# Beyond Benchmarks: A User-Centric Framework for Evaluating Large Language Models


Yasmeen Abdrabou
Munich Center for Machine Learning (MCML)
Technical University of Munich
Munich, Germany
yasmeen.abdrabou@tum.de

Yomna Abdelrahman
University of the Bundeswehr Munich
Munich, Germany
yomna.eldin@gmail.com

Efe Bozkir
Technical University of Munich
Munich, Germany
efe.bozkir@tum.de

Youssef Mazen
Berliner Hochschule für Technik, Siemens Mobility
Berlin, Germany
youssefkhaled02@gmail.com

Florian Alt
LMU Munich
Munich, Germany
florian.alt@ifi.lmu.de

Enkelejda Kasneci
Munich Center for Machine Learning (MCML)
Technical University of Munich
Munich, Germany
enkelejda.kasneci@tum.de

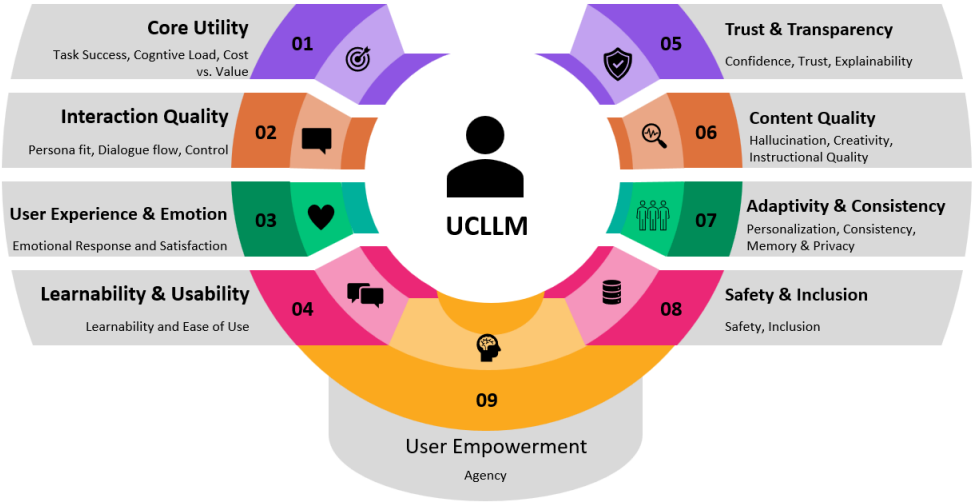




**Figure 1: We introduce UCLLM. A framework to evaluate LLMs from a user-centric perspective. The framework comprises nine user-centric evaluation categories—Core Utility, Interaction Quality, User Experience & Emotion, Learnability & Usability, Trust & Transparency, Content Quality, Adaptivity & Consistency, Safety & Inclusion, and User Empowerment. Each category is represented by an icon to highlight its focus; detailed sub-dimensions and example metrics are provided in Table 1. We also evaluated one aspect of the framework in a preliminary study to compare different LLMs and epistemic stance.**


## Abstract

Large language models (LLMs) are increasingly embedded in everyday life, yet most evaluation efforts still emphasize technical performance rather than how users experience these systems. We present *UCLLM*, a user-centric framework comprising nine dimensions that shift evaluation toward users' perceptions, needs, and interaction goals. Rather than validating the full framework, we demonstrate how one dimension of UCLLM can be operationalized by conducting a pilot blueprint($N = 24$) comparing four widely used LLMs under high- and low-certainty epistemic stances. Low-certainty responses that offered alternatives were often perceived as more confident, reliable, and actionable than high-certainty ones, highlighting a "humility paradox" in user trust. Treating this study as a blueprint illustrates how UCLLM's conceptual dimensions can be translated into measurable user-centered constructs and experimental procedures. Our findings highlight the importance of user-centric evaluation and motivate broader empirical work across



This work is licensed under a Creative Commons Attribution 4.0 International License.


*CUI '26, Bremen, Germany*


© 2026 Copyright held by the owner/author(s).


ACM ISBN 979-8-4007-2741-2/26/07
https://doi.org/10.1145/3816046.3816227

all UCLLM dimensions to guide the design of more trustworthy and responsible LLMs.

## CCS Concepts

• **Human-centered computing → HCI design and evaluation methods**; **Interactive systems and tools**.

## Keywords

Evaluation Framework, User-centric, LLM, Conversational Agents, Human-AI Interaction

**ACM Reference Format:**

Yasmeen Abdrabou, Yomna Abdelrahman, Efe Bozkir, Youssef Mazen, Florian Alt, and Enkelejda Kasneci. 2026. Beyond Benchmarks: A User-Centric Framework for Evaluating Large Language Models. In *ACM Conversational User Interfaces 2026 (CUI '26), July 21–24, 2026, Bremen, Germany*. ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3816046.3816227

## 1 Introduction

Large language models (LLMs) are increasingly embedded in everyday applications [27, 39], supporting tasks across personal assistance [20], education [16, 60], healthcare [13, 54, 76], customer support [70], and creative work [38]. As their use broadens, evaluating their effectiveness becomes critical not only in terms of technical performance but also in terms of how people experience and interpret these interactions in real-world contexts.

Current evaluation approaches focus primarily on system-centric metrics such as computational efficiency [72], memory consumption [28], response latency [58, 63], and task completion accuracy [29, 57]. While necessary, these metrics capture only part of what matters to users. Two models with similar benchmark scores may differ substantially in perceived reliability, trustworthiness, or alignment with social expectations. Without user-centered evaluation, practitioners lack guidance for selecting models suited to specific contexts or for identifying mismatches that can lead to misuse or misinterpretation. Although existing evaluation frameworks cover conversational systems, domain-specific agents, or recommender systems [18, 29, 33, 51, 65, 68], they remain narrow in scope and overlook emergent properties such as hallucination, persona expression, cognitive load, and user agency. A comprehensive structure for evaluating subjective user experience across diverse LLM uses is still missing.

In this work, we address this gap by presenting *UCLLM*, a user-centric framework that organizes LLM evaluation across nine dimensions grounded in HCI constructs and linked to user perspectives, metrics, and methods. UCLLLM synthesizes concepts from prior frameworks such as CoRE [14] and CRS-Que [33] while extending them to accommodate the open-ended, generative, and epistemically uncertain nature of LLM interaction. The nine dimensions were derived through an iterative synthesis of HCI, UX, and conversational evaluation literature (summarized in Section 3.1), guided by the goal of covering distinct yet complementary aspects of user experience rather than replicating existing taxonomies. By articulating Core Utility, Interaction Quality, User Experience & Emotion, Learnability & Usability, Trust & Transparency, Content Quality, Adaptivity & Consistency, Safety & Inclusion, and User Empowerment, UCLLM provides a structured lens for understanding how users engage with LLMs beyond technical performance.

To illustrate how UCLLM can be applied in practice, we conducted a pilot blueprint study ($N = 24$) comparing four widely used LLMs—OpenAI o4-mini[1], Gemini 2.5 Pro[2], Claude Opus 4[3], and DeepSeek V3[4]—under high- and low-certainty epistemic stances. Epistemic stance captures how the model communicates its degree of certainty or uncertainty [12]. Rather than validating the full framework, this pilot demonstrates how one UCLLM dimension, "Trust & Transparency", can be operationalized, serving as a blueprint for evaluating additional dimensions in future work.

Our work is guided by the following research question: **RQ:** How do the underlying LLM and its certainty stance shape users' perceptions of reliability, trust, and interaction quality?

Our findings show that user perceptions vary in ways that purely computational metrics fail to capture. Low-certainty responses were frequently perceived as more confident, reliable, and actionable—a "humility paradox" suggesting that hedging and offering alternatives can enhance users' trust in exploratory tasks. Differences across models were limited, affecting only perceived certainty, indicating that communication style can shape user perceptions as strongly as the underlying model. These results exemplify how evaluating a single UCLLM dimension can surface user-centered insights that system benchmarks overlook, and they illustrate the value of UCLLM as a structured lens for designing future multi-dimensional evaluations.

**Contribution Statement.** This paper makes the following contributions: (1) A user-centric evaluation framework (UCLLM) that organizes LLM assessment across nine user-experience dimensions; (2) A comparative pilot blueprint demonstrating how one UCLLM dimension can be operationalized and used to reveal user-centered differences between models; and (3) Design insights that highlight the importance of user-centered evaluation for selecting, deploying, and designing LLMs aligned with real-world user needs.

## 2 Related Work

This work builds on prior work on 1) User-centric evaluation frameworks and 2) LLMs Evaluation.

### 2.1 User-Centric Evaluation Frameworks

Early surveys such as Peras [51] and Nsaif et al. [49] grouped chatbot evaluation metrics, highlighting the divided and ad hoc nature of early efforts. More recent work moves beyond system benchmarks toward user-centered perspectives. Frameworks such as CoRE [14] and CRS-Que [33] assess conversational recommender systems through hybrid automated–human evaluation, while CONCEPT [29] and URS [67] incorporate social awareness and multi-intent evaluation. Methodological refinements also include ConSiDERS [22], which addresses evaluator bias, and CoPrompter [34], which supports prompt refinement through editable checklists. At a broader level, Ray et al. [55] proposed a six-layered evaluation

[1] https://platform.openai.com/docs/models/o4-mini, last accessed 04-03-2025.
[2] https://chat.chatbot.app/gemini?, last accessed 04-03-2025.
[3] https://chat.ai-pro.org/chat/claude/, last accessed 04-03-2025.
[4] https://www.deepseek.com/en, last accessed 04-03-2025.

framework integrating user feedback and ethical alignment, and Ghassempour[5] critiqued binary feedback systems, advocating for richer, structured user input.

Alongside these general frameworks, domain-specific approaches have emerged. In healthcare, CAPE [62] measures empathy, privacy, and therapeutic alliance, while Denecke [18] and DISCOVER [19] provide structured checklists and iterative design processes for medical agents. In education, Kasneci et al. [36] and Tlili et al. [64] examine opportunities and risks of ChatGPT in classrooms. Ahmed et al. [2] evaluate learning analytics recommendations, and Bahja et al. [5] emphasize collaborative, learner-centered chatbot design. In finance, Yang et al. [71] introduced UCFE, and Urbani et al. [65] assessed organizational readiness using the Technology Acceptance Model. Customer-service research includes qualitative analyses of dialogue relevance by Følstad and Taylor [24].

These frameworks provide valuable insights but are largely context-specific and limited to narrow slices of the interaction space. They tend to emphasize particular facets—such as conversational flow, empathy, organizational readiness, or benchmark-style performance without offering transferable structures that span diverse LLM use cases and interaction contexts. As a result, many important experiential aspects of LLM interaction remain fragmented or underrepresented, including cognitive load, trust calibration, agency, adaptivity, hallucination, and inclusion. Our work addresses this gap by integrating these scattered insights into a unified structure that synthesizes themes across prior frameworks and organizes them into nine complementary user-experience dimensions in UCLLM.

## 2.2 LLMs Evaluation

Research on evaluating LLMs spans three strands benchmark-driven comparisons, user-perception studies, and work on epistemic stance, each illuminating different aspects of LLM behavior yet lacking an integrative structure.

*Benchmark-driven evaluations* provide the dominant paradigm for comparing LLMs, with platforms such as Chatbot Arena [15] and MT-Bench [73] pioneering large-scale preference-based comparisons through LLM-as-judge scoring and crowdsourced votes. While influential, these benchmarks primarily quantify relative performance rankings rather than capturing deeper experiential qualities of user interaction or the nuances of user trust.

