User-Centric Design Paradigms for Trust and Control in Human-LLM-Interactions: A Survey

Milena Belosevic
Bielefeld University
milena.belosevic@uni-bielefeld.de

Abstract

As LLMs become widespread, trust in their behavior becomes increasingly important. For NLP research, it is crucial to ensure that not only AI designers and developers, but also end users, are enabled to control the properties of trustworthy LLMs, such as transparency, privacy, or accuracy. However, involving end users in this process remains a practical challenge. Based on a design-centered survey of methods developed in recent papers from HCI and NLP venues, this paper proposes seven design paradigms that can be integrated in NLP research to enhance end-user control over the trustworthiness of LLMs. We discuss design gaps and challenges of applying these paradigms in NLP and propose future research directions.

1 Introduction

While LLMs bring many advantages, their opacity hinders human agency and trust, as especially end users lack the necessary information and transparency to critically assess system decisions before following or acting on them (Förster et al., 2020). For this reason, there is a growing need in the field of NLP to develop methods that enhance end-user control over AI systems.

At the same time, in the domain of human-computer interaction (HCI), approximately 22 regulations comprising normative principles for mitigating AI risks and enhancing trust in AI systems had been published by 2020 (Hagendorff, 2020). While recent HCI studies explore the attitudes of different groups towards these policies (Agbese et al., 2023), their practical implementation is underexplored (Kaur et al., 2022; Perov and Golovkov, 2024), particularly regarding how to enable end users to proactively participate in controlling the trustworthiness of LLM systems.

This paper proposes seven design paradigms for enhancing end-user control over the trustworthiness of LLM systems based on a design-centered survey of novel methods from recent HCI and NLP studies. We define trustworthy LLMs using the following requirements for trustworthy AI proposed by Ethics Guidelines for Trustworthy AI (HLEG, 2019): (1) *human agency and oversight* (including fundamental rights); (2) *technical robustness and safety* (including resilience to attack and security, fallback plan and general safety, accuracy, reliability, and reproducibility); (3) *privacy and data governance* (including respect for privacy, quality and integrity of data, and access to data); (4) *transparency* (including traceability, explainability, and communication); (5) *diversity, non-discrimination* and *fairness* (including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation); (6) *environmental and societal well-being* (including sustainability and environmental friendliness, social impact, society, and democracy) and (7) *accountability* (including auditability, minimization, and reporting of negative impact, trade-offs, and redress). Although these guidelines were published before the rise of LLMs, we use them to define trust in LLMs because they were among the first to foreground a user-centered approach to trustworthy AI (Usmani et al., 2023) and remain an influential user-centered policy.

This survey contributes to human-centered approaches to LLMs by bridging regulatory perspectives on trustworthy AI from the field of HCI with their practical applications in NLP research on end users' interactions with LLMs.

2 Methodology

We surveyed original research papers (no work in progress, demo papers, posters, provocations, surveys, or extended abstracts) published in English in the ACM Digital Library and the ACL Anthology between January 1, 2022, and August 1, 2025. The start date was selected to include papers published shortly before the release of ChatGPT on November 30, 2022. While the ACM library was selected for its comprehensive coverage of HCI design research and venues (e.g., CHI) relevant to our focus, the ACL Anthology comprises work from some of the most important NLP venues, such as EMNLP and NAACL. This dual-sourced corpus provides a balanced foundation for identifying design patterns at the intersection of HCI and NLP. The following search string was used for the ACM library:

> *trust* OR “agency” OR “oversight” OR robust* OR safe* OR secur* OR accura* OR reliab* OR reproduc* OR “privacy” OR transparen* OR trace* OR explain* OR fair* OR bias* OR sustain* OR accountab* OR audit* or LLM*

The search function in ACL Anthology is limited to simple keyword queries and does not support using this search string. Therefore, we performed multiple keyword-based searches (e.g., trust LLM, transparency LLM, bias LLM) and complemented this with Google site:aclanthology.org searches to approximate Boolean logic and ensure broader coverage. A total of 1781 papers were screened from both databases. At least one of the search words had to appear in the abstract, the title, or the keywords of the paper to be included in the final dataset.

Importantly, the papers that fulfill this criterion were manually inspected to determine whether they have a clear focus on both trust (i.e., the trust aspect mentioned in the abstract) and end-user control in the full text. Accordingly, user-centric papers without a clear relationship to trust and vice versa: trust-related papers without end-user involvement in the design and/or evaluation stage (e.g., Miao and Fang 2025) or papers where the evaluation is conducted based only on datasets, performance comparison of several models, and evaluation metrics, rather than involving users and explicitly addressing how user control is achieved, were not considered. However, papers combining user studies with, for example, comparing the performance of several models, were considered (e.g., Zhou et al. 2024; Koraş et al. 2025; Dong et al. 2025).

Also, papers not explicitly dealing with language models or language model-based applications were excluded (e.g., DeVos et al. 2022). These criteria reduced the number of eligible papers to 773.

Finally, papers that considered user studies as future work (e.g., Hung et al. 2023) were excluded.

In this way, 713 papers were excluded. The final list comprises 60 papers from both sources.

Papers did not need to explicitly address the AI HLEG guidelines, nor did we include studies that analyzed the guidelines themselves. Multimodal LLMs (Zhang et al., 2024; Tang et al., 2024; Chen et al., 2024) were discarded due to the broader use of text-based models. Figure 1, created with a web-based Shiny app (Haddaway et al., 2022), visualizes the PRISMA-compliant search process (Page et al., 2021).

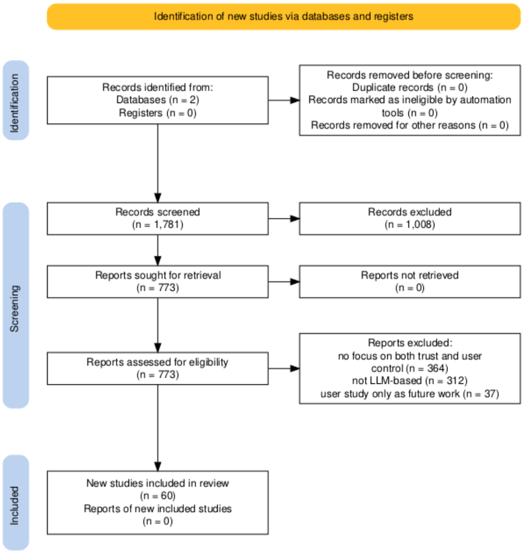
Identification of new studies via databases and registers
Identification
Records identified from: Databases (n = 2) Registers (n = 0)
Records removed before screening: Duplicate records (n = 0) Records marked as ineligible by automation tools (n = 0) Records removed for other reasons (n = 0)
Screening
Records screened (n = 1,781)
Records excluded (n = 1,008)
Reports sought for retrieval (n = 773)
Reports not retrieved (n = 0)
Reports assessed for eligibility (n = 773)
Reports excluded: no focus on both trust and user control (n = 364) not LLM-based (n = 312) user study only as future work (n = 37)
Included
New studies included in review (n = 60) Reports of new included studies (n = 0)

Figure 1: Overview of the literature search and screening process, following PRISMA-style structure.

Note that this is not a systematic review aiming for completeness, but a design-centered survey of recent work focusing on a synthesis of paradigms that support user control and trust in LLM systems.

