	✓	✓	✓	✓	✓
Navarro and Casacuberta (2023)	✓		✓			✓		✓	
Rajashekar et al. (2024)		✓		✓	✓		✓	✓	✓
Wu et al. (2022)		✓	✓	✓	✓		✓	✓	✓
Zhang et al. (2024)		✓	✓		✓		✓	✓	✓
AXNav (Taeb et al., 2024)		✓	✓		✓	✓	✓	✓	✓
ChainForge (Arawjo et al., 2024)		✓	✓		✓		✓	✓	✓
ChatReport (Ni et al., 2023)	✓		✓			✓		✓	✓
ClarifAI (Zavolokina et al., 2024)		✓	✓	✓	✓	✓	✓	✓	✓
CoEdit (Raheja et al., 2023)	✓	✓				✓		✓	
CollabKG (Wei et al., 2024)		✓	✓	✓			✓	✓	
ContextCam (Fan et al., 2024)		✓	✓		✓	✓	✓	✓	✓
CoPoet (Chakrabarty et al., 2022)	✓	✓	✓			✓	✓	✓	
CoQuest (Liu et al., 2024d)		✓	✓	✓	✓	✓	✓	✓	✓
Dialight (Hu et al., 2024)	✓		✓			✓	✓	✓	
DuetSim (Luo et al., 2024)	✓		✓			✓	✓	✓	✓
EmpathyEar (Fei et al., 2024)	✓	✓				✓		✓	
EvalLM (Kim et al., 2024b)	✓	✓	✓	✓	✓	✓	✓	✓	✓
Human I/O (Liu et al., 2024c)		✓	✓	✓	✓	✓	✓	✓	✓
InsightPilot (Ma et al., 2023)		✓	✓				✓	✓	✓
Low-code LLM (Cai et al.)		✓	✓				✓		✓
MEGAnno+ (Kim et al., 2024a)		✓	✓				✓		✓
Memoro (Zulfikar et al., 2024)	✓	✓	✓	✓	✓	✓	✓	✓	✓
NarrativePlay (Zhao et al., 2024)	✓		✓			✓		✓	
NavNudge (Yu and Chattopadhyay, 2024)		✓	✓	✓	✓		✓	✓	
PaperWeaver (Lee et al., 2024b)	✓	✓	✓	✓	✓	✓	✓	✓	✓
PlanLLM (Glória-Silva et al., 2024)	✓	✓	✓			✓		✓	
PopBlends (Wang et al., 2023)	✓	✓	✓	✓	✓	✓	✓	✓	✓
Proofread (Liu et al., 2024b)	✓		✓			✓		✓	
Rehearsal (Shaikh et al., 2024)	✓	✓	✓			✓	✓	✓	✓
RELIC (Cheng et al., 2024)	✓	✓	✓	✓	✓	✓	✓	✓	✓
Selenite (Liu et al., 2024a)	✓	✓	✓	✓	✓	✓	✓	✓	✓
VAL (Lawley and Maclellan, 2024)	✓	✓	✓		✓	✓	✓	✓	✓
VirtuWander (Wang et al., 2024c)		✓			✓	✓	✓	✓	✓
Weaver (Yang et al., 2023)	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 2: Annotations on the 39 human-LLM systems covering the dimensions of what is being evaluated and how evaluation is conducted.