A complementary line of work examines *how users perceive and interpret LLM behavior*, focusing on user impressions beyond accuracy scores. Liu et al. [44] collected in-the-moment feedback identifying variation in verbosity, personalization, and credibility across models. Li et al. [42] also demonstrated that human judgments often diverge from LLM-based evaluators. Controlled studies by Miller et al. [48] and Rudd et al. [57] show that trust, usability, and satisfaction often diverge from technical performance, underscoring the importance of user-centered comparative evaluation.

Prior work highlights *epistemic stance and communication style as critical experiential factors*, showing how expressions of certainty or doubt shape users' trust and reliance. Frameworks such as Epistemic Alignment [17], REL-AI [75], and Trust Gate [26] emphasize the importance of communicative cues in human–AI alignment. Although users can distinguish qualitative differences between models, systematic experimental comparisons remain limited.

Taken together, existing LLM evaluations are fragmented: benchmark approaches emphasize technical capability, while user-centered studies highlight perceptions of trust, stance, and conversational style. Yet no unifying structure connects these experiential dimensions across tasks, contexts, and interaction goals. UCLLM addresses this gap by integrating recurring themes from prior work—such as trust, interaction quality, cognitive effort, adaptivity, and safety—into a nine-dimension framework that supports systematic and comparable user-centered evaluation.

## 2.3 Summary and Research Gap

Across both framework-driven and empirical evaluation research, prior work demonstrates clear interest in user-centered assessment of LLMs but remains fragmented across benchmarks, domain-specific tools, and isolated experiential studies. Existing approaches typically focus on narrow facets such as conversational flow, empathy, or preference alignment—and therefore fail to offer a transferable structure that captures the breadth of user experience with generative models. What is missing is a unified, multi-dimensional framework that integrates these perspectives and supports systematic evaluation across tasks, contexts, and interaction goals.

UCLLM addresses this gap by synthesizing recurring themes from prior work into nine complementary user-experience dimensions and by providing a structured foundation for designing empirical studies. In the remainder of the paper, we introduce the framework and demonstrate, through a pilot study of epistemic stance, how one dimension can be operationalized as a blueprint for evaluating the others.

# 3 UCLLM: A User-Centric LLM Evaluation Framework

In this section, we introduce *UCLLM*, a user-centric framework for evaluating LLMs. The framework contains nine dimensions, each defined through its scope, user perspective, and representative evaluation metrics (Table 1). These dimensions were not chosen ad hoc; they emerged from an iterative synthesis of HCI, UX, and conversational-system evaluation literature, combined with our analysis of concrete LLM use cases and interaction failures. UCLLM captures the breadth of experiential factors that repeatedly surface in prior work such as trust, cognitive effort, adaptivity, and safety, while organizing them into a coherent structure that spans diverse interaction contexts. The framework is designed to be conceptual and actionable: it offers a vocabulary for discussing user-centered evaluation and a practical foundation for designing empirical studies, as we illustrate by operationalizing one dimension in our pilot.

## 3.1 Methodological Derivation of UCLLM

We derived UCLLM through an integrative synthesis of prior work and observed LLM interaction patterns. Our starting point was a broad set of constructs drawn from established evaluation frameworks for dialog systems and recommender systems (e.g., PARADISE [66], CRS-Que [33], CoRE [14], UCFE [71]), HCI and UX

[5]https://medium.com/data-science-at-microsoft/beyond-thumbs-up-and-thumbs-down-a-human-centered-approach-to-evaluation-design-for-llm-products-d2df5c821da5, last accessed 04-03-2025.

models of usability and user experience (e.g., ISO 9241-11[6], adaptive interface design [31]), and human–AI guidelines on trust, safety, inclusion, and reliability [4, 41, 61, 69]. We complemented these sources with constructs emerging from recent LLM application domains (e.g., education, healthcare, and creative work), ensuring coverage of generative-AI–specific concerns such as hallucinations, prompt literacy, and user agency.

The derivation process combined deductive and inductive reasoning. Initially, constructs were deductively collected from prior evaluation frameworks, HCI theories, and human-centered AI guidelines. The authors then collaboratively conducted iterative conceptual grouping and refinement sessions to identify recurring experiential patterns, resolve overlaps, and refine conceptual boundaries. Related constructs such as success metrics, conversational qualities, cognitive demands, confidence cues, personalization, and safety were clustered into higher-level experiential themes.

Throughout the refinement process, candidate dimensions were evaluated with respect to clarity, distinctiveness, applicability across LLM contexts, and relevance to end-user interaction. Constructs that were overly system-centric, redundant, or insufficiently actionable from a user-evaluation perspective were merged, reframed, or excluded. Through multiple refinement rounds, we converged on nine dimensions that collectively provide broad coverage while minimizing conceptual redundancy.

We additionally established conceptual boundaries between closely related themes, such as distinguishing task success from factual correctness, cognitive load from learnability, and user control from model agency. These boundaries are summarized in Section 3.3 and detailed in Appendix A. Together, the resulting nine dimensions form a structured foundation for user-centered LLM evaluation and highlight the framework described in Section 3.2. The final dimensions were retained when they appeared repeatedly across the reviewed literature, captured a distinct user-centered concern, and could be operationalized through empirical evaluation.

## 3.2 UCLLM Dimensions

Building on the synthesis described in Section 3.1, UCLLM comprises nine dimensions that reflect recurring experiential themes across prior work and observed LLM use. Each dimension captures a distinct aspect of user experience, framed through a user perspective and linked to concrete evaluation methods. Below, we provide an overview of each dimension, with detailed metrics in Table 1.

*3.2.1 Core Utility.* This dimension captures an LLM's fundamental usefulness. Building on HCI and usability standards (ISO 9241-11 [6]) and classic dialog evaluation [66], usefulness is best judged in open-ended, multi-goal tasks such as brainstorming or sense-making, where success emerges through dialogue rather than fixed end points. Core Utility can be examined through three complementary sub-dimensions:

- *Task Success*: Evaluates whether the model helps users achieve their goals (e.g., completion rate, time-to-answer, usefulness rating), typically measured with scenario tasks and expert judgment.
- *Cognitive Load*: Assesses the effort required to interpret or apply the output, using metrics such as NASA-TLX scores, clarification counts, or verbosity.
- *Cost vs. Value*: Assesses whether the benefit of the output justifies the user's time and effort, via indicators like time saved, cost per task, or willingness to pay.

*3.2.2 Interaction Quality.* This dimension reflects the conversational dynamics between user and system. Conversational quality has long been evaluated in dialog systems and CRSs (e.g., PARADISE; CRS-Que's user-centric constructs) [33, 66]. For LLMs, we examine whether models stay coherent over extended exchanges, follow instructions well, and recover from misunderstandings:

- *Persona Fit*: Evaluates whether the LLM's tone and style match user expectations (e.g., tone appropriateness, empathy, cultural sensitivity).
- *Dialogue Flow*: Assesses the naturalness of multi-turn conversations, using metrics such as context retention, turn coherence, and breakdown occurrences.
- *Control*: Measures how easily users can shape outputs to their needs, via style matching, constraint adherence, or edit distance.
- *Error Recovery*: Evaluates whether users can correct or redirect the model effectively, using recovery rate and the number of correction turns.

*3.2.3 User Experience & Emotion.* This dimension captures how interacting with the model shapes users' feelings and overall satisfaction. This dimension includes valence–arousal ratings using the Self-Assessment Manikin (SAM) [9] and satisfaction scores collected through short surveys, diary entries, or post-task reflections.

*3.2.4 Learnability & Usability.* This dimension describes how quickly and comfortably users can adopt and operate the system. Since LLMs rely on open-ended prompting rather than fixed UI elements, we focus on developing "prompt literacy" and adapting to different skill levels. Measures include time to proficiency, initial error rates, and standardized usability scales such as SUS [10] or UEQ [40].

*3.2.5 Trust & Transparency.* This dimension concerns how the system signals certainty and supports appropriate user trust. An important component is epistemic stance—how the model communicates confidence and acknowledges uncertainty [41]. This dimension includes:

- *Confidence*: How clearly models indicate certainty and support verification (confidence cues, citation quality).
- *Trust*: Whether users develop appropriate trust—neither overtrust nor distrust—measured via trust-alignment scores or behavioral indicators.
- *Explainability*: Clarity and usefulness of the model's explanations, typically evaluated with ask-for-reason tasks.

*3.2.6 Content Quality.* This dimension focuses on the factual accuracy, novelty, and instructional clarity of outputs.

- *Hallucination*: Accuracy and grounding, measured through hallucination rate, citation correctness, or fact-checking against documents.
- *Creativity*: The novelty and usefulness of generated ideas, using originality ratings and diversity metrics.

[6]ISO 9241-11:https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en

**Table 1: UCLLM: A user-centric framework for evaluating large language models across nine dimensions with associated user perspectives and evaluation metrics.**

| Category | Dimension | User Perspective | Evaluation Metrics | Example Methods |
|---|---|---|---|---|
| **Core Utility** | Task Success | Did the model help users reach their goal? | Completion rate; time-to-answer; usefulness rating | Scenario tasks; human judgment |
| | Cognitive Load | How much effort is needed to understand or use the output? | NASA-TLX; number of clarifications; verbosity | Usability testing; time-on-task |
| | Cost vs. Value | Is the output worth the user's time and effort? | Time saved; cost per task; willingness-to-pay | Baseline comparison; diary studies |
| **Interaction Quality** | Persona Fit | Does the LLM's tone and style match user expectations? | Tone appropriateness; empathy rating; cultural sensitivity | Role-based prompts; A/B tests |
| | Dialogue Flow | Are multi-turn conversations smooth and natural? | Context retention; turn coherence; breakdowns | Multi-turn dialogue tests |
| | Control | Can users shape outputs to their needs? | Style match; constraint adherence; edit distance | Style/constraint prompts |
| | Error Recovery | Can users easily correct mistakes or redirect the model? | Recovery rate; number of correction turns | Iterative task prompts |
| **User Experience & Emotion** | Emotional Response & Satisfaction | How does interacting with the model influence users emotionally? Do they find it satisfying or frustrating? | Valence/arousal ratings; satisfaction surveys | Self-Assessment Manikin (SAM); user satisfaction questionnaires |
| **Learnability & Usability** | Learnability & Ease of Use | How quickly can users learn to use the system, and how easy is it to use? | Time to proficiency; ease-of-use ratings; SUS/UEQ scores | Time-on-task analysis; System Usability Scale (SUS); User Experience Questionnaire (UEQ) |
| **Trust & Transparency** | Confidence | Does the model clearly indicate certainty and support verification? | Helpfulness of confidence cues; citation quality | Confidence reporting tasks |
| | Trust | Do users trust the model appropriately (not over/under)? | Trust alignment; change in trust after cues | Questionnaires; behavior analysis |
| | Explainability | Are model explanations clear and helpful? | Rationale clarity; usefulness of explanations | Ask-for-reason tasks |
| **Content Quality** | Hallucination | Are outputs accurate and grounded? | Hallucination rate; citation correctness | Grounded Question-Answering; document-based tasks |
| | Creativity | Can it generate new and useful ideas? | Originality rating; diversity across outputs | Creative tasks; human judgment |
| | Instructional Quality | Can it teach or adapt explanations to users? | Learning gain; use of analogies | Pre/post tests; explanation scoring |
| **Adaptivity & Consistency** | Personalization | Does it adjust to user preferences or level? | Preference alignment; fewer clarifications needed | Progressive tasks; adaptation latency |
| | Consistency | Are outputs stable across prompts and sessions? | Variance across paraphrases; reproducibility | Prompt families; repeated sessions |
| | Memory & Privacy | Does it use past context correctly without privacy issues? | Memory accuracy; safe retention | Long-session tests; privacy checks |
| **Safety & Inclusion** | Safety | Does it avoid harmful or biased outputs? | Harm incidents; refusal quality | Red-team scenarios; demographic sampling |
| | Inclusion | Can diverse users benefit equally? | Accessibility and usability across groups | Diverse samples; assistive tech workflows |
| **User Empowerment** | Agency | Does the model empower users rather than replace them? | Suggestions vs. prescriptions; perceived agency | Decision-support tasks; surveys |

- *Instructional Quality*: The model's ability to teach or clarify information, using learning gains or analogy use.