Two annotators searched for the papers in the databases described above. They discussed and refined the inclusion criteria following the PRISMA paradigm (Page et al., 2021). The included papers were then annotated for their primary trust aspects (multiple assignment was allowed), and design paradigm. Annotation decisions were reached via iterative discussion. No formal inter-annotator agreement was calculated, as the focus was on interpretive synthesis. As a result, seven design paradigms are proposed (Section 3) and discussed in terms of their applications in NLP (Section 5). Note that although multiple paradigm assignments were theoretically possible, each paper was assigned to exactly one primary design paradigm based on annotator agreement.

Table 1: Design paradigms and primary user goals

Paradigm	Primary User Goal
Interface-level accuracy control	Verify factual correctness of LLM output
Workflow-aligned & domain-adapted LLM assistance	Maintain control in expert workflows
Explanation-centered approaches	Understand how and why outputs are produced
Participatory designs	Learn about LLMs; shape behavior
Interactive authoring & co-creation	Co-generate or revise outputs with AI
Style-based trust calibration	Calibrate trust based on how outputs are expressed
Privacy-aware architectures & tools	Control what personal data is exposed

3 Results

Table 1 provides an overview of design paradigms identified through inductive coding by two annotators (see Section 2). A detailed mapping of the reviewed approaches to trust aspects is provided in the Appendix. The observed skew in the distribution of papers across paradigms (interactive authoring/co-creation (13 papers) and explanation-centered approaches (12) vs. interface-level accuracy control and privacy-aware tools, four studies each) may in part reflect the methodological choices of this survey, such as the single-label annotation protocol (see Section 2). Furthermore, trust dimensions, such as accuracy, transparency, and reliability, are overrepresented probably because they are easier to operationalize through measurable interventions (e.g., confidence scores), aligning well with existing evaluation practices in NLP and HCI. In contrast, underrepresented dimensions, such as environmental and societal wellbeing, require long-term stakeholder engagement and more resource-intensive methods that are harder to implement within the scope of typical research prototypes.

Interface-level accuracy control Interface-level accuracy control refers to design approaches that equip users with interactive tools and visual cues at the interface level to help inspect, verify, and guide the factual accuracy of LLM outputs. These interfaces do not require altering the model itself, but instead focus on enhancing user control, interpretability of outputs, and trust calibration through features such as consistency checks, confidence scores, source attributions, and interactive verification workflows.

The primary goal of this paradigm is to foster accuracy and user agency by integrating transparent control mechanisms directly into the interface rather than modifying the LLM architecture.

Core strategies include tools for output verification, hallucination detection, and user-led content auditing. For example, Cheng et al. (2024) enable users to compare the factual consistency of multiple LLM outputs. Laban et al. (2024) introduce a factual editing framework that alerts users to new content, supports verification via web search, and enables tracing of model-generated edits. Formal verification has also been integrated into LLM planning tasks: Lee et al. (2025) combine model checking with user oversight. Other interfaces visualize hallucination risks or confidence scores to help users identify unreliable content (Leiser et al., 2024).

Despite promising interaction designs, this paradigm faces several challenges. First, many studies prioritize surface-level model accuracy without systematically examining how interface interventions influence other trust dimensions such as fairness, transparency, or robustness. Second, tools like confidence scores (Leiser et al., 2024) assume a high degree of AI literacy and decision-making capacity, potentially excluding non-expert users or overburdening them with the responsibility to correctly interpret, evaluate, and act on information provided by an AI system. Third, the usability and cognitive demands of these systems remain under-evaluated, as it is often unclear whether users meaningfully benefit from features like verification workflows or simply ignore them in practice.

Workflow-aligned and domain-adapted LLM assistance This design approach integrates LLMs into real-world tasks or professional practices, such as education (Kazemitabaar et al., 2024), qualitative analysis (Dai et al., 2023), legal consultation (Hu et al., 2024), banking (Gupta et al., 2025), coding (Dong et al., 2025), or clinical settings (Koraş et al., 2025), addressing domain-specific challenges of the LLM application. Users are typically given mechanisms to adapt, guide, or verify outputs in-situ, through plan-then-execute pipelines (He et al., 2025), interface-level guardrails (Liffiton et al., 2023), or feedback loops involving humans in iterative roles (Dong et al., 2025; Dai et al., 2023). Unlike generic chat interfaces, these systems align generation with domain goals and domain-specific verification routines and constraints.

The goal is to integrate LLMs into domain-specific workflows in ways that preserve user control, ensure output reliability, and align with domain-specific goals.

Examples include the restriction of LLM outputs to pseudocode in educational contexts to prevent over-reliance and support learning (Kazemitabaar et al., 2024), real-time human feedback (Gupta et al., 2025), iterative human verification Dong et al. (2025), or guardrails that prevent programmers' over-reliance (Liffiton et al., 2023), collaborative human-LLM thematic analysis and topic modeling (Dai et al., 2023; Akter et al., 2025; Choi et al., 2024).

While these paradigms offer promising forms of human-in-the-loop control, several limitations remain. First, they often assume static domain knowledge and well-formed tasks and do not adapt to rapid changes in domains like coding. Second, despite placing high cognitive demands on users (e.g., verifying assertions (Dong et al., 2025) or interpreting multi-step plans (He et al., 2025)), most designs treat users as uniformly skilled and do not assess or adjust for varying levels of domain expertise and AI literacy. This creates risks of misalignment between tool complexity and user capability and of mismatched support (either under-serving novice users or constraining experts). Finally, the integration of these designs in professional workflows raises epistemic and normative concerns since the normative assumptions integrated in designs (e.g., what counts as a "good" summary or acceptable pseudocode) are rarely made explicit or empirically evaluated. As a result, these designs may reinforce domain conventions (e.g., legal templates) without enabling critical reflection, for example in qualitative analysis (Akter et al., 2025; Choi et al., 2024).

In sum, workflow-aligned assistance offers a promising direction for domain-specific LLM use, but often relies on hidden assumptions about task stability, user capability, and normative correctness. Future work should investigate how designs could better adapt to user diversity and task ambiguity.

Participatory designs Participatory designs aim to empower users through learning and reflection, engaging them not just as passive recipients of AI output but as active collaborators, educators, assessors, or learners.

The goal is to foster AI literacy, critical awareness of LLM capabilities, and trust calibration by giving users tools to customize, question, and steer LLM behavior, particularly in educational, reflective, or interpersonal contexts.

Common strategies include user-controlled editable outputs (Chun et al., 2025), scaffolded interaction via AI literacy workshops (Theophilou et al., 2023), user-led evaluation through comparisons, subjective trust metrics (Pan et al., 2024; Zhu et al., 2025; Nguyen et al., 2024), or human-LLM evaluation of social appropriateness (Rao et al., 2025). Expert-in-the-loop approaches include collaborative prompt refinement for educational content (Reza et al., 2025) or feedback-driven role-play simulation in counseling (Louie et al., 2024).

Despite the user-centered intent, several gaps persist. First, participatory mechanisms are often introduced without sufficient onboarding or scaffolding. Users are asked to judge, configure, or collaborate with LLMs before acquiring a conceptual understanding of model behavior, which may lead to overtrust. AI literacy, while a core aim, is rarely embedded as a design prerequisite—Theophilou et al. (2023) being a notable exception. Second, customization and feedback are typically limited to surface-level tuning (e.g., tone or behavior), with little support for questioning underlying assumptions, biases, or system limitations. Third, while many systems frame participation as empowering, they may implicitly rely on user labor, placing the burden of correction, verification, and ethical reflection onto the user without adequate institutional or system-side accountability.