System	User	Expert	Static	Generative	Immediate	Short	Long	Reliable	Valid
Addlesee et al. (2024)	✓				✓				
Calle et al. (2024)	✓		✓	✓	✓				
Ding et al. (2023)		✓			✓	✓			✓
Inan et al. (2024)	✓		✓	✓	✓	✓	✓		
Liu et al. (2023)	✓	✓				✓		✓	
Navarro and Casacuberta (2023)	✓				✓				
Rajashekar et al. (2024)	✓	✓			✓	✓			✓
Wu et al. (2022)	✓	✓			✓	✓			
Zhang et al. (2024)		✓				✓			
AXNav (Taeb et al., 2024)	✓	✓	✓		✓	✓		✓	
ChainForge (Arawjo et al., 2024)		✓				✓			
ChatReport (Ni et al., 2023)	✓							✓	
ClarifAI (Zavolokina et al., 2024)	✓	✓				✓		✓	
CoEdit (Raheja et al., 2023)	✓		✓		✓				
CollabKG (Wei et al., 2024)				✓		✓			
ContextCam (Fan et al., 2024)		✓	✓		✓	✓	✓		
CoPoet (Chakrabarty et al., 2022)	✓			✓	✓	✓		✓	✓
CoQuest (Liu et al., 2024d)	✓	✓				✓		✓	
Dialight (Hu et al., 2024)	✓			✓	✓				
DuetSim (Luo et al., 2024)		✓	✓		✓				
EmpathyEar (Fei et al., 2024)	✓			✓	✓				
EvalLM (Kim et al., 2024b)	✓	✓	✓		✓	✓		✓	
Human I/O (Liu et al., 2024c)		✓	✓		✓	✓		✓	✓
InsightPilot (Ma et al., 2023)		✓				✓			
Low-code LLM (Cai et al.)									
MEGAnno+ (Kim et al., 2024a)	✓	✓				✓			
Memoro (Zulfikar et al., 2024)		✓			✓	✓			✓
NarrativePlay (Zhao et al., 2024)		✓		✓	✓	✓		✓	
NavNudge (Yu and Chattopadhyay, 2024)				✓		✓			✓
PaperWeaver (Lee et al., 2024b)	✓	✓				✓		✓	✓
PlanLLM (Glória-Silva et al., 2024)	✓	✓		✓	✓	✓		✓	
PopBlends (Wang et al., 2023)	✓	✓				✓		✓	
Proofread (Liu et al., 2024b)	✓								
Rehearsal (Shaikh et al., 2024)		✓	✓		✓	✓		✓	
RELIC (Cheng et al., 2024)		✓	✓		✓	✓			
Selenite (Liu et al., 2024a)		✓	✓		✓	✓			✓
VAL (Lawley and Maclellan, 2024)		✓	✓		✓	✓			
VirtuWander (Wang et al., 2024c)		✓				✓			
Weaver (Yang et al., 2023)	✓	✓	✓		✓	✓		✓	✓

Table 3: Annotations on the 39 human-LLM systems covering the dimensions of who is participating in the evaluation, when is the evaluation (duration), and how evaluation is evaluated.

For example, in LearnLM, experts provided ratings for each turn generated during a tutoring session. We could also measure changes in student performance using pre-post test scores, grades, or drop-out rates.

Second, both studies only perform immediate or short-term evaluations. Thus, we do not know the downstream impact that these systems might have. For example, for AngleKindling, a longitudinal evaluation could help us understand how journalists integrate the system into their workflow and the potential impact on productivity or the types of stories being written.

Finally, while there is some acknowledgment of meta-evaluation done in the case of LearnLM, overall discussion on validating the evaluation paradigm is not common.

SPHERE Evaluation Card: LearnLM-Tutor

System Overview: *LearnLM-Tutor (Jurenka et al., 2024) is conversational AI designed to serve as a personalized tutor for learners and a teaching assistant for educators. The model component is a fine-tuned Gemini 1.0 (Team et al., 2023) using a custom dataset. LearnLM-Tutor is evaluated using seven different methods of which we only present two here as examples: turn-level pedagogy (TLP, §5.2), and language model evaluation (LME, §6.1).*

What is being evaluated?
- **Component:** TLP & LME both evaluated the *model* component of LearnLM-Tutor.
- **Design Goal:** Evaluate LearnLM-Tutor's *effectiveness* (TLP: how good is each turn's pedagogy use, and LME: how well it performs on different pedagogical tasks).

How is evaluation conducted?
- **Scope:** For both TLP & LME, an *intrinsic* evaluation is conducted to examine the LearnLM model's capacity.
- **Method:** For TLP, each turn of unguided tutoring sessions between real learners and either LearnLM-Tutor or Gemini 1.0 was rated on yes/no/na for nine items of customized pedagogy rubrics (e.g., promotes engagement, monitors motivation, etc.). Welch t-test with Holm-Bonferroni adjustment was used to compare the turn-level *quantitative* scores of LearnLM-Tutor and Gemini 1.0.
  For LME, they make the LLM critic generate *quantitative* binary scores on an expert-curated dataset of different customized pedagogical tasks (e.g., stay on topic, don't reveal the answer, etc.), using task-specific prompts (including task description, reference answer, context, and the generated response).

Who is participating in the evaluation?
- **Automated:** PaLM 2.0 is used as a *generative* evaluator for the model for LME.
- **Human:** Approximately 60 human pedagogical *experts* were recruited to give ratings for TLP (*intended users* for LearnLM-Tutor would be the students). There were about 1.5k overall turns for each model rated by at least three different raters, and a majority vote was used for each model's response.

When is evaluation conducted?
- **Time Scale**: LME is conducted *immediately* since it uses a *generative* evaluator. For TLP, there is no explicit mention of the time it takes to label each turn for a human expert, so we mark it as *immediate* as well.