*3.2.7 Adaptivity & Consistency.* This dimension describes how the model personalizes interaction while maintaining stable behavior. Personalization has deep roots in recommender systems [56] and adaptive UI design [31], and consistency is central in dialogue evaluation [35]. This dimension includes:

- *Personalization*: Alignment with user preferences or skill level.
- *Consistency*: Stability of outputs across paraphrased prompts or sessions.
- *Memory & Privacy*: Appropriateness of using past context and safe handling of sensitive information.

*3.2.8 Safety & Inclusion.* This dimension ensure equitable, non-harmful interactions across diverse user groups. Guided by FAccT principles [4] and accessibility standards such as WCAG 2.2 [69], this dimension includes:

- *Safety*: Avoidance of harmful or biased outputs, assessed through red-teaming or sensitive scenarios.
- *Inclusion*: Accessibility and usability across demographics, dialects, and ability groups.

*3.2.9 User Empowerment.* This dimension concerns whether the model strengthens users' judgment and agency rather than replacing them. Rooted in human-centered AI principles [61] and participatory design [25], this dimension includes:

- *Agency*: The extent to which the model supports user decision-making, measurable through the balance between suggestions and prescriptions or perceived agency ratings.

## 3.3 Dimension Boundaries and Overlaps

Because experiential aspects of LLM interaction often co-occur, UCLLM defines boundaries between dimensions to support consistent evaluation and avoid double-counting. The goal is not strict statistical independence, but ensuring that each dimension addresses a distinct user-centered question and can be assessed separately.

Table 2 summarizes conceptual relationships among sub-dimensions (✓ indicates overlap). The overlaps were identified through iterative conceptual comparison of sub-dimensions based on their user-facing concerns, interaction outcomes, and evaluation focus. Detailed boundary notes are provided in Appendix A. The examples below illustrate how related concepts are distinguished.

*Task Success vs. Hallucination.* Task Success measures whether users achieve their goals, whereas Hallucination concerns factual correctness and grounding. An interaction may help users complete a task despite minor inaccuracies, while a correct answer may still fail to address the user's objective. Treating these separately prevents conflating usefulness with factual accuracy.

*Control vs. Agency.* Control refers to the user's ability to steer the model (e.g., through prompts or corrections), while Agency reflects the model's initiative in shaping the interaction, such as offering suggestions. Although both affect collaboration, they represent opposite dynamics—user-directed versus model-driven behavior—and are therefore evaluated independently.

*Cognitive Load vs. Learnability.* Cognitive Load describes the immediate effort required to interpret or use a response, whereas Learnability concerns how quickly users develop proficiency over time. Distinguishing these dimensions separates momentary effort from longer-term learning processes.

These examples illustrate UCLLM's guiding principle: when interactions influence multiple experiential factors—such as factuality, effort, trust, or control—each should be evaluated within its respective dimension. Table 2 and Appendix A provide further guidance for consistent application of the framework.

The overlaps shown in Table 2 primarily represent *partial conceptual overlap* rather than substantial redundancy. An overlap indicates that two sub-dimensions may be influenced by related interaction phenomena or may co-occur during evaluation, while still addressing different user-centered concerns. We considered overlap to be substantial only when two constructs captured nearly identical evaluative goals or user outcomes; such cases were merged or excluded during refinement. Consequently, the remaining overlaps in UCLLM reflect complementary relationships rather than duplicated constructs.

## 3.4 Applying UCLLM in Empirical Work

UCLLM is designed not only as a conceptual structure but also as a practical guide for user-centered evaluation. Each dimension can be operationalized by identifying its sub-dimensions, selecting metrics that capture the associated user perspective, and designing tasks or interactions in which these metrics can be meaningfully observed.

The pilot study in Section 4 illustrates this process for one dimension, Trust & Transparency by manipulating epistemic stance and measuring users' perceptions of certainty, reliability, and actionability. While focused on a single dimension, the same methodological pattern can be extended to other dimensions in UCLLM, enabling systematic and comparable evaluation across models and contexts.

# 4 Evaluation

We illustrate the empirical use of UCLLM by operationalizing one of its dimensions, Trust & Transparency, in a pilot study that manipulates epistemic stance and measures users' perceptions of certainty, reliability, and actionability. Rather than validating the entire framework, this study demonstrates how individual UCLLM dimensions can be translated into measurable experimental constructs.

## 4.1 Study Design

We used a within-subjects design to examine how users perceive differences in LLMs' epistemic stance, specifically the level of expressed certainty. Prior work shows that confidence cues and linguistic markers of certainty influence user trust, reliance, and satisfaction [45, 59]. As these cues correspond directly to the Trust & Transparency dimension within UCLLM, epistemic stance provides a concrete way to operationalize this dimension empirically.

Our pilot manipulated two independent variables: (1) *LLM identity* and (2) *certainty level* (high vs. low epistemic stance). We selected four widely deployed and publicly accessible LLMs—OpenAI o4-mini[1], Gemini 2.5 Pro[2], Claude Opus 4[3], and DeepSeek V3[4]—which represent diverse model families and consistently appear in the top tier of human preference benchmarks such as LMSYS Chatbot Arena[7]. This diversity makes them suitable for comparative user-centered evaluation.

We measured four dependent variables aligned with Trust & Transparency: *perceived certainty*, *reliability*, and *actionability*, and *enjoyment* to contextualize participants' overall experience. Participants completed eight conditions (four models × two certainty levels), counterbalanced via a Latin square to mitigate order effects.

To reduce extraneous variability while preserving natural interaction, several aspects of the study were held constant across conditions: all participants received the same initial prompt and scenario, interactions were limited to five participant–model exchanges, and the interface and instructions remained identical across conditions. Participants could ask natural follow-up questions within the travel-planning task, but were instructed to remain within the assigned scenario. The stance manipulation was implemented through the high-certainty and low-certainty pre-prompts described in Section 4.3, which were applied across all study conditions to induce different epistemic response styles. The study was approved by the institutional ethics board and lasted approximately 45 minutes.

While our focus here is on epistemic stance, UCLLM encompasses additional dimensions such as safety, adaptivity, and interaction quality. Evaluating these dimensions requires different

[7] Chatbot Arena: https://lmarena.ai/leaderboard, last accessed 04-03-2025.

**Table 2: Comprehensive orthogonality matrix for all UCLLM sub-dimensions. ✓ indicates a conceptual overlap and empty cells indicate independence.**

| UCLLM Sub-dimensions | TS | CL | VAL | PF | DF | CTRL | REC | EMO | LEARN | CONF | TRUST | EXPL | HALL | CREAT | INSTR | PERS | CONS | MEM | SAFE | INCL | AGCY |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Task Success (TS) | | ✓ | ✓ | | ✓ | | ✓ | | | | | | ✓ | | ✓ | | | | | | ✓ |
| Cognitive Load (CL) | ✓ | | ✓ | | ✓ ✓ | ✓ | ✓ | ✓ | | | ✓ | | | ✓ | | | | | | ✓ | |
| Value (VAL) | ✓ | ✓ | | | | | | ✓ | | | | | | ✓ | | ✓ | | | | ✓ | |
| Persona Fit (PF) | | | | | ✓ | | | ✓ | | | | | | | | ✓ | | | | ✓ | |
| Dialogue Flow (DF) | ✓ | ✓ | | ✓ | | | ✓ | ✓ | ✓ | | | | | | | | | | | | |
| Control (CTRL) | | ✓ | | | | | | | | | | | | | | ✓ | | ✓ | | | ✓ |
| Error Recovery (REC) | ✓ | ✓ | | | ✓ | | | | | | | | | | | | | | | | |
| Emotion (EMO) | | ✓ | ✓ | ✓ | ✓ | | | | | | | | | | | ✓ | | | ✓ | ✓ | ✓ |
| Learnability (LEARN) | | ✓ | | | ✓ | | | | | | | | | | ✓ | ✓ | | | | | |
| Confidence (CONF) | | | | | | | | | | | ✓ | | | | | | | | | | |
| Trust (TRUST) | | | | | | | | | | ✓ | | ✓ | ✓ | | | | | | ✓ | | |
| Explainability (EXPL) | | ✓ | | | | | | | | | ✓ | | | | ✓ | | | | | | |
| Hallucination (HALL) | ✓ | | | | | | | | | | ✓ | | | | ✓ | | ✓ | | ✓ | | |
| Creativity (CREAT) | | | ✓ | | | | | | | | | | | | | | | | | | ✓ |
| Instructional Quality (INSTR) | ✓ | ✓ | | | | | | | ✓ | | | ✓ | ✓ | | | | | | | | |
| Personalization (PERS) | | | ✓ | ✓ | | ✓ | | ✓ | ✓ | | | | | | | | | | | ✓ | |
| Consistency (CONS) | | | | | | | | | | | | | ✓ | | | | | | | | |
| Memory & Privacy (MEM) | | | | | | ✓ | | | | | | | | | | | | | | | |
| Safety (SAFE) | | | | | | | | ✓ | | | ✓ | | ✓ | | | | | | | | |
| Inclusion (INCL) | | | | ✓ | | | | ✓ | | | | | | | | | | | | | |
| Agency (AGCY) | ✓ | ✓ | ✓ | | ✓ | ✓ | | ✓ | | | | | | ✓ | | ✓ | | | | | |

manipulations and measures; thus, fully testing all nine dimensions is part of future work.

This pilot study demonstrates how one UCLLM dimension can be systematically operationalized and evaluated, providing a concrete example of how the framework may be applied in empirical settings.

## 4.2 Scenario

We selected a travel-planning scenario as a focused, low-stakes domain that naturally elicits reasoning under uncertainty and recommendation behavior. This makes it well-suited for examining epistemic stance, which is central to the Trust & Transparency dimension. The goal of the scenario was not to achieve broad generalizability, but to provide a tractable and ecologically valid context in which stance-related cues could be expressed and evaluated.

Participants were asked to plan a 3-day solo trip. This task offers a goal-oriented yet open-ended structure, allowing for variation in preference elicitation while maintaining comparable interactions. Similar travel-planning tasks have been widely used in HCI and conversational-agent research to study decision support and recommendation quality [23, 32], making it an appropriate context for cross-model comparison.

## 4.3 Epistemic Stance Creation via Pre-Prompting

To manipulate epistemic stance, conditions began with a system-level pre-prompt designed to induce either a high-certainty or low-certainty response style. This manipulation highlights confidence-related interaction cues associated with the Trust & Transparency dimension of UCLLM. The *High-certainty* prompt instructed the model to use decisive language and provide singular recommendations without hedging. The *Low-certainty* prompt encouraged acknowledging uncertainty, presenting alternatives, and using hedging markers. This style-based behavioral steering aligns with prior work showing that pre-prompts can reliably influence LLM response characteristics [50, 52].

**High-certainty** You are a confident and assertive travel planning assistant. You speak with clarity and certainty in your recommendations. Respond to the user's questions by selecting the best possible option and stating it as a definitive recommendation. Avoid hedging phrases like "maybe," "it depends," or "you might want to." Your tone should be that of a knowledgeable travel expert who knows what works best. Do not present multiple alternatives unless asked to. When giving advice, use strong and direct language, e.g., "You should go to Kyoto" or "This is the perfect itinerary for a solo traveler." Do not focus on your limitations or express doubt unless absolutely necessary. Remain calm, professional, and confident.