Overall, participatory designs signal a significant shift toward user agency and transparency, but remain underdeveloped in terms of empowering user AI literacy and critical engagement with model limitations.

Interactive authoring and co-creation This paradigm focuses on enabling users to collaborate with LLMs during complex or creative tasks (e.g., writing, prompt design, workflow creation) by enabling real-time interaction, iterative refinement, and mixed-initiative control. These systems support back-and-forth exchanges between users and LLMs, allowing users to guide, steer, edit, or evaluate intermediate outputs through customizable workflows.

The primary goal is to empower users as co-authors, prompt designers, or evaluators in creative or analytical tasks by enabling interactive, transparent, and customizable collaboration with LLMs. These systems seek to enhance human agency, reduce cognitive load, and make LLM-powered generation more interpretable and aligned with users' goals and values.

This paradigm centers on prompt chaining (Arawjo et al., 2024; Wu et al., 2022), co-auditing LLM-behavior in general (Rastogi et al., 2023), or LLM-generated biases (Prabhudesai et al., 2025) and personality traits (Zheng et al., 2025a) in particular, LLM- and human-based disinformation evaluation (Zugecova et al., 2025), co-creative authoring (Ding et al., 2023; Liu et al., 2024; Hoque et al., 2024), direct manipulation (Masson et al., 2024), and mixed-initiative control, enabling users to collaboratively shape LLM behavior via initiative-sharing interfaces (Overney et al., 2025), LLM-initiated prompt pipelines (Zhang and Arawjo, 2025) and editable preference profiles created based on user preferences (Liu et al., 2025).

Despite their promise, interactive authoring designs raise several unresolved questions. First, while many interfaces emphasize modularity, prompt chaining, or editable outputs (Arawjo et al., 2024; Wu et al., 2022; Zhang and Arawjo, 2025), it remains unclear how much initiative users actually retain in practice. Systems often alternate initiative without clearly defining the boundaries of user agency, and few studies examine whether users can override or put the model's underlying assumptions into question. Customization is usually limited to surface-level, such as prompt components, without affording deeper user control or interpretability of generation mechanisms.

Second, user literacy and feedback quality are assumed rather than supported. Designs empower users to filter outputs, flag disinformation, or assess persuasiveness (Zugecova et al., 2025; Liu et al., 2025), but offer limited scaffolding to support critical evaluation. Since there are no clear scaffolds for critical reflection, user perception of biases or auditing personality traits (Zheng et al., 2025a; Prabhudesai et al., 2025) risks being subjective and culturally dependent.

Third, while some systems highlight transparency and provenance (e.g., via interface visualizations or think-aloud protocols (Hoque et al., 2024; Rastogi et al., 2023)), it remains unclear whether such interventions are always desirable and whether more transparency always leads to better trust calibration.

Finally, there is limited evidence that these approaches generalize beyond low-stakes, exploratory domains. Many studies involve small participant samples (e.g., thirteen participants in Hoque et al. 2024), leaving open the question of how co-creation behaves under real-world constraints such as time pressure or conflicting user goals.

In sum, interactive authoring represents a promising design approach to expanding human–AI collaboration, but current work underestimates the dynamics of control and overlooks users' cognitive limitations.

Explanation-centered approaches Explanation-centered approaches aim to make LLM behavior more interpretable by providing human-understandable justifications for model predictions, such as rationales (Mishra et al., 2024), contrastive explanations (Buçinca et al., 2025; Si et al., 2024), multilevel and contextualized explanations (Monteiro Paes et al., 2025; Mei et al., 2023; Di Bonaventura et al., 2024), anchored in situ explanations (Yan et al., 2024), explanations with different confidence levels (Wang et al., 2025), saliency explanations (Pafla et al., 2024) or visualization of internal states (Spinner et al., 2024), at various stages of interaction (Kim et al., 2025; Yao et al., 2023) to help users understand how and why a model generated a particular output.

The primary goal is to empower users to interpret, question, and calibrate trust in LLM outputs by integrating user-relevant explanations into the human-LLM interaction. Rather than being a post-hoc feature, explanations are regarded as an integral part of the user experience.

However, several design limitations remain underexplored. First, explanation quality is uneven, and users are often asked to trust model-generated justifications without support for interrogating the explanation itself. For instance, saliency maps or ranked rationales assume that the model's attention aligns with human reasoning, but users are not empowered to put this alignment into question. Most designs present a single explanation type, limiting opportunities for comparison (Pafla et al., 2024).

Second, explanation interfaces often rely on static visualizations or textual input. While a few designs allow users to manipulate explanations (e.g., editable search trees or contrastive comparisons), these remain exceptions. Moreover, explanations are usually presented as final, and users can not contribute to the model's reasoning. This risks reinforcing overreliance on explanation rather than promoting interactivity and critical engagement.

Third, the cognitive demands of interpreting explanations are often overlooked. Visualizations, importance heatmaps, or rationales may be challenging to interpret for non-experts or minoritized groups, and some studies suggest that users make better decisions with external references (e.g., Wikipedia) than with model-generated explanations (Si et al., 2024). The assumption that explanations automatically enhance trust or understanding must be validated across diverse user groups and domains.

Finally, most explanation-centered designs explain one output at a time (for example, why the model gave a specific answer), but they usually don't help users understand a general model behavior, such as whether the model is biased, how it was trained, or what kinds of mistakes it tends to make overall. An exception is Yao et al. (2023) where human-annotated explanations are integrated into active learning loops for annotation support, involving users in both training and evaluation phases.

In sum, while explanation-centered interfaces enhance transparency, they risk oversimplifying the complexity of LLM behavior and limiting user agency if not designed with deeper interactivity, explanation pluralism, and user education in mind.

Style-based trust calibration Style-based trust calibration refers to design strategies that shape users' trust in LLM outputs by varying the communicative style of the output. Rather than changing the factual content, these approaches manipulate how information is conveyed, for example, by presenting the output in an assertive or a hesitant tone, or showing confidence cues and visually marking lexical indicators of uncertainty to help users form more accurate mental models of LLM reliability. The central assumption is that stylistic framing and contextual cues strongly influence user reliance, perceived transparency, and decision confidence.

The primary goal is to support better alignment between perceived and actual model capabilities. This is especially crucial in settings involving uncertainty or risk, such as healthcare, legal advice, or career guidance.

Rather than improving accuracy directly, these interventions calibrate user perception of models. Studies have tested expressions of uncertainty (e.g., first-person: "I'm not sure..." vs. impersonal: "It is not sure...") (Kim et al., 2024b), confidence disclaimers (Metzger et al., 2024), comparing hesitant versus assertive tones (Kadoma et al., 2024), trust repair techniques through apologies, denials, and promises (Pareek et al., 2024), stylistic variations across chatbot types (LLM-based vs. intent-based vs. form-based) (Zylowski et al., 2025), uncertainty markers (Zhou et al., 2024; Chen et al., 2025b), model-generated greetings (Zhou et al., 2025b), and visual disclaimers or highlights (Bo et al., 2025). These features are tested in calibrated and miscalibrated scenarios to assess their influence on user trust.

However, most work remains narrowly focused on whether a given stylistic manipulation influences trust, rather than how users can be supported in recognizing and critically engaging with such cues in everyday use. For instance, while many features are shown to affect trust in experimental setups, they are rarely integrated into interface systems with guidance or educational scaffolding. Ma et al. (2025) address this by proposing a deliberation-based interface that encourages users to reason through LLM suggestions. Yet, dealing with insufficient analytical engagement of users with AI recommendations remains an exception.