How is evaluation validated?
- **Validation**: For TLP, the Krippendorff's α of 0.3 across all attributes showed a low *inter-rater reliability*. For LME, the paper reported that the average LME score highly correlates with humans for different iterations of the tutor model. However, in terms of *validity*, the paper lacks details on the score calculation, correlation statistics, and how the pedagogy rubrics for LME were developed.

Turn-Level Pedagogy (TLP, §5.2)

What is being evaluated?

Model component's *effectiveness*: how good is each turn's pedagogy use.

How is evaluation conducted?

Intrinsic evaluation: each turn of unguided tutoring sessions was rated on yes/no/na for nine rubric items (e.g., promotes engagement, monitors motivation, etc.). Welch t-test with Holm-Bonferroni adjustment was used to compare the turn-level *quantitative* scores of LearnLM-Tutor and Gemini 1.0.

Who is participating in the evaluation?

Approximately 60 pedagogical *experts*. About 1.5k overall turns rated by at least three different raters, and a majority vote was used for each model's response.

When is evaluation conducted?

No explicit mention of time it takes to label each turn for a human expert, so we mark it as *immediate*.

How is evaluation validated?

The Krippendorff's α of 0.3 across all attributes showed a low *inter-rater reliability*

Language Model Evaluation (LME, §6.1)

What is being evaluated?

Model component's *effectiveness*: how well it performs on different pedagogical tasks.

How is evaluation conducted?

Intrinsic evaluation: LLM critic generates *quantitative* binary scores on an expert-curated dataset of different customized pedagogical tasks (e.g., stay on topic, don't reveal the answer, etc.), using task-specific prompts (including task description, reference answer, context, and the generated response).

Who is participating in the evaluation?

PaLM 2.0 is used as a *generative* evaluator for the model for LME.

When is evaluation conducted?

Immediate since it uses a *generative* evaluator.

How is evaluation validated?

Average LME score highly correlates with humans for different iterations of the tutor model. However, the paper lacks details on the *validity* for score calculation, correlation statistics and rubrics development.

Figure 4: Example SPHERE evaluation card for LearnLM-Tutor (Jurenka et al., 2024). One can apply SPHERE with one card per human-AI system as in Fig. 5, or one card per evaluation method for cleaner separation.

Figure 5: Example SPHERE evaluation card for AngleKindling (Petridis et al., 2023).

"We analyzed the transcripts through a combination of inductive thematic analysis through affinity diagramming, augmented with a spreadsheet to list participants' ideas, behaviors (nodes added, process of their exploration, whether they imported data, etc), and answers to post-interview questions. For our in-lab study, three coauthors separately affinity diagrammed three transcripts each, then met and joined the clusters through mutual discussion. The merged cluster was iteratively expanded with more participant data until clusters reached saturation. For interviews, the first author affinity diagrammed all transcripts to determine themes."– Arawjo et al. (2024)

"The research team also took field notes during the session and used the notes to guide the analysis...We performed a thematic analysis on the qualitative data from the user study. Two authors of the paper first individually coded all the transcripts, then presented the codes to each other and collaboratively and iteratively constructed an affinity diagram of quotes and codes together to develop themes."– Taeb et al. (2024)

"Think-aloud data was primarily used for understanding how users generated and interpreted RQs. One researcher first generated a codebook through open coding using videos and transcripts from three randomly selected participants, and then three other researchers independently coded the data of the same three participants, reaching an inter-rater agreement of 0.83 in Krippendorf's alpha. The annotators then discussed and refined the codebook again until they reached full agreement. Then, four researchers proceeded to annotate the remaining 17 participants' behavior data separately. In the final codebook, whether users interacted with the system was annotated and used for quantitative analysis in RQ3 as "Acted During Wait". The final codebook also included sense-making behavior (e.g., reasons for (not) waiting, reason for providing certain feedback) as qualitative results."– Liu et al. (2024d)

"All study sessions were recorded and transcribed. Two authors read through the text script of three randomly selected participants together to understand their user experience of the prototype. Then, they independently coded the script using an open-coding approach. They combined deductive and inductive coding techniques to form the codebook. The two coders regularly discussed the codes and resolved disagreements to create a consolidated codebook. Further meetings were scheduled with the whole research team to discuss the codes and how they should be grouped into themes. The whole team iterated on the codes and their grouping until they reached consensus. In the end, we arrived at four themes: overall user behavioral patterns, engagement, diverse information, and in-depth information processing"– Zhang et al. (2024)

Table 4: Set of examples of how qualitative methods were described in papers from HCI venues.