**Low-certainty** You are a cautious and thoughtful travel planning assistant. Your goal is to acknowledge the uncertainty that often comes with travel advice. Offer multiple options and highlight factors that may affect the user's choice (e.g., preferences, season, budget). Use hedging phrases such as "one option could be. . . ," "it might depend on. . . ," or "you may want to consider. . . " Your tone should be friendly, balanced, and informative—not overly assertive. Avoid strong claims. Let the user feel in control of the decision-making process by presenting possibilities rather than definitive answers. Feel free to acknowledge when you are unsure or when preferences may vary. Avoid assuming too much about the user's needs unless they provide specifics.

**Table 3: Average number of hedging markers per user across models and stance conditions.**

| Model | Stance | Avg. Hedging per user | Count of Hedging |
|---|---|---|---|
| Claude | High | 1.72 | 42 |
| Claude | Low | 2.25 | 54 |
| DeepSeek | High | 0.66 | 16 |
| DeepSeek | Low | 0.71 | 17 |
| Gemini | High | 3.88 | 93 |
| Gemini | Low | 4.67 | 112 |
| GPT-4o-mini | High | 1.21 | 29 |
| GPT-4o-mini | Low | 1.33 | 32 |
| All models | High | 1.5 | 180 |
| All models | Low | 1.79 | 215 |

As models differ in how they interpret prompts, the exact linguistic expression of the stance may vary across LLMs. Our goal was not to produce identical outputs, but to test whether stance-related cues could be systematically manipulated across models.
To verify that the manipulation influenced linguistic behavior, we conducted a post-hoc analysis of stance markers (Table 3), which showed modest but consistent differences between conditions. These effect sizes were small, reflecting both natural model variation and the subtlety of linguistic stance cues. Future work will complement this automated analysis with coder-based annotation to provide a more robust manipulation check. These pre-prompts were directly used as the experimental manipulation in the pilot study described above, where participants interacted with each model under both certainty conditions.

## 4.4 Participants and Recruitment

We recruited 24 participants through university mailing lists, social media, and word of mouth (9 females, 15 males). Ages ranged from 18 to 25 years ($M = 25.1$, $SD = 1.9$). Participants reported basic to intermediate experience with LLMs, with an average daily use of 13 out of 24 possible interactions. Most participants primarily used ChatGPT ($N = 19$), followed by Claude ($N = 2$), while others reported no regular use. Backgrounds spanned engineering ($N = 11$), computer science ($N = 8$), medicine ($N = 3$), and arts ($N = 2$).

## 4.5 Procedure and Task

Participants first reviewed the study aim and provided informed consent. They were told they would interact with several LLMs across a real-world task but were not informed about the epistemic-stance manipulation. After completing a demographics questionnaire, participants proceeded through all eight conditions (four models × two stance levels), presented in a Latin-square order.

Each condition began with a standardized prompt:

> "I want to plan a 3-day solo trip. Can you help me figure out where to go and what to do?"

Participants were allowed to introduce minor preference-based adjustments during subsequent conversational exchanges (e.g., interests in nature, food, budget, or activities) in order to preserve ecological validity and support natural interaction. These adjustments were limited to elaborations within the travel-planning task and did not alter the overall scenario structure.

Each interaction consisted of five participant–model exchanges (i.e., five participant messages and five corresponding model responses). Following the initial standardized prompt, participants were free to formulate follow-up questions naturally within the task scope rather than following a fixed script. To maintain comparability across conditions, all interactions used the same task framing, interface, and turn limit.

After each interaction, participants completed four 7-point Likert-scale items assessing: (1) *perceived certainty* ("The assistant seemed confident in its recommendations"), (2) *trust and reliability* (The assistant's recommendations felt reliable and trustworthy"), (3) *actionability* ("The assistant provided advice I would feel comfortable following"), and (4) *overall enjoyment* ("I enjoyed interacting with this assistant."). Responses ranged from 1 (*strongly disagree*) to 7 (*strongly agree*). These items were adapted from validated trust-in-automation and human-robot interaction measures [6, 7] to fit the conversational LLM context and the Trust & Transparency dimension operationalized in this study. As this study serves as a pilot operationalization of one UCLLM dimension, we used concise single-item measures targeting distinct experiential constructs rather than full multi-item psychometric scales.

Following each condition, participants additionally provided brief open-ended reflections about the interaction, including comments on response quality, confidence, clarity, and overall experience. At the end of the study, participants also completed a broader comparative reflection across systems. These open-ended responses formed the basis of the qualitative analysis described in Section 5.4.

# 5 Results

Table 4 shows the descriptive statistics for each dependent variable, perceived *certainty*, *reliability*, *actionability*, and *satisfaction*, across the eight experimental conditions (4 LLMs × 2 stance levels).

Across measures, ratings were uniformly high (many cell means near 6 on a 7-point scale). A consistent pattern emerged: for almost every model, scores in the *Low-certainty* condition slightly exceeded those in the *High-certainty* condition. For instance, participants rated Gemini's certainty at 6.46 ± 0.78 (high) and 6.33 ± 0.87 (low), while OpenAI o4-mini averaged 5.58 ± 1.56 (high) and 6.04 ± 1.00 (low). Similar trends appeared for reliability and actionability, with DeepSeek low-certainty responses (6.25 ± 1.11 reliability) exceeding their high-certainty counterparts (5.71 ± 1.30). Satisfaction showed the same direction but with smaller absolute differences. Table 5 summarizes all inferential results.

## 5.1 Epistemic Stance Effects

To test whether the descriptive differences were reliable, we ran repeated-measures ANOVAs with *stance* (high vs. low) and *model* (Claude, DeepSeek, Gemini, OpenAI o4-mini) as within-subject factors. We found significant main effects of *stance* on **certainty**, $F(1, 23) = 6.30$, $p = .020$, $\eta_p^2 = .22$; *reliability*, $F(1, 23) = 6.50$, $p = .018$, $\eta_p^2 = .22$; and *actionability*, $F(1, 23) = 4.83$, $p = .038$, $\eta_p^2 = .17$. No stance effect was observed for *satisfaction*, $F(1, 23) = 2.81$,

**Table 4: Descriptive statistics (means ± SD) for LLM × Epistemic Stance interaction.**

| LLM | Stance | Certainty | Reliability | Actionability | Satisfaction |
|---|---|---|---|---|---|
| Claude | Low | 6.12 ± 1.12 | 6.17 ± 1.13 | 6.12 ± 1.12 | 6.12 ± 1.12 |
| | High | 5.50 ± 1.59 | 5.42 ± 1.74 | 5.46 ± 1.61 | 5.42 ± 1.82 |
| DeepSeek | Low | 6.21 ± 1.06 | 6.25 ± 1.11 | 6.17 ± 1.09 | 6.04 ± 1.37 |
| | High | 5.96 ± 1.16 | 5.71 ± 1.30 | 5.75 ± 1.51 | 5.42 ± 1.93 |
| Gemini | Low | 6.33 ± 0.87 | 6.00 ± 1.38 | 6.21 ± 1.18 | 5.83 ± 1.34 |
| | High | 6.46 ± 0.78 | 6.33 ± 0.96 | 6.38 ± 1.01 | 6.29 ± 1.08 |
| o4-mini | Low | 6.04 ± 1.00 | 5.83 ± 1.49 | 5.83 ± 1.37 | 5.62 ± 1.58 |
| | High | 5.58 ± 1.56 | 5.58 ± 1.56 | 5.62 ± 1.64 | 5.62 ± 1.69 |

$p = .11$. Participants thus tended to rate low-certainty responses as more certain, reliable, and actionable.

To verify that the certainty manipulation was reflected in the model outputs, we conducted a post-hoc Hyland stance marker analysis where we counted hedging words and phrases namely, "might", "may", "could", "perhaps", "possibly", "apparently", "generally", "often", "sometimes", "seems", "appear", "likely" [30]. We found that low-Certainty conversations contained more hedging expressions on average ($M = 1.79$) than High-Certainty conversations ($M = 1.5$), suggesting that the manipulation influenced linguistic style in the intended direction (see Table 3). However, a Welch's t-test did not show statistical significant difference ($t(585.28) = 0.95$, $p = .341$), indicating that the effect remained modest. This helps explain why some participants reported noticing clear differences while others perceived the models as more similar. This is also reflected in our qualitative analysis (Section 5.4).

## 5.2 Model Effects

A main effect of *model* emerged for **certainty**, $F(3, 69) = 3.07$, $p = .033$, $\eta_p^2 = .12$. Holm-corrected pairwise comparisons showed that *Gemini* ($M = 6.46$) was perceived as more certain than both *Claude* ($M = 5.81$) and *OpenAI o4-mini* ($M = 5.81$), $p < .05$ for each. No model differences were found for reliability, actionability, or satisfaction (all $p > .16$). Descriptively, Gemini showed slightly higher reliability ($M = 6.17$) and actionability ($M = 6.29$), though these differences were not statistically distinguishable.

## 5.3 Interaction Effects

Stance × model interactions were non-significant for all dependent variables (all $p > .10$), indicating that the stance-related pattern held across models.

Overall, participants favored low-certainty responses, judging them slightly more certain, reliable, and actionable than high-certainty ones. Among models, only perceived certainty differed reliably, with *Gemini* consistently rated higher than *Claude* and *OpenAI o4-mini*. Perceptions of reliability, actionability, and satisfaction were uniformly high across models and stance combinations.

To complement the ANOVA results, we ran supplementary linear mixed-effects models with *stance*, *model*, and their interaction as fixed effects, and *Participant* as a random intercept. This approach provides a more conservative check by modeling individual differences directly. The mixed-effects results showed the same overall trends as the ANOVA—low-certainty responses were generally preferred, and Gemini tended to score higher, but none of the effects reached significance once participant variability was taken into account. For *Certainty*, neither *model* ($\chi^2(3) = 5.92$, $p = .115$), *stance* ($\chi^2(1) = 2.19$, $p = .139$), nor their interaction ($\chi^2(3) = 3.19$, $p = .364$) was significant. The same pattern appeared for *Reliability* (all $p > .10$) and *Actionability* (all $p > .30$). For *Satisfaction*, the results did not meet the threshold for significance, ($\chi^2(3) = 7.12$, $p = .068$). Full results are shown in Table 6. Overall, the mixed-effects models suggest that the group-level patterns observed in the ANOVA are present but vary across participants. This aligns with the qualitative Feedback in the next section, where some users noticed clear differences between models, while others felt the systems behaved similarly.

**Table 5: Two-way repeated-measures ANOVA results for each dependent variable (*Stance*: high vs. low-certainty; *Model*: Claude, DeepSeek, Gemini, OpenAI o4-mini). For Certainty, significant Holm-adjusted pairwise model comparisons are shown in the lower block. Partial $\eta_p^2$ is reported as an effect size. Significant $p$ values (< .05) are bold.**

| DV | Effect / Comparison | Num df | Den df | F | Mean Diff. | $p$ | $\eta_p^2$ |
|---|---|---|---|---|---|---|---|
| Certainty | Stance | 1 | 23 | **6.30** | – | **.020** | .22 |
| | Model | 3 | 69 | **3.07** | – | **.033** | .12 |
| | Stance×Model | 3 | 69 | 1.29 | – | .29 | .05 |
| | Gemini – Claude | – | – | – | 0.59 | **.030** | – |
| | Gemini – o4-mini | – | – | – | 0.65 | **.024** | – |
| Reliability | Stance | 1 | 23 | **6.50** | – | **.018** | .22 |
| | Model | 3 | 69 | 1.21 | – | .31 | .05 |
| | Stance×Model | 3 | 69 | 2.07 | – | .11 | .08 |
| Actionability | Stance | 1 | 23 | **4.83** | – | **.038** | .17 |
| | Model | 3 | 69 | 1.73 | – | .17 | .07 |
| | Stance×Model | 3 | 69 | 1.27 | – | .29 | .05 |
| Satisfaction | Stance | 1 | 23 | 2.81 | – | .11 | .11 |
| | Model | 3 | 69 | 0.75 | – | .52 | .03 |
| | Stance×Model | 3 | 69 | 1.93 | – | .13 | .08 |

## 5.4 Qualitative Analysis

We conducted an exploratory thematic analysis of participants' written comments to contextualize the quantitative findings and understand how users interpreted stance cues and model behavior. We first open-coded the responses to identify recurring perceptions related to confidence, trust, clarity, reliability, and conversational experience. Related codes were then iteratively grouped into higher-level experiential themes through collaborative refinement among the authors. The analysis focused on identifying recurring interpretive patterns rather than solely quantifying response frequencies.