A further limitation is the tension between helping users calibrate trust and the risk of unintentionally manipulating them or introducing new ethical problems. Specifically, stylistic cues may encode cultural or gender biases, reinforce stereotypes, or mask unreliable model behavior behind persuasive style. Future work should examine how style interacts with power, and whether certain user groups are more vulnerable to over-reliance due to stylistic calibration alone.

In conclusion, while style-based approaches offer promising mechanisms for aligning user-perceived trust with actual model reliability, they raise critical open questions about fairness and the long-term effects of such calibration.

Privacy-aware architectures and tools Privacy-aware architectures and tools are systems, interfaces, or frameworks that aim to detect, minimize, or prevent privacy risks in human–LLM interaction. They enhance user awareness, control, and protection by implementing privacy safeguards either before, during, or after data exchange with LLMs. These approaches consider input redaction, output inspection, system-level manipulation detection, and user education, often grounded in user-centered design and participatory development. Unlike general security methods, this category focuses on end-user-facing privacy measures, enabling users to actively participate in managing their personal data exposure and autonomy in LLM-mediated environments.

The primary goal is to empower users to manage and protect their personal data by providing controllable tools that mitigate privacy risks at every stage of the interaction pipeline. These systems aim to increase user agency and awareness while reducing unintended data leakage, over-disclosure, or manipulation in AI-mediated communication. They address not only what LLMs can "know" or "leak", but how users can actively participate in preventing harm and making informed choices about data use, visibility, and trustworthiness.

Core strategies in this paradigm span the full privacy lifecycle, from input-level privacy control (Ngong et al., 2025) through self-disclosure detection (Dou et al., 2024), user-led data minimization via browser extensions (Zhou et al., 2025a) to post-hoc inspection (e.g., leaking personal identifiers through LLM outputs Kim et al. 2024a or detecting prompt injection attacks Lin et al. 2025) and user education (Chen et al., 2025a).

However, these designs may face adoption challenges. Many tools assume that users are both willing and able to engage in privacy management, although users may sometimes prioritize convenience or utility over caution, especially in low-stakes contexts. Moreover, privacy-aware interfaces can disrupt the user experience if they demand too much time, technical understanding, or attention. To be effective, they must be carefully adapted to the context of use and the user's mental workload, for example by being paired with automation, personalization, or persuasive design. Finally, some designs risk offloading the responsibility for privacy onto the user without addressing underlying system-level weaknesses in how LLMs handle user data. For example, asking users to identify sensitive content assumes they understand what counts as risky in the context of opaque model behavior, but this assumption may not hold. It is also unclear how such tools perform across user groups with varying levels of sensitivity to privacy issues.

In sum, more research is needed to assess how to communicate privacy risks without overwhelming users or discouraging them from critical use of LLMs. Privacy-aware tools play a crucial role in shifting privacy control closer to users, but must be designed to balance protection, usability, and psychological trust across varied real-world scenarios.

4 Theoretical perspectives

To synthesize the design strategies identified through inductive coding, we draw on three complementary frameworks from HCI and cognitive science: Activity Theory (Kuutti, 1996), Distributed Cognition (Hollan et al., 2000), and Mental Models (see an overview in Payne, 2003). These descriptive theories are suited for analyzing user-centred paradigms across NLP and HCI research.

Activity theory highlights how users engage with LLMs as tools to achieve specific goals (e.g., writing, learning). It aligns closely with interactive authoring & co-creation and workflow-aligned designs where LLMs support domain-specific tasks (e.g., Masson et al. 2024; Kazemitabaar et al. 2024), enabling users to shift from passive prompting to active participation. Participatory designs also empower users by emphasizing their agency in shaping system behavior (e.g., Theophilou et al. 2023).

Distributed cognition frames trust as emerging from the interaction between the user, the LLM system, and the interventions (e.g., visualizations, warnings), such as in interface-level accuracy control (e.g., Leiser et al. 2024) and style-based trust calibration (e.g., Zhou et al. 2024). Trust calibration is distributed across the model's suggestions, system-generated evidence, and design interventions rather than by internal understanding alone.

Referring to users' internal understandings of how LLMs work, mental models are central to explanation-centered approaches (e.g., Yan et al. 2024) that aim to scaffold reasoning about model logic, privacy-aware designs (e.g., Dou et al. 2024) that help users understand what LLMs might infer from personal data, and style-based trust calibration, which influences users' conceptual models of LLM reliability.

Additionally, our classification aligns with the more recent human-centered AI (HCAI) framework proposed by Shneiderman (2022), particularly in treating user control not only as an outcome (product) but also as a participatory design process.

The proposed design paradigms also align with principles from classical HCI, such as Norman's gulfs of execution and evaluation (see Norman, 2013, 38–40), which describe the barriers users face in acting on and interpreting system behavior. Several designs aim to reduce Norman's gulf of execution by simplifying prompt design (Zhang and Arawjo, 2025) or providing scaffolds that guide users in expressing their intentions. Others address the gulf of evaluation by offering visualizations (Spinner et al., 2024) of model decisions or contrastive explanations (Buçinca et al., 2025) to help users interpret outputs. Furthermore, activity theory helps reduce the gulf of execution by analyzing whether users can meaningfully act on interfaces to achieve their goals. Distributed cognition addresses the gulf of evaluation by highlighting how trust and understanding are mediated through interface-level cues, external visualizations, and interaction history. Finally, mental models support both gulfs by determining how users understand what actions are possible and how outputs should be interpreted. Together, these theories provide a layered perspective on user control in LLM interactions.

5 Discussion and Conclusions

This paper identified and systematized seven design paradigms that promote user control in human–LLM interaction and reflect design strategies grounded in different user goals, ranging from verifying factuality and shaping model output to managing trust and data exposure. Our design-centered perspective complements current discussions on human involvement in post-training by emphasizing user control during deployment and interaction.

While empirical studies have offered scattered examples of user-centered designs and most recent related surveys do not primarily focus on trust or have a broader scope (e.g., human-model cooperation in Huang et al. 2025), our contribution lies in synthesizing these efforts into a coherent framework that centers user goals as the organizing principle of human trust in LLMs. Across paradigms, we observe a shift from one-shot prompting toward interactive, iterative, and increasingly user-configurable LLM workflows. These designs foreground a broad spectrum of control types: perceptual (e.g., accuracy cues), procedural (e.g., workflow pausing), epistemic (e.g., explanations, varying linguistic style), and protective (e.g., privacy screening).

Yet, critical gaps remain. Although many studies mention cross-domain application (e.g. Louie et al. 2024), the variety of tested scenarios is limited. We also observe a lack of design frameworks that help practitioners balance automation and human agency. For example, many tools mediate control through additional LLMs (e.g., Pan et al. 2024), which risks reinforcing automation bias rather than supporting user autonomy. To address this, future systems could incorporate trust calibration strategies (e.g., communicative framing, interactive uncertainty visualization) that help users reflect on when and how to trust outputs. Most studies assume AI-literate end users with a high level of technical literacy. Designs rarely account for diverse user needs, e.g., those with low reading/writing literacy, limited technical expertise, or from marginalized communities. This limits the accessibility and generalizability of proposed methods. Users are often expected to interpret complex cues (e.g., factuality scores) without training. It remains unclear how to prevent over-reliance on automation while avoiding user frustration and how to balance control vs. usability, or privacy vs. personalization.

For the research at the intersection of HCI and NLP, we identify several promising directions for future work:

- Explicitly address interaction design patterns that foster meaningful user oversight (e.g., modular prompt chaining, co-creation loops).
- Expand design efforts to underexplored trust dimensions (e.g., fairness, social well-being).
- Develop participatory methods that involve diverse users in the co-design of trust-aware LLM interfaces.
- Develop systems that calibrate trust in LLMs not only by LLMs but also include human-in-the-loop review.
- To support low-literacy users, consider, for example, visual metaphors to reduce cognitive burden or interaction logging, or user interfaces with a toggle to simplify responses.
- Replace binary on/off controls with graded or layered control (e.g., co-authoring steps or adjustable initiative).
- Move beyond controlled studies to assess how trust and control evolve during prolonged, real-world interaction (in-the-wild evaluation)
- Consider long-term, real-world deployment studies to assess how interaction designs shape trust over time.

Finally, we advocate for design that enables not just enhanced control, but critical engagement with LLM behavior, especially through scaffolds that support users in questioning and modifying model output.

6 Limitations

We identify three main limitations of this study. First, as this is a design-centered survey rather than a systematic meta-analysis, two types of constraints apply: those related to paper selection and those associated with the derivation of the proposed design typology. The final scope of included papers was based on a qualitative assessment by two annotators, followed by iterative discussion to reach consensus on inclusion. Consequently, not all papers containing search terms in the abstract, title, or keywords were included. Both the paper selection and the resulting classification are thus shaped by human judgment and interpretability. In particular, some papers at the boundary between metric-driven evaluation and user-centered design were included if they contained at least partial user evaluation components, such as in Koraş et al. (2025), where the user study was exploratory and not systematic. Although many papers could plausibly be assigned to multiple paradigms, annotators were instructed to assign each paper to a single primary category. The proposed design paradigms were qualitatively derived and require further empirical validation.

Second, due to limitations in the ACL Anthology search interface (see Section 2), it was not possible to apply an identical search string across both databases. While the ACM Digital Library search allowed for complex Boolean queries, the ACL Anthology search relied on simpler keyword combinations (see Section 2). This discrepancy may have introduced a bias by potentially missing relevant ACL papers that would have matched the full ACM query. A brief comparative test or validation of coverage was not feasible, but we acknowledge that this search asymmetry could affect the completeness and balance of the corpus. Furthermore, the review does not include papers from other sources such as arXiv, which means that unpublished or in-progress work was not considered.

Third, the reviewed studies are predominantly situated in English-speaking and Western contexts, as only papers published in English were included. This limits the cultural and linguistic diversity of the findings.

7 Ethical statement

This work is a meta-analysis of published research at the intersection of HCI and NLP. We do not present or process personal data, nor do we involve human participants. All surveyed papers were selected from publicly accessible, peer-reviewed sources, excluding preprints. Where user studies are reported in the cited literature, we rely on the original authors' ethical approvals and disclosures. Care was taken to fairly represent a diverse set of approaches and to avoid overgeneralizing results.

We acknowledge that relying solely on published, English-language sources may introduce publication and cultural bias, leading to an overrepresentation of Western perspectives. This is not only a methodological limitation (see Section 6), but also an ethical concern for the generalizability and inclusivity of our findings.

References

Mamia Agbese, Rahul Mohanani, Arif Khan, and Pekka Abrahamsson. 2023. Implementing AI ethics: Making sense of the ethical requirements. In *Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering*, EASE '23, page 62–71, New York, NY, USA. Association for Computing Machinery.

Syeda Sabrina Akter, Seth Hunter, David Woo, and Antonios Anastasopoulos. 2025. Costs and benefits of AI-enabled topic modeling in P-20 research: The case of school improvement plans. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 460–476, Vienna, Austria. Association for Computational Linguistics.

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 *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, Chi '24, New York, NY, USA. ACM.

Jessica Y Bo, Sophia Wan, and Ashton Anderson. 2025. To rely or not to rely? Evaluating interventions for appropriate reliance on large language models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Zana Buçinca, Siddharth Swaroop, Amanda E. Paluch, Finale Doshi-Velez, and Krzysztof Z. Gajos. 2025. Contrastive explanations that anticipate human misconceptions can improve human decision-making skills. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Chaoran Chen, Daodao Zhou, Yanfang Ye, Toby Jia-Jun Li, and Yaxing Yao. 2025a. Clear: Towards contextual LLM-empowered privacy policy analysis and risk generation for large language model applications. In *Proceedings of the 30th International*

Conference on Intelligent User Interfaces, IUI '25, page 277–297, New York, NY, USA. Association for Computing Machinery.

Cheng Chen, Sangwook Lee, Eunchae Jang, and S. Shyam Sundar. 2024. Is your prompt detailed enough? exploring the effects of prompt coaching on users' perceptions, engagement, and trust in text-to-image generative AI tools. In *Proceedings of the Second International Symposium on Trustworthy Autonomous Systems*, pages 1–12, Austin TX USA. ACM.

Rex Chen, Ruiyi Wang, Norman Sadeh, and Fei Fang. 2025b. Missing pieces: How do designs that expose uncertainty longitudinally impact trust in AI decision aids? An in situ study of gig drivers. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, page 790–816, New York, NY, USA. Association for Computing Machinery.

Furui Cheng, Vilém Zouhar, Simran Arora, Mrinmaya Sachan, Hendrik Strobelt, and Mennatallah El-Assady. 2024. Relic: Investigating large language model responses using self-consistency. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–18, Honolulu HI USA. ACM.

Alexander S. Choi, Syeda Sabrina Akter, JP Singh, and Antonios Anastasopoulos. 2024. The LLM effect: Are humans truly using LLMs, or are they being influenced by them instead? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22032–22054, Miami, Florida, USA. Association for Computational Linguistics.

Jiwon Chun, Yankun Zhao, Hanlin Chen, and Meng Xia. 2025. Planglow: Personalized study planning with an explainable and controllable LLM-driven system. In *Proceedings of the Twelfth ACM Conference on Learning @ Scale*, L@S '25, page 116–127, New York, NY, USA. Association for Computing Machinery.

Shih-Chieh Dai, Aiping Xiong, and Lun-Wei Ku. 2023. LLM-in-the-loop: Leveraging large language model for thematic analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9993–10001, Singapore. Association for Computational Linguistics.

Alicia DeVos, Aditi Dhabalia, Hong Shen, Kenneth Holstein, and Motahhare Eslami. 2022. Toward user-driven algorithm auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Chiara Di Bonaventura, Lucia Siciliani, Pierpaolo Basile, Albert Merono Penuela, and Barbara Mcgillivray. 2024. Is explanation all you need? an expert survey on LLM-generated explanations for abusive language detection. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 280–288, Pisa, Italy. CEUR Workshop Proceedings.

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 *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3321–3339, Singapore. Association for Computational Linguistics.

Jinhao Dong, Jun Sun, Wenjie Zhang, Jin Song Dong, and Dan Hao. 2025. Contested: Consistency-aided tested code generation with LLM. *Proc. ACM Softw. Eng.*, 2(ISSTA).