*5.4.1 Post-Interaction Comments.* Post-interaction reflections revealed several recurring themes related to how participants interpreted stance cues and overall interaction quality. One prominent theme concerned comparative *Preference / Ranking*, where participants directly contrasted systems in terms of conciseness, usefulness, and interaction style. These comparative judgments often shaped broader perceptions of quality and trustworthiness and appeared frequently across the comments (18 references). For example,

**Table 6: Mixed-effects likelihood-ratio test results for each dependent variable. None of the fixed effects reached statistical significance once participant-level variability was modeled.**

| DV | Effect | $\chi^2$ | df | $p$ |
|---|---|---|---|---|
| Certainty | LLM | 5.92 | 3 | .115 |
| | Stance | 2.19 | 1 | .139 |
| | Interaction | 3.19 | 3 | .364 |
| Reliability | LLM | 0.60 | 3 | .896 |
| | Stance | 1.44 | 1 | .229 |
| | Interaction | 6.21 | 3 | .102 |
| Actionability | LLM | 2.86 | 3 | .415 |
| | Stance | 1.04 | 1 | .308 |
| | Interaction | 3.65 | 3 | .301 |
| Satisfaction | LLM | -0.12 | 3 | 1.000 |
| | Stance | -0.18 | 1 | 1.000 |
| | Interaction | 7.12 | 3 | .068 |

one participant noted: "I did not like this one much; the other was more concise and to the point." (P4, OpenAI high-certainty).

Another recurring theme involved *Speed / Latency* (7 references), where participants commented on responsiveness and interaction fluency. For instance, one participant observed: "It was systematic, but a little slow to respond." (P10, OpenAI high-certainty). These reflections suggest that responsiveness influenced perceived interaction quality independently of stance-related confidence cues.

Participants also reflected on *Overconfidence / Unwarranted Certainty* (4 references), particularly when confident-sounding responses lacked sufficient justification. One participant described a response as "Very detailed and confident, but not always sure if it's accurate." (P6, Gemini low-certainty). These comments suggest that while assertive language increased perceived expertise, it simultaneously raised concerns about unsupported confidence.

Other themes concerned *Tone & Politeness* (3 references), *Clarity / Readability* (3), and *Actionability / Usefulness* (3). Participants often associated well-structured and polite responses with smoother interactions, while overly verbose explanations were sometimes perceived as less effective or practical. For example, one participant commented: "Seemed to understand my question, but explanations were a bit long-winded." (P14, DeepSeek high-certainty).

Further comments reflected broader evaluative concerns beyond stance alone, including *factual accuracy*, *personalization*, *structural clarity*, and *cultural appropriateness*. Although these themes appeared less consistently, they illustrate how participants simultaneously evaluated multiple interacting aspects of the conversational experience rather than focusing on confidence cues alone.

These themes reflect several sub-dimensions of UCLLM's Trust & Transparency category. Concerns related to confidence, justification, clarity, and perceived trustworthiness appeared repeatedly throughout participant reflections, supporting the relevance of epistemic stance as an operationalization of this dimension.

*5.4.2 Post-Study Comments.* Post-study reflections mirrored many of the interaction-level themes while also revealing broader comparative impressions across systems. Participants frequently discussed comparative *Preference / Ranking* (6 references), *Speed / Latency* (6), and instances where they perceived *No Difference* between models (6). Some participants reported strong preferences for particular systems, often associating them with greater trustworthiness or smoother interaction. For example, one participant stated: "3 and 7 were the best in my opinion, I would trust those more." (P22). Others emphasized responsiveness and interaction fluency: "I liked the first couple of AI models, they were fast and smooth." (P3).

At the same time, several participants perceived only limited distinctions between systems, suggesting variability in sensitivity to stance-related cues. One participant noted: "The differences were not significant." (P13). Less common themes included *Accuracy / Factuality*, *Clarity / Readability*, and *Transparency / Reasoning Quality*, as illustrated by comments such as: "Not all of them seemed to be up to date with information." (P14).

These reflections reinforce that users form trust based not only on explicit confidence cues, but also on broader aspects of communicative style and response behavior. They additionally highlight variability in how strongly users perceived stance differences, which aligns with the variability observed in the mixed-effects models.

Overall, the qualitative findings complement the quantitative results by showing that stance-related cues—such as hedging, tone, confidence, and clarity—influenced how users assessed reliability and actionability. At the same time, themes related to speed, verbosity, factuality, and personalization demonstrate that user experience is inherently multi-dimensional and shaped by factors beyond stance alone, further motivating the need for a broader framework such as UCLLM.

**Reproducibility.** To support transparency and reproducibility, we release anonymized chat logs from all study conditions and the post-study questionnaire [8]

## 5.5 Summary of Key Findings

The results provide initial insights into how epistemic stance shapes users' perceptions of LLMs. The manipulation of certainty was effective: the Low-certainty stance, characterized by offering alternatives and acknowledging uncertainty, was perceived as more confident, reliable, and actionable than the High-certainty stance. This "humility paradox" indicates that, in exploratory tasks such as travel planning, hedging can enhance rather than undermine perceived trustworthiness.

Model identity also played a role. Models, such as Gemini, elicited higher certainty ratings, suggesting that baseline model behavior interacts with stance cues. Across conditions, participants generally favored responses that balanced clarity, detail, and optionality. The qualitative analysis reinforces these results: participants frequently noted differences in confidence, justification, clarity, and actionability—factors closely aligned with *Trust & Transparency*.

Although the results were statistically significant, the effect sizes were modest ($\eta_p^2 < .06$), so these findings should be interpreted cautiously and viewed as consistent with the study's role as a pilot blueprint. To complement the ANOVA, we ran exploratory

[8] Repo: https://github.com/yasmeenessam/UCLLM-Pilot-Study-Dataset.git

mixed-effects models. These replicated the same overall pattern of effects, but the stance differences did not reach significance once participant-level variability was taken into account, which is consistent with the small effect sizes and the pilot nature of the study.

# 6 Discussion

We interpret our findings through the lens of user-centered LLM evaluation and the UCLLM Trust & Transparency dimension. We outline how epistemic stance and model behavior shaped user perceptions and reflect on methodology.

## 6.1 Certainty Effects on User Perceptions

A notable finding of this study was that the *Low-certainty* epistemic stance, which is designed to hedge and offer options, was consistently judged as more certain, reliable, and actionable than the *High-certainty* stance. Rather than interpreting hedging as indecisiveness, participants often treated it as a sign of thoughtfulness and collaboration. This pattern aligns with the idea of a "humility paradox," in which acknowledging uncertainty can enhance perceived confidence. In subjective, exploratory domains like travel planning, users may equate openness with competence. This contrasts with long-standing assumptions in decision-support research, where higher confidence is typically associated with greater trust [37]. It also suggests that stance preferences may be task-dependent: in everyday scenarios, openness may build comfort, while in higher-stakes settings decisiveness may still be expected.

The implication is that LLMs may benefit from adapting their epistemic stance to situational goals and user expectations. Looking ahead, adaptive systems that personalize stance to context [11] could better calibrate trust, moving beyond static personas toward dynamically aligned communication strategies [47].

Our supplementary mixed-effects models showed the same overall pattern as the ANOVA but also indicated variability across participants, suggesting that users differ in how strongly they respond to stance cues. This is consistent with the qualitative findings: some participants described clear differences between stances, while others perceived the systems as behaving similarly. Together, these results suggest that stance influences perceptions, but the degree to which users draw on these cues varies across individuals. These findings illustrate why user-centered evaluation dimensions such as Trust & Transparency are important alongside traditional benchmark-oriented metrics. While the models often produced similarly plausible travel recommendations, participants differentiated them based on confidence calibration, clarity, actionability, and conversational style. This reinforces the need for evaluation frameworks such as UCLLM that capture experiential and interaction-centered aspects of LLM use beyond correctness alone.

More broadly, the results demonstrate how seemingly subtle interaction cues—such as hedging, decisiveness, and conversational framing—can influence user perceptions in ways that are not easily captured through system-centric evaluation alone. By organizing these experiential factors into complementary dimensions, UCLLM provides a structure for studying how users experience and interpret LLM behavior across contexts.

## 6.2 Model Effects Beyond Epistemic Stance

We also found that model identity influenced perceived certainty: Gemini was rated more certain than Claude and OpenAI o4-mini ($p < .05$), whereas reliability, actionability, and satisfaction showed no differences across models. This contrasts with the stance-related effects and suggests that underlying model behavior still shapes how confident responses appear, even with identical stance prompts.

The stance effects were broadly consistent across models, indicating that epistemic stance cues can influence user perceptions across model families, even if the strength of these cues varies. This aligns with Zheng et al. [74], who report that stance-related manipulations interact with each model's conversational habits in distinct ways. Prompt-engineering work similarly shows that base model characteristics often dominate over stylistic adjustments [46].

Taken together, these patterns highlight the value of considering both stance and model identity when designing user-centered evaluations: stance cues shape perceptions, but they do so against a backdrop of model-specific tendencies. This also underscores the importance of verifying prompt effects directly such as through stance-marker analyses or other behavioral indicators, to understand how different models realize the same communicative prompt.

## 6.3 Designing for Trust in Interaction

Participants emphasized that their trust in the LLM was shaped not only by what the model said but by how it communicated. Responses that acknowledged ambiguity or explained their reasoning were often seen as more supportive, whereas single "best" answers sometimes felt less helpful. This aligns with research showing that users often prefer systems that enhance their sense of agency rather than providing directive solutions [3, 8, 53].

Our findings highlight a design trade-off: hedging can convey thoughtfulness and collaboration in open-ended tasks, yet may reduce clarity in high-stakes contexts such as healthcare or legal decision-making. While hedging was associated with greater reliability and thoughtfulness, some participant comments suggested that overly cautious or lengthy responses could reduce perceived decisiveness, whereas highly assertive responses could appear overly authoritative. Prior work on trust in automation similarly emphasizes balancing empathy, clarity, and factual reliability [41].

The qualitative feedback also suggests that trust judgments arise from multiple cues beyond stance—such as verbosity, responsiveness, and clarity—indicating that stance interacts with broader aspects of interaction quality. Future work could investigate how these factors interact to shape users' perceptions of trust and appropriateness across contexts.

## 6.4 Reflections on Applying UCLLM in Empirical Work

Our experience operationalizing the Trust & Transparency dimension offers insight into how UCLLM can inform empirical evaluation. The framework provides the conceptual structure, while the study demonstrates one concrete way of translating a dimension into measurable constructs, task designs, and outcome variables. In this sense, the approach used here serves as a blueprint for how other UCLLM dimensions may be operationalized in future studies.

First, designing an experiential manipulation required targeting subtle communicative cues, as illustrated by how stance was shaped through linguistic framing. This highlights the importance of combining prompt-based manipulations with complementary checks such as the stance-marker analysis. Second, participant comments indicated that stance cues were interpreted in conjunction with factors such as clarity, verbosity, and responsiveness, suggesting that evaluations informed by UCLLM should consider interactions across these dimensions. Third, combining quantitative and qualitative methods provided a richer understanding of how users interpret certainty cues.