Yao Dou, Isadora Krsek, Tarek Naous, Anubha Kabra, Sauvik Das, Alan Ritter, and Wei Xu. 2024. Reducing privacy risks in online self-disclosures with language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13732–13754, Bangkok, Thailand. Association for Computational Linguistics.

Maximilian Förster, Mathias Klier, Kilian Kluge, and Irina Sigler. 2020. *Fostering human agency: A process for the design of user-centric XAI systems.*

Abhinav Gupta, Devendra Singh, Greig A Cowan, N Kadhiresan, Siddharth Srivastava, Yagneswaran Sriraja, and Yoages Kumar Mantri. 2025. AUTOSUMM: A comprehensive framework for LLM-based conversation summarization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 500–509, Vienna, Austria. Association for Computational Linguistics.

Neal R Haddaway, Matthew J Page, Chris C Pritchard, and Luke A McGuinness. 2022. PRISMA2020: An R package and shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and open synthesis. *Campbell Syst. Rev.*, 18(2):e1230.

Thilo Hagendorff. 2020. The ethics of AI ethics: An evaluation of guidelines. *Minds Mach.*, 30(1):99–120.

Gaole He, Gianluca Demartini, and Ujwal Gadiraju. 2025. *Plan-then-execute: An empirical study of user trust and team performance when using LLM agents as a daily assistant.* Association for Computing Machinery, New York, NY, USA.

HLEG. 2019. Ethics guidelines for trustworthy AI. Expert-group report, European Commission, Directorate-General for Communications Networks, Content and Technology, Brussels.

James Hollan, Edwin Hutchins, and David Kirsh. 2000. Distributed cognition: toward a new foundation for human-computer interaction research. *ACM Trans. Comput.-Hum. Interact.*, 7(2):174–196.

Md Naimul Hoque, Tasfia Mashiat, Bhavya Ghai, Cecilia D. Shelton, Fanny Chevalier, Kari Kraus, and Niklas Elmqvist. 2024. The HaLLMark effect: Supporting provenance and transparent use of large language models in writing with interactive visualization. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, Honolulu HI USA. ACM.

Yutong Hu, Kangcheng Luo, and Yansong Feng. 2024. ELLA: Empowering LLMs for interpretable, accurate and informative legal advice. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 374–387, Bangkok, Thailand. Association for Computational Linguistics.

Chen Huang, Yang Deng, Wenqiang Lei, Jiancheng Lv, Tat-Seng Chua, and Jimmy Huang. 2025. How to enable effective cooperation between humans and NLP models: A survey of principles, formalizations, and beyond. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 466–488, Vienna, Austria. Association for Computational Linguistics.

Chia-Chien Hung, Wiem Ben Rim, Lindsay Frost, Lars Bruckner, and Carolin Lawrence. 2023. Walking a tightrope – evaluating large language models in high-risk domains. In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, pages 99–111, Singapore. Association for Computational Linguistics.

Kowe Kadoma, Marianne Aubin Le Quere, Xiyu Jenny Fu, Christin Munsch, Danaë Metaxa, and Mor Naaman. 2024. The role of inclusion, control, and ownership in workplace AI-mediated communication. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, Chi '24, New York, NY, USA. ACM.

Davinder Kaur, Suleyman Uslu, Kaley J. Rittichier, and Arjan Durresi. 2022. Trustworthy artificial intelligence: A review. *ACM Comput. Surv.*, 55(2).

Majeed Kazemitabaar, Runlong Ye, Xiaoning Wang, Austin Zachary Henley, Paul Denny, Michelle Craig, and Tovi Grossman. 2024. CodeAid: Evaluating a classroom deployment of an LLM-based programming assistant that balances student and educator needs. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–20, Honolulu HI USA. ACM.

Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. 2024a. ProPILE: Probing privacy leakage in large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Nips '23, Red Hook, NY, USA. Curran Associates Inc.

Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024b. "I'm not sure, but...": Examining the impact of large language models' uncertainty expression on user reliance and trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 822–835, New York, NY, USA. Association for Computing Machinery.

Sunnie S. Y. Kim, Jennifer Wortman Vaughan, Q. Vera Liao, Tania Lombrozo, and Olga Russakovsky. 2025. Fostering appropriate reliance on large language models: The role of explanations, sources, and inconsistencies. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Osman Alperen Koraş, Rabi Bahnan, Jens Kleesiek, and Amin Dada. 2025. Towards conditioning clinical text generation for user control. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10549–10569, Vienna, Austria. Association for Computational Linguistics.

Kari Kuutti. 1996. *Activity theory as a potential framework for human-computer interaction research*, chapter 2. MIT press.

Philippe Laban, Jesse Vig, Marti Hearst, Caiming Xiong, and Chien-Sheng Wu. 2024. Beyond the chat: Executable and verifiable text-editing with LLMs. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, pages 1–23, Pittsburgh PA USA. ACM.

Christine P. Lee, David Porfirio, Xinyu Jessica Wang, Kevin Chenkai Zhao, and Bilge Mutlu. 2025. VeriPlan: Integrating formal verification and LLMs into end-user planning. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Florian Leiser, Sven Eckhardt, Valentin Leuthe, Merlin Knaeble, Alexander Mädche, Gerhard Schwabe, and Ali Sunyaev. 2024. HILL: A hallucination identifier for large language models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–13, Honolulu HI USA. ACM.

Mark Liffiton, Brad E Sheese, Jaromir Savelka, and Paul Denny. 2023. CodeHelp: Using large language models with guardrails for scalable support in programming classes. In *Proceedings of the 23rd Koli Calling International Conference on Computing Education Research*, pages 1–11, Koli Finland. ACM.

Weiran Lin, Anna Gerchanovsky, Omer Akgul, Lujo Bauer, Matt Fredrikson, and Zifan Wang. 2025. LLM whisperer: An inconspicuous attack to bias LLM responses. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Jiahao Liu, Yiyang Shao, Peng Zhang, Dongsheng Li, Hansu Gu, Chao Chen, Longzhi Du, Tun Lu, and Ning Gu. 2025. Filtering discomforting recommendations with large language models. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 3639–3650, New York, NY, USA. Association for Computing Machinery.

Yiren Liu, Si Chen, Haocong Cheng, Mengxia Yu, Xiao Ran, Andrew Mo, Yiliu Tang, and Yun Huang. 2024. How AI processing delays foster creativity: Exploring research question co-creation with an LLM-based agent. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, Chi '24, New York, NY, USA. ACM.

Ryan Louie, Ananjan Nandi, William Fang, Cheng Chang, Emma Brunskill, and Diyi Yang. 2024. Roleplay-doh: Enabling domain-experts to create LLM-simulated patients via eliciting and adhering to principles. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10570–10603, Miami, Florida, USA. Association for Computational Linguistics.

Shuai Ma, Qiaoyi Chen, Xinru Wang, Chengbo Zheng, Zhenhui Peng, Ming Yin, and Xiaojuan Ma. 2025. Towards human-AI deliberation: Design and evaluation of LLM-empowered deliberative AI for AI-assisted decision-making. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Damien Masson, Sylvain Malacria, Géry Casiez, and Daniel Vogel. 2024. DirectGPT: A direct manipulation interface to interact with large language models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, Chi '24, New York, NY, USA. ACM.

Alex Mei, Sharon Levy, and William Yang Wang. 2023. Foveate, attribute, and rationalize: Towards physically safe and trustworthy AI. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11021–11036, Toronto, Canada. Association for Computational Linguistics.