These reflections arise from applying UCLLM to a single dimension and may extend to other dimensions. More broadly, they illustrate how UCLLM can guide researchers in linking conceptual dimensions to empirical designs and revealing interaction patterns that may remain underexplored in narrower evaluation approaches.

Unlike benchmark-oriented or narrowly task-focused evaluation frameworks, UCLLM emphasizes how users experience and interpret LLM behavior during interaction. For example, our findings showed that participants differentiated systems based not only on recommendation quality, but also on confidence calibration, conversational framing, actionability, and perceived reliability. These interaction-centered effects—including the observed "humility paradox," where hedged responses were perceived as more trustworthy—would likely remain underexplored in evaluations focused solely on correctness, preference rankings, or task completion. By organizing these experiential factors into complementary dimensions, UCLLM supports richer analysis of how users engage with and evaluate LLM systems in practice.

## 6.5 Rethinking Evaluation Frameworks

We designed our study to illustrate how one UCLLM dimension, Trust & Transparency, can be examined empirically through epistemic stance. The findings also show how challenging such evaluations become when model behavior, task context, and user expectations intersect. Prior work on contextualized evaluation has shown that assessments disconnected from the setting of use risk overlooking important interaction dynamics [21]. Similarly, research on epistemic agency highlights that reasoning, reflection, and belief updating can vary substantially with task framing [43]. These observations raise important questions: Is epistemic stance a stable property of an LLM, or does it depend on user goals and context? How can evaluations capture such variability while still enabling cross-model comparison?

These complexities highlight that our study does not validate UCLLM as a whole; rather, it tests the feasibility of operationalizing a single subdimension within the framework. A full validation of UCLLM would require systematic tasks and metrics for each dimension, studies across multiple domains with varying stakes, and evidence that the dimensions capture distinct yet complementary aspects of user experience. Longitudinal and multi-factor evaluations would also be needed to assess how dimensions interact over time and across contexts.

Viewed in this light, our findings emphasize the relevance of a contextualized, multi-dimensional framework: they highlight how context, stance cues, and model characteristics jointly shape user experience, and why single-metric or purely system-focused evaluations often miss these nuances.

## 6.6 Challenges, Opportunities, and Ethical Concerns

Evaluating epistemic stance is not straightforward. Models differ in how they interpret the same prompt, user expectations shift with context, and even minor system updates can alter behavior in ways that complicate replication. These challenges make it difficult to isolate variables and draw generalizable conclusions.

Looking more broadly, these difficulties reveal opportunities for user-centered evaluation beyond epistemic stance. Table 1 illustrates how all nine dimensions can be operationalized through concrete user perspectives, metrics, and methods. We position the study as a proof of concept, with UCLLM offering a roadmap for future empirical work. Study designs may integrate behavioral, qualitative, and physiological measures to capture user experience more fully, and domain-sensitive approaches can examine how LLMs behave in exploratory, sensitive, or socially complex settings. These observations suggest that frameworks such as UCLLM should incorporate guidance for evaluating dynamic, context-adaptive model behavior, including how multiple experiential dimensions interact in real use.

Future work could systematically validate the remaining UCLLM dimensions by pairing each with appropriate tasks and measurement strategies, for example, evaluating Interaction Quality through expert-coded multi-turn dialogues, Hallucination through grounded fact-checking protocols, Adaptivity through longitudinal personalization tasks, User Agency through controlled manipulations of user control, and Safety through expert-rated, risk-sensitive scenarios. Such evaluations can combine behavioral measures, qualitative feedback, and potentially physiological or attention-based signals to capture user experience comprehensively across contexts.

Finally, evaluations carry ethical implications. Miscalibrated certainty may mislead users in high-risk settings, while excessive hedging can reduce confidence when decisive guidance is needed. Personalization also raises questions of fairness and autonomy: tailoring stance or personas may improve trust but risks reinforcing biases or narrowing perspectives. Ethical evaluation therefore requires balancing accuracy, usability, and responsibility to ensure that LLMs foster trust without encouraging misplaced reliance [1].

## 6.7 Design Insights

Drawing on the UCLLM framework and the patterns observed in our study, we outline implications for designing and selecting LLM behaviors. These insights reflect tendencies in our data, which showed modest stance effects and some variability across participants and should be interpreted as considerations rather than prescriptive recommendations.

**Match certainty to context.** Low-certainty responses characterized by offering alternatives and acknowledging uncertainty were consistently, though modesty perceived as more certain, reliable, and actionable in our exploratory task. This suggests that open, option-oriented phrasing may be appropriate for creative or planning scenarios, yet more decisive communication may be needed in high-stakes domains.

**Test personas on each model.** Although stance effects appeared across models, perceived certainty varied by model identity (e.g., Gemini elicited higher certainty ratings). As models differ in how they realize the same prompt, evaluating personas on a per-model basis may help ensure consistent expression of intended styles.

**Balance accuracy with interaction style.** Participants' comments highlighted that trust was shaped not only by correctness but also by communicative qualities such as clarity, justification, and tone. This suggests that providing brief reasoning, acknowledging uncertainty where appropriate, and supporting user agency can complement factual accuracy.

**Monitor model drift.** Because commercial LLMs evolve rapidly, stance expression and baseline communication behaviors may shift over time. Periodic re-evaluation can help ensure continued alignment between desired personas and actual model behavior.

**Combine multiple metrics.** Perceptions of certainty and trust were influenced by several cues, including verbosity and responsiveness. Evaluations may therefore benefit from pairing subjective user-perception measures with behavioral or task-based metrics such as task success, latency, or error patterns, to capture a complete picture of interaction quality.

Together, these insights highlight how stance cues interact with broader experiential factors, reinforcing the value of multi-dimensional, user-centered evaluation approaches as in UCLLM.

### 6.8 Limitations

Although this work provides useful insights, it is important to acknowledge the limitations. First, the study focused on a single, low-stakes domain, travel planning, which may not generalize to higher-stakes contexts such as medical or legal advice. We chose this domain as a controlled setting for examining certainty communication; future work could extend UCLLM-based evaluations to additional domains. Second, interactions were restricted to five turns, limiting how fully an LLM's epistemic stance could manifest over time. Third, the study operationalized only one UCLLM dimension. Our aim was to demonstrate feasibility and methodological alignment rather than to validate the full framework; evaluating the remaining dimensions is part of planned future work.

We did not control for response length, fluency, or latency across models, all of which can shape users' judgments of certainty and reliability. For example, Gemini produced substantially longer responses (up to 12.7k characters) compared to Claude and DeepSeek (around 5.4k), and interaction durations varied by several minutes across models. Such differences may partially inflate stance-related perceptions—e.g., longer or more elaborated responses may appear more thoughtful or reliable independent of stance. Future work could normalize outputs (e.g., through length-matched prompts, truncation strategies, or latency controls) or model these factors explicitly to better isolate the effects of stance.

Stance prompts may yield slightly different behavioral expressions across models, complicating comparisons. Although we conducted an automated stance-marker analysis to verify the manipulation, future studies could use coder-based annotation and more fine-grained linguistic checks to strengthen manipulation validity.

As a pilot, the study relied on a small and relatively homogeneous sample. A larger and more diverse sample will be needed to strengthen generalizability and to examine how multiple UCLLM dimensions interact across user populations. Finally, we note that future empirical studies are needed to validate and refine the overlaps and boundaries we propose in the framework and Appendix A.

## 7 Conclusion

We introduce *UCLLM*, a framework for evaluating LLMs, and demonstrate how one of its dimensions—Trust & Transparency—can be operationalized through an empirical study of epistemic stance. Comparing four LLMs under high- and low-certainty response styles, we found that lower-certainty framing was often judged as more certain, reliable, and actionable than higher-certainty framing. Model identity also shaped perceptions, with some models eliciting higher certainty ratings, while reliability, actionability, and satisfaction remained similar across systems.

Our findings illustrate the feasibility of applying UCLLM to a single dimension and show how the framework can guide the translation of conceptual constructs into empirical measures. Fully validating UCLLM will require domain-diverse, multi-dimensional studies examining how dimensions interact across tasks, contexts, and users.

By foregrounding user experience and communicative qualities that shape trust, UCLLM offers a foundation for future evaluations aimed at designing more transparent, adaptive, and supportive LLM interactions.

## Acknowledgment of AI Use

GPT-5 was used to paraphrase, reword, and shorten text, improve writing style, and perform grammar and spelling checks to improve the overall text quality.

## References

[1] Yasmeen Abdrabou, Süleyman Özdel, Virmarie Maquiling, Efe Bozkir, and Enkelejda Kasneci. 2025. From Gaze to Data: Privacy and Societal Challenges of Using Eye-tracking Data to Inform GenAI Models. In *Proceedings of the 2025 Symposium on Eye Tracking Research and Applications (ETRA '25)*. Association for Computing Machinery, New York, NY, USA, Article 109, 9 pages. doi:10.1145/3715669.3726788

[2] Hesham Ahmed, Halil Kayaduman, Sonsoles López-Pernas, Markku Tukiainen, and Mohammed Saqr. 2025. User-centric Evaluation of GenAI Alignment and Recommendations based on Predictive Learning Analytics. In *Proceedings of the 2nd Workshop on Generative AI for Learning Analytics (GenAI-LA)*.

[3] Mohammed Alsobay, David M Rothschild, Jake M Hofman, and Daniel G Goldstein. 2025. Bringing Everyone to the Table: An Experimental Study of LLM-Facilitated Group Decision Making. *arXiv preprint arXiv:2508.08242* (2025).

[4] Association for Computing Machinery [n. d.]. *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. Association for Computing Machinery. https://facctconference.org/ Annual conference series on fairness, accountability, and transparency in sociotechnical systems.

[5] Mohammed Bahja, Rawad Hammad, and Gibran Butt. 2020. A User-Centric Framework for Educational Chatbots Design and Development. In *HCI International 2020 - Late Breaking Papers: Multimodality and Intelligence*, Constantine Stephanidis, Masaaki Kurosu, Helmut Degen, and Lauren Reinerman-Jones (Eds.). Springer International Publishing, Cham, 32–43.

[6] Christoph Bartneck. 2023. *Godspeed Questionnaire Series: Translations and Usage*. 1–35. doi:10.1007/978-3-030-89738-3_24-1

[7] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics* 1, 1 (2009), 71–81.

[8] Amanda Baughan, Xuezhi Wang, Ariel Liu, Allison Mercurio, Jilin Chen, and Xiao Ma. 2023. A Mixed-Methods Approach to Understanding User Trust after Voice Assistant Failures. In *Proceedings of the 2023 CHI Conference on Human Factors in*

*Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 7, 16 pages. doi:10.1145/3544548.3581152

[9] Margaret M. Bradley and Peter J. Lang. 1994. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry* 25, 1 (1994), 49–59. doi:10.1016/0005-7916(94)90063-9

[10] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[11] Wanling Cai, Yucheng Jin, and Li Chen. 2022. Impacts of personal characteristics on user trust in conversational recommender systems. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–14.

[12] Marcella Caprario. 2022. Multifunctionality of epistemic stance markers: Variation across disciplines and speaker roles in classroom discourse. *Contrastive Pragmatics* 4, 2 (2022), 213–242.

[13] Marco Cascella, Jonathan Montomoli, Valentina Bellini, and Elena Bignami. 2023. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. *Journal of medical systems* 47, 1 (2023), 33.

[14] Nuo Chen, Quanyu Dai, Xiaoyu Dong, Xiao-Ming Wu, and Zhenhua Dong. 2025. Evaluating conversational recommender systems with large language models: A user-centric evaluation framework. *arXiv e-prints* (2025), arXiv–2501.

[15] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*.