Luise Metzger, Linda Miller, Martin Baumann, and Johannes Kraus. 2024. Empowering calibrated (dis-)trust in conversational agents: A user study on the persuasive power of limitation disclaimers vs. authoritative style. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–19, Honolulu HI USA. ACM.

Qijun Miao and Zhixuan Fang. 2025. User-side model consistency monitoring for open source large language models inference services. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11610–11622, Vienna, Austria. Association for Computational Linguistics.

Aditi Mishra, Sajjadur Rahman, Kushan Mitra, Hannah Kim, and Estevam Hruschka. 2024. Characterizing large language models as rationalizers of knowledge-intensive tasks. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8117–8139, Bangkok, Thailand. Association for Computational Linguistics.

Lucas Monteiro Paes, Dennis Wei, Hyo Jin Do, Hendrik Strobelt, Ronny Luss, Amit Dhurandhar, Manish Nagireddy, Karthikeyan Natesan Ramamurthy, Prasanna Sattigeri, Werner Geyer, and Soumya Ghosh. 2025. Multi-level explanations for generative language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32291–32317, Vienna, Austria. Association for Computational Linguistics.

Ivoline C. Ngong, Swanand Ravindra Kadhe, Hao Wang, Keerthiram Murugesan, Justin D. Weisz, Amit Dhurandhar, and Karthikeyan Natesan Ramamurthy. 2025. Protecting users from themselves: Safeguarding contextual privacy in interactions with conversational agents. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26196–26220, Vienna, Austria. Association for Computational Linguistics.

Vincent Nguyen, Sarvnaz Karimi, Willow Hallgren, Ashley Harkin, and Mahesh Prakash. 2024. My climate advisor: An application of NLP in climate adaptation for agriculture. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 27–45, Bangkok, Thailand. Association for Computational Linguistics.

Don Norman. 2013. *The design of everyday things*, 2 edition. Basic Books, London, England.

Cassandra Overney, Daniel T Kessler, Suyash Pradeep Fulay, Mahmood Jasim, and Deb Roy. 2025. Coalesce: An accessible mixed-initiative system for designing community-centric questionnaires. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, IUI '25, page 366–389, New York, NY, USA. Association for Computing Machinery.

Marvin Pafla, Kate Larson, and Mark Hancock. 2024. Unraveling the dilemma of ai errors: Exploring the effectiveness of human and machine explanations for large language models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, and 7 others. 2021. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Syst. Rev.*, 10(1):89.

Qian Pan, Zahra Ashktorab, Michael Desmond, Martín Santillán Cooper, James Johnson, Rahul Nair, Elizabeth Daly, and Werner Geyer. 2024. Human-centered design recommendations for LLM-as-a-judge. In *Proceedings of the 1st Human-Centered Large Language Modeling Workshop*, pages 16–29, TBD. ACL.

Saumya Pareek, Eduardo Velloso, and Jorge Goncalves. 2024. Trust development and repair in AI-assisted decision-making during complementary expertise. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 546–561, New York, NY, USA. Association for Computing Machinery.

Stephen J Payne. 2003. Users' mental models: The very ideas. In *HCI Models, Theories, and Frameworks*, pages 135–156. Elsevier.

Vadim Perov and Vladislav Golovkov. 2024. Ethics documents in the field of AI. Concepts, achievements and problems. In *2024 IEEE Ural-Siberian Conference on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT)*, volume 7, pages 196–199, Yekaterinburg. Eee.

Snehal Prabhudesai, Ananya Prashant Kasi, Anmol Mansingh, Anindya Das Antar, Hua Shen, and Nikola Banovic. 2025. "Here the GPT made a choice, and every choice can be biased": How students critically engage with LLMs through end-user auditing activity. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Abhinav Sukumar Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2025. NormAd: A framework for measuring the cultural adaptability of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2373–2403, Albuquerque, New Mexico. Association for Computational Linguistics.

Charvi Rastogi, Marco Tulio Ribeiro, Nicholas King, Harsha Nori, and Saleema Amershi. 2023. Supporting human-AI collaboration in auditing LLMs with LLMs. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, Aies '23, page 913–926, New York, NY, USA. ACM.

Mohi Reza, Ioannis Anastasopoulos, Shreya Bhandari, and Zachary A. Pardos. 2025. PromptHive: Bringing subject matter experts back to the forefront with collaborative prompt engineering for educational content creation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Ben Shneiderman. 2022. *Human-centered AI*. Oxford University Press, London, England.

Chenglei Si, Navita Goyal, Tongshuang Wu, Chen Zhao, Shi Feng, Hal Daumé Iii, and Jordan Boyd-Graber. 2024. Large language models help humans verify truthfulness – except when they are convincingly wrong. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1459–1474, Mexico City, Mexico. Association for Computational Linguistics.

Thilo Spinner, Rebecca Kehlbeck, Rita Sevastjanova, Tobias Stähle, Daniel A. Keim, Oliver Deussen, and Mennatallah El-Assady. 2024. -generAItor: Tree-in-the-loop text generation for language model explainability and adaptation. *ACM Trans. Interact. Intell. Syst.*, 14(2).

Yi Tang, Chia-Ming Chang, and Xi Yang. 2024. PDFchatannotator: A human-LLM collaborative multi-modal data annotation tool for PDF-format catalogs. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, Iui '24, page 419–430, New York, NY, USA. ACM.

Emily Theophilou, Cansu Koyutürk, Mona Yavari, Sathya Bursic, Gregor Donabauer, Alessia Telari, Alessia Testa, Raffaele Boiano, Davinia Hernandez-Leo, Martin Ruskov, Davide Taibi, Alessandro Gabbiadini, and Dimitri Ognibene. 2023. Learning to prompt in the classroom to understand AI limits: A pilot study. In *AIxIA 2023 – Advances in Artificial Intelligence: XXIInd International Conference of the Italian Association for Artificial Intelligence, AIxIA 2023, Rome, Italy, November 6–9, 2023, Proceedings*, page 481–496, Berlin, Heidelberg. Springer-Verlag.

Usman Ahmad Usmani, Ari Happonen, and Junzo Watada. 2023. Human-centered artificial intelligence: Designing for user empowerment and ethical considerations. In *2023 5th international congress on human-computer interaction, optimization and robotic applications (HORA)*, pages 1–7, Istanbul. Ieee.

Xinru Wang, Mengjie Yu, Hannah Nguyen, Michael Iuzzolino, Tianyi Wang, Peiqi Tang, Natasha Lynova, Co Tran, Ting Zhang, Naveen Sendhilnathan, Hrvoje Benko, Haijun Xia, and Tanya R. Jonker. 2025. Less or more: Towards glanceable explanations for LLM recommendations using ultra-small devices. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, IUI '25, page 938–951, New York, NY, USA. Association for Computing Machinery.

Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. AI chains: Transparent and controllable human-AI interaction by chaining large language model prompts. In *CHI Conference on Human Factors in Computing Systems*, pages 1–22, New Orleans LA USA. ACM.

Litao Yan, Alyssa Hwang, Zhiyuan Wu, and Andrew Head. 2024. Ivie: Lightweight anchored explanations of just-generated code. In *Proceedings of the*

2024 CHI Conference on Human Factors in Computing Systems, Chi '24, New York, NY, USA. ACM.

Bingsheng Yao, Ishan Jindal, Lucian Popa, Yannis Katsis, Sayan Ghosh, Lihong He, Yuxuan Lu, Shashank Srivastava, Yunyao Li, James Hendler, and Dakuo Wang. 2023. Beyond labels: Empowering human annotators with natural language explanations through a novel active-learning architecture. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11629–11643, Singapore. Association for Computational Linguistics.