[16] Zhendong Chu, Shen Wang, Jian Xie, Tinghui Zhu, Yibo Yan, Jinheng Ye, Aoxiao Zhong, Xuming Hu, Jing Liang, Philip S Yu, et al. 2025. Llm agents for education: Advances and applications. *arXiv preprint arXiv:2503.11733* (2025).

[17] Nicholas Clark, Hua Shen, Bill Howe, and Tanushree Mitra. 2025. Epistemic Alignment: A Mediating Framework for User-LLM Knowledge Delivery. *arXiv preprint arXiv:2504.01205* (2025).

[18] Kerstin Denecke. 2023. Framework for guiding the development of high-quality conversational agents in healthcare. In *Healthcare*, Vol. 11. MDPI, 1061.

[19] Dhakshenya Ardhithy Dhinagaran, Laura Martinengo, Moon-Ho Ringo Ho, Shafiq Joty, Tobias Kowatsch, Rifat Atun, and Lorainne Tudor Car. 2022. Designing, developing, evaluating, and implementing a smartphone-delivered, rule-based conversational agent (DISCOVER): development of a conceptual framework. *JMIR mHealth and uHealth* 10, 10 (2022), e38740.

[20] Xin Luna Dong, Seungwhan Moon, Yifan Ethan Xu, Kshitiz Malik, and Zhou Yu. 2023. Towards next-generation intelligent assistants leveraging llm techniques. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 5792–5793.

[21] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[22] Aparna Elangovan, Ling Liu, Lei Xu, Sravan Bodapati, and Dan Roth. 2024. Considers-the-human evaluation framework: Rethinking human evaluation for generative large language models. *arXiv preprint arXiv:2405.18638* (2024).

[23] Asbjørn Følstad and Petter Bae Brandtzæg. 2017. Chatbots and the new world of HCI. *Interactions* 24, 4 (June 2017), 38–42. doi:10.1145/3085558

[24] Asbjørn Følstad and Cameron Taylor. 2021. Investigating the user experience of customer service chatbot interaction: a framework for qualitative analysis of chatbot dialogues. *Quality and User Experience* 6, 1 (2021), 6.

[25] Joan Greenbaum and Morten Kyng. 2020. *Design at work: Cooperative design of computer systems*. CRC Press.

[26] Matthew Hall. 2025. The Trust Gate: A Confidence-Based Framework to Prevent AI Hallucinations. doi:10.5281/zenodo.15347057

[27] Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating Large Language Models in Generating Synthetic HCI Research Data: a Case Study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 433, 19 pages. doi:10.1145/3544548.3580688

[28] Yuanzhe Hu, Yu Wang, and Julian McAuley. 2025. Evaluating memory in llm agents via incremental multi-turn interactions. *arXiv preprint arXiv:2507.05257* (2025).

[29] Chen Huang, Peixin Qin, Yang Deng, Wenqiang Lei, Jiancheng Lv, and Tat-Seng Chua. 2024. Concept–An Evaluation Protocol on Conversational Recommender Systems with System-centric and User-centric Factors. *arXiv preprint arXiv:2404.03304* (2024).

[30] Ken Hyland. 2005. Stance and engagement: a model of interaction in academic discourse. *Discourse Studies* 7, 2 (2005), 173–192. arXiv:https://doi.org/10.1177/1461445605050365 doi:10.1177/1461445605050365

[31] Anthony Jameson. 2007. Adaptive interfaces and agents. In *The human-computer interaction handbook*. CRC press, 459–484.

[32] Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2021. A Survey on Conversational Recommender Systems. *ACM Comput. Surv.* 54, 5, Article 105 (May 2021), 36 pages. doi:10.1145/3453154

[33] Yucheng Jin, Li Chen, Wanling Cai, and Xianglin Zhao. 2024. CRS-Que: A User-centric Evaluation Framework for Conversational Recommender Systems. *ACM Trans. Recomm. Syst.* 2, 1, Article 2 (March 2024), 34 pages. doi:10.1145/3631534

[34] Ishika Joshi, Simra Shahid, Shreeya Manasvi Venneti, Manushree Vasu, Yantao Zheng, Yunyao Li, Balaji Krishnamurthy, and Gromit Yeuk-Yin Chan. 2025. CoPrompter: User-Centric Evaluation of LLM Instruction Alignment for Improved Prompt Engineering. In *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25)*. Association for Computing Machinery, New York, NY, USA, 341–365. doi:10.1145/3708359.3712102

[35] Daniel Jurafsky and James H. Martin. 2009. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall.

[36] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences* 103 (2023), 102274.

[37] John Kerr, Anne-Marthe Bles, Sarah Dryhurst, Claudia Schneider, Vivien Chopurian, Alexandra Freeman, and Sander van der Linden. 2023. The effects of communicating uncertainty around statistics, on public trust. *Royal Society Open Science* 10 (11 2023). doi:10.1098/rsos.230604

[38] Sungeun Kim and Dongsuk Oh. 2025. Evaluating Creativity: Can LLMs Be Good Evaluators in Creative Writing Tasks? *Applied Sciences* 15, 6 (2025), 2971.

[39] Thomas Kosch and Sebastian Feger. 2024. Risk or Chance? Large Language Models and Reproducibility in HCI Research. *Interactions* 31, 6 (Oct. 2024), 44–49. doi:10.1145/3695765

[40] Bettina Laugwitz, Theo Held, and Martin Schrepp. 2008. Construction and evaluation of a user experience questionnaire. In *Symposium of the Austrian HCI and usability engineering group*. Springer, 63–76.

[41] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[42] Junlong Li, Fan Zhou, Shichao Sun, Yikai Zhang, Hai Zhao, and Pengfei Liu. 2024. Dissecting human and llm preferences. *arXiv preprint arXiv:2402.11296* (2024).

[43] Lingyu Li, Yixu Wang, Haiquan Zhao, Shuqi Kong, Yan Teng, Chunbo Li, and Yingchun Wang. 2024. Reflection-Bench: Evaluating Epistemic Agency in Large Language Models. *arXiv preprint arXiv:2410.16270* (2024).

[44] Mengqiao Liu, Tevin Wang, Cassandra A. Cohen, Sarah Li, and Chenyan Xiong. 2025. Understand User Opinions of Large Language Models via LLM-Powered In-the-Moment User Experience Interviews. *arXiv preprint arXiv:2502.15226* (2025). https://arxiv.org/abs/2502.15226

[45] Melanie J McGrath, Oliver Lack, James Tisch, and Andreas Duenser. 2025. Measuring trust in artificial intelligence: validation of an established scale and its short form. *Frontiers in Artificial Intelligence* 8 (2025), 1582880.

[46] Lennart Meincke, Ethan R. Mollick, Lilach Mollick, and Dan Shapiro. 2025. *Prompt Engineering is Complicated and Contingent*. Technical Report. Wharton Generative AI Labs. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5165270 Downloaded from Wharton Generative AI Labs website.

[47] Sabrina J Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics* 10 (2022), 857–872.

[48] Justin K Miller and Wenjia Tang. 2025. Evaluating llm metrics through real-world capabilities. *arXiv preprint arXiv:2505.08253* (2025).

[49] Wassem Saad Nsaif, Huda Mohammed Salih, Hassan Hadi Saleh, and Bashar Talib Al-Nuaimi. 2024. Chatbot development: Framework, platform, and assessment metrics. *The Eurasia Proceedings of Science Technology Engineering and Mathematics* 27 (2024), 50–62.

[50] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.

[51] Dijana Peras. 2018. Chatbot evaluation metrics. *Economic and Social Development: Book of Proceedings* (2018), 89–97.

[52] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286* (2022).

[53] J Brian Pickering, Vegard Engen, and Paul Walland. 2017. The interplay between human and machine agency. In *International Conference on Human-Computer Interaction*. Springer, 47–59.

[54] Cheng Qian, Zuxin Liu, Akshara Prabhakar, Zhiwei Liu, Jianguo Zhang, Haolin Chen, Heng Ji, Weiran Yao, Shelby Heinecke, Silvio Savarese, et al. 2025. UserBench: An Interactive Gym Environment for User-Centric Agents. *arXiv preprint arXiv:2507.22034* (2025).

[55] Partha Pratim Ray. 2023. Benchmarking, ethical alignment, and evaluation framework for conversational AI: Advancing responsible development of ChatGPT. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations* 3, 3 (2023), 100136. doi:10.1016/j.tbench.2023.100136

[56] Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor. 2011. *Introduction to recommender systems handbook*. Springer.

[57] Ethan M Rudd, Christopher Andrews, and Philip Tully. 2025. A Practical Guide for Evaluating LLMs and LLM-Reliant Systems. *arXiv preprint arXiv:2506.13023* (2025).

[58] Yasmeen Saleh, Manar Abu Talib, Qassim Nasir, and Fatima Dakalbab. 2025. Evaluating large language models: a systematic review of efficiency, applications, and future directions. *Frontiers in Computer Science* 7 (2025), 1523699.
[59] Pouria Salehi, Yang Ba, Nayoung Kim, Ahmadreza Mosallanezhad, Anna Pan, Myke C Cohen, Yixuan Wang, Jieqiong Zhao, Shawaiz Bhatti, James Sung, et al. 2024. Towards trustworthy AI-enabled decision support systems: Validation of the multisource AI scorecard table (MAST). *Journal of Artificial Intelligence Research* 80 (2024), 1311–1341.
[60] Kathrin Seßler, Ozan Kepir, and Enkelejda Kasneci. 2024. Enhancing Student Motivation Through LLM-Powered Learning Environments: A Comparative Study. In *European Conference on Technology Enhanced Learning*. Springer, 156–162.
[61] Ben Shneiderman. 2020. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction* 36, 6 (2020), 495–504.
[62] Kunmi Sobowale and Daniel Kevin Humphrey. 2025. Evaluating the Quality of Psychotherapy Conversational Agents: Framework Development and Cross-Sectional Study. *JMIR Formative Research* 9 (2025). doi:10.2196/65605
[63] Chunlin Tian, Xinpeng Qin, Kahou Tam, Li Li, Zijian Wang, Yuanzhe Zhao, Minglei Zhang, and Chengzhong Xu. 2025. CLONE: Customizing LLMs for Efficient Latency-Aware Inference at the Edge. *arXiv preprint arXiv:2506.02847* (2025).
[64] Ahmed Tlili, Boulus Shehata, Michael Agyemang Adarkwah, Aras Bozkurt, Daniel T Hickey, Ronghuai Huang, and Brighter Agyemang. 2023. What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart learning environments* 10, 1 (2023), 15.
[65] Roberto Urbani, Caitlin Ferreira, and Joey Lam. 2024. Managerial framework for evaluating AI chatbot integration: Bridging organizational readiness and technological challenges. *Business Horizons* 67, 5 (2024), 595–606. doi:10.1016/j.bushor.2024.05.004 SPECIAL ISSUE: WRITTEN BY CHATGPT.
[66] Marilyn Walker, Diane Litman, Candace A Kamm, and Alicia Abella. 1997. PARADISE: A framework for evaluating spoken dialogue agents. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*. 271–280.
[67] Jiayin Wang, Fengran Mo, Weizhi Ma, Peijie Sun, Min Zhang, and Jian-Yun Nie. 2024. A user-centric multi-intent benchmark for evaluating large language models. *arXiv preprint arXiv:2404.13940* (2024).
[68] Anna Wolters, Arnold F Arz von Straussenburg, and Dennis M Riehle. 2024. Evaluation framework for large language model-based conversational agents. (2024).
[69] World Wide Web Consortium (W3C). 2023. Web Content Accessibility Guidelines (WCAG) 2.2. https://www.w3.org/TR/WCAG22/. W3C Recommendation.
[70] Jochen Wulf and Jürg Meierhofer. 2024. Exploring the potential of large language models for automation in technical customer service. *arXiv preprint arXiv:2405.09161* (2024).
[71] Yuzhe Yang, Yifei Zhang, Yan Hu, Yilin Guo, Ruoli Gan, Yueru He, Mingcong Lei, Xiao Zhang, Haining Wang, Qianqian Xie, Jimin Huang, Honghai Yu, and Benyou Wang. 2025. UCFE: A User-Centric Financial Expertise Benchmark for Large Language Models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, Luis Chiruzzo, Alan Ritter, and Lu Wang (Eds.). Association for Computational Linguistics, Albuquerque, New Mexico, 5429–5448. doi:10.18653/v1/2025.findings-naacl.300
[72] Zhengqing Yuan, Weixiang Sun, Yixin Liu, Huichi Zhou, Rong Zhou, Yiyang Li, Zheyuan Zhang, Wei Song, Yue Huang, Haolong Jia, et al. 2025. EfficientLLM: Efficiency in Large Language Models. *arXiv preprint arXiv:2505.13840* (2025).
[73] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems* 36 (2023), 46595–46623.
[74] Mingqian Zheng, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. 2024. When" a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 15126–15154.
[75] Kaitlyn Zhou, Jena D Hwang, Xiang Ren, Nouha Dziri, Dan Jurafsky, and Maarten Sap. 2024. Rel-AI: An Interaction-Centered Approach To Measuring Human-LM Reliance. *arXiv preprint arXiv:2407.07950* (2024).
[76] Li Zhui, Nina Yhap, Liu Liping, Wang Zhengjie, Xiong Zhonghao, Yuan Xiaoshu, Cui Hong, Liu Xuexiu, and Ren Wei. 2024. Impact of large language models on medical education and teaching adaptations. *JMIR Medical Informatics* 12, 1 (2024), e55933.

# A Appendix

## A.1 Boundaries Between UCLLM Dimensions

To clarify how the UCLLM dimensions relate to one another, this section highlights key boundaries between them and explains why each pair should remain conceptually distinct.

### *A.1.1 Task Success (TS).*

- **TS-CL.** Task Success evaluates whether the user achieves their goal, where Cognitive Load reflects the mental effort required. Easy-to-comprehend responses reduce load and indirectly support success, but a task may still succeed even under high cognitive effort; the two should therefore be counted independently.
- **TS-VAL.** Value reflects the user's subjective sense of benefit, which increases when tasks are completed efficiently. However, a highly valuable suggestion may occur even when no task is involved; thus Value captures perceived utility, not functional accomplishment.
- **TS-DF.** Smooth Dialogue Flow supports successful completion by maintaining continuity, but flow is about interaction dynamics, not outcomes. Disrupted flow may still lead to a correct final answer.
- **TS-REC.** Error Recovery prevents task completion when the model makes errors, but REC measures recovery behavior, not task outcome. A task may succeed despite poor recovery if the user compensates manually for it.
- **TS-HALL.** Hallucinations compromise Task Success on factual tasks, but they remain distinct: hallucination concerns factual correctness, whereas success concerns whether the user's goal is achieved. A user might still complete a simple task despite errors, and vice versa.
- **TS-INSTR.** Instructional Quality influences a user's ability to complete tasks requiring procedural guidance, but INSTR measures pedagogical clarity rather than task completion itself.
- **TS-AGCY.** Model Agency can help or hinder task success depending on initiative alignment with user goals. Agency captures the model's proactiveness, while TS focuses solely on whether the task is achieved.

### *A.1.2 Cognitive Load (CL).*

- **CL-VAL.** Users perceive greater Value when effort is minimized, but CL captures mental burden and VAL captures subjective benefit; high value can still occur with substantial effort.
- **CL-DF.** Breaks in flow increase cognitive effort; however, DF reflects interaction smoothness rather than mental workload.
- **CL-CTRL.** A lack of Control raises cognitive demand as users must compensate manually, but CTRL measures the availability of controlling mechanisms rather than mental effort itself.
- **CL-REC.** Recovery adds effort, though CL and REC differ: REC evaluates how the model responds to errors, while CL concerns the user's mental demands.
- **CL-EMO.** High cognitive effort often produces frustration; however, emotional response (EMO) reflects affect, not workload.

- **CL-LEARN.** Load slows learning, but LEARN focuses on the user's ability to understand the model's behavior over time. A dense response may raise cognitive load without affecting learnability at all.
- **CL-EXPL.** Good explanations reduce cognitive load by clarifying reasoning, but EXPL reflects explanation quality rather than effort.
- **CL-INSTR.** Poor instruction raises load, but INSTR measures pedagogical clarity independent of cognitive effort.
- **CL-AGCY.** Model Agency can increase or decrease effort depending on appropriateness; however, agency captures initiative rather than workload.

#### *A.1.3 Value (VAL).*

- **VAL-EMO.** Emotional tone influences perceived usefulness, but EMO concerns affective experience while VAL concerns perceived utility.
- **VAL-CREAT.** Creative or novel responses can feel more valuable, but CREAT measures generative innovation rather than perceived benefit.
- **VAL-PERS.** Personalization increases relevance, enhancing value, but PERS concerns tailoring mechanisms independent of user evaluation.
- **VAL-AGCY.** Helpful initiative raises value, but AGCY reflects proactiveness irrespective of subjective usefulness.

#### *A.1.4 Persona Fit (PF).*

- **PF-DF.** A consistent persona improves conversational flow; on the other side, DF concerns turn coherence rather than persona alignment.
- **PF-EMO.** Persona tone shapes emotional reaction, but PF concerns tone matching while EMO concerns affective state.
- **PF-PERS.** Tailored tone improves persona alignment; however, PERS deals with customization beyond persona.
- **PF-INCL.** Inclusive language supports persona fit, although INCL measures equity and fairness rather than tone consistency.

#### *A.1.5 Dialogue Flow (DF).*

- **DF-REC.** Recovery depends on restoring flow, but REC assesses correction behavior, not flow smoothness.
- **DF-EMO.** Flow disruptions frustrate users; however, EMO captures emotional state rather than conversational structure.
- **DF-LEARN.** Predictable flow helps learning, but LEARN captures understanding of system behavior rather than flow quality.
- **DF-AGCY.** Agency affects pacing, but AGCY measures initiative rather than coherence.

#### *A.1.6 Control (CTRL).*

- **CTRL-PERS.** User-controlled preferences influence personalization, but CTRL refers to steering ability rather than tailored output.
- **CTRL-MEM.** Memory enhances control by stabilizing the model's behavior, however, MEM concerns persistent information rather than user control.
- **CTRL-AGCY.** Agency may conflict with user control, but AGCY reflects model-driven action rather than user-driven.

#### *A.1.7 Error Recovery (REC).*

- **REC-TS.** Recovery helps keep task progress but remains distinct from success, which depends on final outcomes.
- **REC-CL.** Correction increases workload but REC evaluates error-handling behavior rather than cognitive overload.
- **REC-DF.** Recovery influences flow; however, DF captures continuity, independent of recovery.

#### *A.1.8 Emotion (EMO).*

- **EMO-SAFE.** Unsafe content causes negative emotion, but SAFE is not about how users feel. Safety is about the risk, harm, or danger in the content, even if a user doesn't feel bad.
- **EMO-INCL.** Inclusive language supports positive emotion, however INCL concerns equity rather than affective response.
- **EMO-AGCY.** Initiative influences emotion, but AGCY evaluates proactive behavior rather than emotional impact.

#### *A.1.9 Learnability (LEARN).*

- **LEARN-INSTR.** Instruction shapes learning but INSTR evaluates pedagogical clarity, not long-term understanding.
- **LEARN-PERS.** Personalized difficulty improves learnability, but PERS concerns customization mechanisms.

#### *A.1.10 Confidence (CONF).*

- **CONF-TRUST.** Confidence cues help calibrate trust, yet CONF assesses expressed certainty while TRUST assesses user belief in reliability.

#### *A.1.11 Trust (TRUST).*

- **TRUST-EXPL.** Explanations strengthen trust, but EXPL concerns reasoning quality independent of relational judgment.
- **TRUST-HALL.** Factual errors reduce trust, however, HALL assesses correctness itself, not the trust outcome.
- **TRUST-SAFE.** Unsafe outputs affect trust, though SAFE concerns harm potential independent of user belief.

#### *A.1.12 Explainability (EXPL).*

- **EXPL-INSTR.** Explanations support instruction, yet EXPL concerns the rationale quality rather than pedagogical structure.

#### *A.1.13 Hallucination (HALL).*

- **HALL-CONS.** A hallucination remains a hallucination even when consistent; CONS measures stability, not correctness.
- **HALL-SAFE.** Harmful incorrect content is unsafe, though SAFE captures risk beyond factual quality.

#### *A.1.14 Creativity (CREAT).*

- **CREAT-AGCY.** Creative initiative influences interaction tone, but CREAT concerns generative novelty while AGCY concerns proactiveness.

#### *A.1.15 Instructional Quality (INSTR).*

- **INSTR-TS.** Instructional Quality can improve or hinder task completion, but INSTR focuses on how clearly the model

teaches or guides, whereas TS focuses on whether the user actually achieves their goal.
- **INSTR-CL.** Poorly structured instruction raises cognitive effort, but INSTR concerns pedagogical design while CL concerns mental demand.
- **INSTR-LEARN.** Instruction shapes how much users learn, yet INSTR evaluates teaching quality while LEARN evaluates users' resulting proficiency.
- **INSTR-EXPL.** Explanations contribute to good instruction, but EXPL focuses on rationale clarity while INSTR focuses on the overall teaching strategy.
- **INSTR-HALL.** Incorrect teaching content is a hallucination, however, INSTR concerns pedagogical form while HALL concerns factual grounding.

*A.1.16 Personalization (PERS).*

- **PERS-AGCY.** Adaptive initiative is a form of personalized behavior, yet PERS focuses on tailoring while AGCY emphasizes initiative.

*A.1.17 Consistency (CONS).*

- **CONS-HALL.** A hallucination remains a hallucination even when consistent; CONS measures stability, not correctness.

*A.1.18 Memory & Privacy (MEM).*

- **MEM-CTRL.** Memory enhances control by stabilizing the model's behavior over time, yet MEM concerns what is retained while CTRL concerns how easily users can steer the system.

*A.1.19 Safety (SAFE).*

- **SAFE-HALL.** Harmful inaccuracies are unsafe, yet HALL captures factual invention while SAFE concerns harm even in correct statements.

*A.1.20 Inclusion (INCL).*

- **INCL-PF.** Inclusive tone supports persona fit; PF captures stylistic persona shape while INCL addresses fairness.

*A.1.21 Agency (AGCY).*

- **AGCY-TS.** Model initiative can help or hinder whether tasks are completed; however, AGCY measures proactiveness while TS measures goal completion.
- **AGCY-CL.** Initiative may simplify or complicate interactions, but AGCY concerns who drives the interaction while CL concerns the mental effort required.
- **AGCY-VAL.** Proactive suggestions can increase perceived value, yet AGCY captures initiative while VAL captures subjective benefit.
- **AGCY-DF.** Initiative shapes conversational pacing and turn-taking, but AGCY concerns proactivity while DF concerns coherence and flow.
- **AGCY-CTRL.** High agency can conflict with user control, yet AGCY reflects model-driven behavior while CTRL reflects user-directed steering.
- **AGCY-EMO.** Initiative can feel supportive or overwhelming, yet AGCY concerns proactiveness while EMO concerns users' emotional reactions.
- **AGCY-CREAT.** Creative initiative changes the style of proactive behavior, but AGCY measures initiative while CREAT measures originality.
- **AGCY-PERS.** Adaptive initiative can be personalized, yet AGCY focuses on self-directed behavior while PERS focuses on tailoring to the user.