Jingyue Zhang and Ian Arawjo. 2025. Chainbuddy: An AI-assisted agent system for generating LLM pipelines. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Yuechen Zhang, Shengju Qian, Bohao Peng, Shu Liu, and Jiaya Jia. 2024. Prompt highlighter: Interactive control for multi-modal LLMs.

Jingyao Zheng, Xian Wang, Simo Hosio, Xiaoxian Xu, and Lik-Hang Lee. 2025a. LMLPA: Language model linguistic personality assessment. *Computational Linguistics*, 51:599–640.

Xi Zheng, Zhuoyang Li, Xinning Gui, and Yuhan Luo. 2025b. Customizing emotional support: How do individuals construct and interact with LLM-powered chatbots. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Jijie Zhou, Eryue Xu, Yaoyao Wu, and Tianshi Li. 2025a. Rescriber: Smaller-LLM-powered user-led data minimization for LLM-based chatbots. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Kaitlyn Zhou, Jena Hwang, Xiang Ren, and Maarten Sap. 2024. Relying on the unreliable: The impact of language models' reluctance to express uncertainty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3623–3643, Bangkok, Thailand. Association for Computational Linguistics.

Kaitlyn Zhou, Jena D. Hwang, Xiang Ren, Nouha Dziri, Dan Jurafsky, and Maarten Sap. 2025b. REL-A.I.: An interaction-centered approach to measuring human-LM reliance. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11148–11167, Albuquerque, New Mexico. Association for Computational Linguistics.

Tiffany Zhu, Iain Weissburg, Kexun Zhang, and William Yang Wang. 2025. Human bias in the face of AI: Examining human judgment against text labeled as AI generated. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25907–25914, Vienna, Austria. Association for Computational Linguistics.

Aneta Zugecova, Dominik Macko, Ivan Srba, Robert Moro, Jakub Kopál, Katarína Marcinčinová, and Matúš Mesarčík. 2025. Evaluation of LLM vulnerabilities to being misused for personalized disinformation generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 780–797, Vienna, Austria. Association for Computational Linguistics.

Thorsten Zylowski, Nathalia Sautchuk-Patricio, Wladimir Hettmann, Katharina Anderer, Karl Fischer, Matthias Wölfel, and Peter Henning. 2025. User study on the trustworthiness, usability and explainability of intent-based and large language model-based career planning conversational agents. In *Proceedings of the 2024 16th International Conference on Education Technology and Computers*, ICETC '24, page 46–53, New York, NY, USA. Association for Computing Machinery.

A Appendix

Interface-level accuracy control. User-led verification based on consistency of LLM responses (Cheng et al., 2024); user control of LLM edits (Laban et al., 2024); user study of LLM-based planning systems (Lee et al., 2025); user-centered development of hallucination identifier for LLMs (Leiser et al., 2024). *Primary trust aspect:* Accuracy, transparency.

Workflow-aligned and domain-adapted AI assistance. LLM-assisted topic modeling for qualitative analysis (Akter et al., 2025; Choi et al., 2024); human–LLM collaboration for thematic analysis (Dai et al., 2023); LLM code generation with user feedback (Dong et al., 2025); human-in-the-loop conversation summarization for financial advisors (Gupta et al., 2025); plan-then-execute LLM collaboration with user-in-the-loop control (He et al., 2025); LLM-based legal advice with user intervention (Hu et al., 2024); human evaluation of LLM programming assistant (Kazemitabaar et al., 2024); human evaluation of LLM-generated texts in clinical settings (Koraş et al., 2025); user evaluation of LLM-based code assistance with guardrails (Liffiton et al., 2023). *Primary trust aspect:* Accuracy, transparency, oversight; auditability.

Participatory designs. Co-designed self-directed learning planner (Chun et al., 2025); co-design of roleplay prompts with domain experts (Louie et al., 2024); climate advice via co-designed LLM interaction (Nguyen et al., 2024); user involvement in the LLM-as-a-judge concept (Pan et al., 2024); comparing human and LLM judgements of cultural adaptability (Rao et al., 2025); collaborative prompt authoring interface for homework problems (Reza et al., 2025); AI literacy education (Theophilou et al., 2023); co-creation of chatbot personas for emotional reliance (Zheng et al., 2025b); user preference of texts with different labels (LLM-generated vs. human) (Zhu et al., 2025). *Primary trust aspect:* Reliability, fairness, bias.

Interactive authoring & co-creation. Interactive prompt engineering and evaluation (Arawjo et al., 2024); human–AI co-creation of news headlines (Ding et al., 2023); provenance-driven co-writing (Hoque et al., 2024); human–LLM co-creation of research questions (Liu et al., 2024); user-aligned co-filtering of discomforting recommendations (Liu et al., 2025); direct manipulation interface (Masson et al., 2024); human–LLM co-creation of questionnaires (Overney et al., 2025); end-user auditing scaffolds for identifying LLM biases (Prabhudesai et al., 2025); LLM-based human–AI auditing (Rastogi et al., 2023); human–LLM modular prompt chaining (Wu et al., 2022); LLM-based human–AI evaluation of LLM behavior (Zhang and Arawjo, 2025); LLM-assisted user evaluation of LLM personalities (Zheng et al., 2025a); human evaluation of LLM-generated personalized disinformation (Zugecova et al., 2025). *Primary trust aspect:* Reliability, transparency.

Explanation-centered approaches. User evaluation of LLM explanations for abusive language detection tasks (Di Bonaventura et al., 2024); user evaluation of contrastive explanations (Buçinca et al., 2025); impact of LLM explanations on user reliance (Kim et al., 2025); user evaluation of safety-related LLM rationales (Mei et al., 2023); evaluation of LLM rationale quality (Mishra et al., 2024); user study with multi-level model explanations (Monteiro Paes et al., 2025); user evaluation of human vs. XAI explanations (Pafla et al., 2024); user evaluation of LLM explanations and search engines (Si et al., 2024); user evaluation of tree-of-thought visualization (Spinner et al., 2024); in-situ anchored code explanations (Yan et al., 2024); human vs. LLM rationales (Yao et al., 2023); spatially structured and temporally adaptive explanations (Wang et al., 2025). *Primary trust aspect:* Explainability, transparency, reliability.

Style-based trust calibration. Reliance interventions (Bo et al., 2025); hesitant vs. self-assured auto-complete LLM suggestions (Kadoma et al., 2024); certain vs. uncertain LLM responses (Kim et al., 2024b); interactive AI–human deliberation (Ma et al., 2025); disclaimers + high vs. low authority style in LLM responses (Metzger et al., 2024); LLM-generated trust repair strategies (Pareek et al., 2024); LLM-generated emphatic expressions of politeness (Zhou et al., 2025b); LLM-generated uncertainty markers (Zhou et al., 2024). *Primary trust aspect:* Transparency, reliability, biases.

Privacy-aware architectures and tools. User-centered self-disclosure abstraction (Dou et al., 2024); threat model for user-centered mitigation of adversarial prompts (Lin et al., 2025); user-led data minimization (Zhou et al., 2025a); privacy-safeguarding intermediary between users and LLMs (Ngong et al., 2025). *Primary trust aspect:* Privacy.