Analyzing the Presentation, Content, and Utilization of References in LLM-powered Conversational AI Systems

Jianheng Ouyang
The Hong Kong University of Science and Technology
Hong Kong, China
jouyangae@connect.ust.hk

Arpit Narechania
The Hong Kong University of Science and Technology
Hong Kong, China
arpit@ust.hk

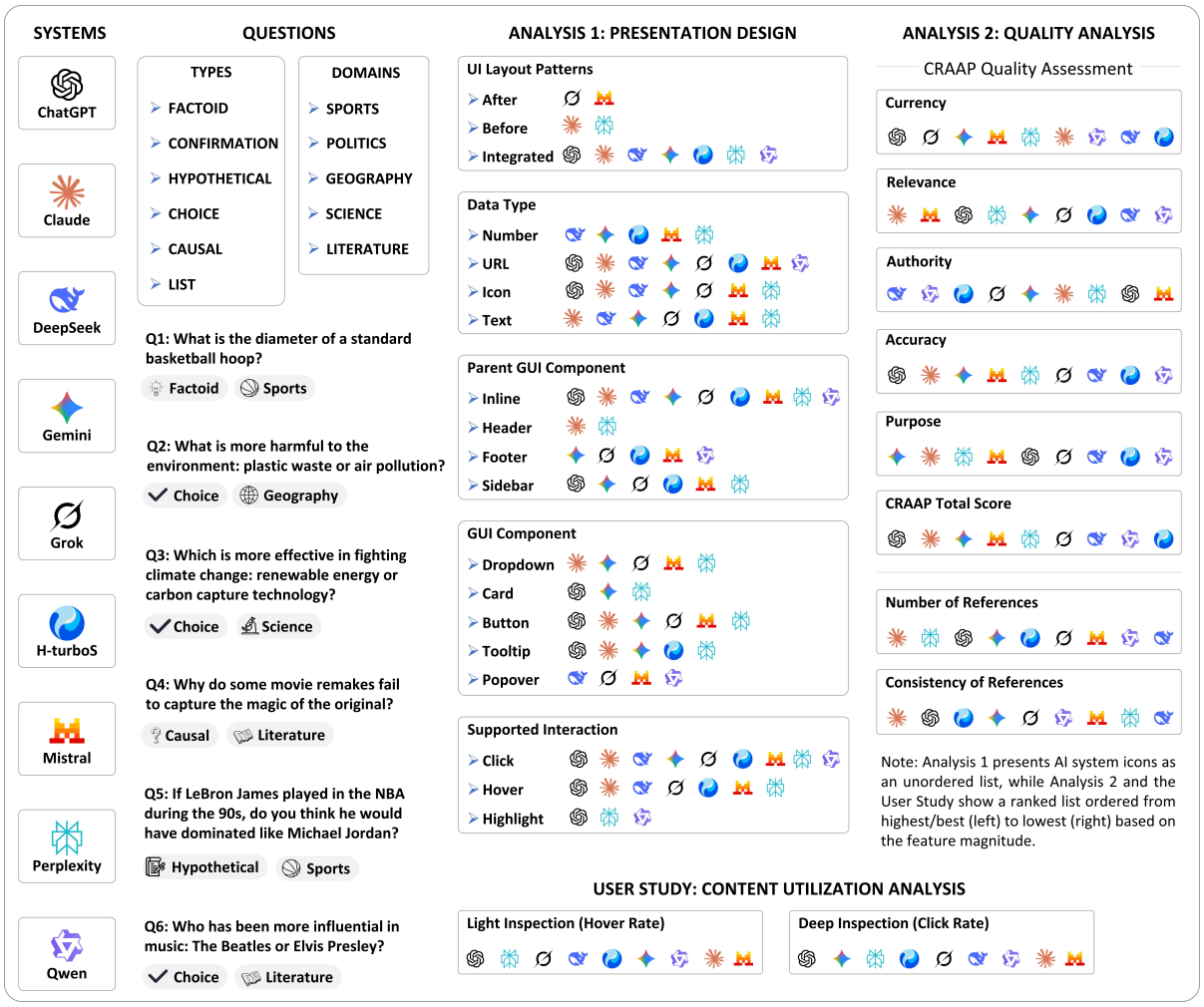

Figure 1: Summary of findings from three analyses of references across nine conversational AI systems. Analysis 1: Presentation Design illustrates how and where references are presented in the UI. Analysis 2: Quality Analysis evaluates the credibility of references using CRAAP criteria, alongside quantity and consistency metrics. Content Utilization Analysis reports results from a preliminary user study measuring participants' engagement with references. Together, these analyses reveal notable disparities in how references are presented, their reliability, and how users interact with them.

CHI EA '26, Barcelona, Spain

ACM ISBN 979-8-4007-2281-3/26/04
https://doi.org/10.1145/3772363.3798415

Abstract

As conversational AI systems become popular for information retrieval and question-answering, the references they cite are key to ensuring their answers are reliable and trustworthy. Yet, no prior work systematically analyzes how these references are presented or their quality. We examine 1,517 references from 30 question–answer pairs across nine systems, focusing on their (1) presentation in the user interface and (2) quality using the CRAAP criteria. We find notable variations in the presentation, quality, and quantity of references across systems. For instance, ChatGPT provides more references (9.5 per response on average) with higher quality (15.48/20 CRAAP score), while Hunyuan-TurboS provides fewer references (4.0) and lower quality (11.65/20). Additionally, a preliminary user study shows that people rarely interact with these references and that their behavior differs across systems. These findings highlight the need for better interface designs that help users engage with and trust references more effectively.

CCS Concepts

• **Human-centered computing → Human computer interaction (HCI)**; • **Computing methodologies → Artificial intelligence**.

Keywords

Large language model, Question-answering, Conversational AI systems, Content analysis, References, User interface, User experience, Design

ACM Reference Format:

Jianheng Ouyang and Arpit Narechania. 2026. Analyzing the Presentation, Content, and Utilization of References in LLM-powered Conversational AI Systems. In *Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems (CHI EA '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3772363.3798415

1 Introduction and Background

Search engines such as Google[1] and Bing[2] have long supported information retrieval on the web [17]. They present relevant hyperlinks, leaving users to find answers by themselves. In contrast, and more recently, conversational AI systems such as ChatGPT[3] and Gemini[4] facilitate question-answering by directly synthesizing answers, saving users significant effort in finding information manually [6, 46]. Today, billions of users worldwide rely on these systems [9]: everyday users turn to them for general or specific information lookup [49] and decision-making [22]; students use them to learn new concepts [51]; developers apply them for coding assistance [4]; and researchers employ them to find scientific literature [3]. A key distinction between search engines and conversational AI systems lies in the role of references [31]. In search engines, hyperlinks act as implicit references, delegating trust to the original websites. Conversely, conversational AI systems synthesize answers and select specific references to support them. Because AI-generated answers can sometimes be incorrect or incomplete [43, 45], it becomes important for users to consult these references to verify the credibility of the answers and gain further knowledge [24, 27]. Consequently, references play a more central role in shaping user trust and perceptions of credibility in conversational AI systems compared to traditional search engines [8, 13, 14, 20].

Notably, like the AI-generated answers, AI-generated references can also sometimes be flawed, i.e., they can be inaccurate, biased, irrelevant, fabricated, or poorly formatted [2, 10, 18, 47]. For example, LLMs have been shown to exhibit gender bias by disproportionately referencing articles authored by male writers [18]. Beyond accuracy, other characteristics also matter. For example, reference quantity (number of references) can influence perceived thoroughness; consistency (the reliability of sources across repeated queries) can affect trust and reproducibility [29]. These issues highlight the need to systematically analyze how AI systems provide references. We therefore ask: *"What is the quality of references provided by conversational AI systems?"*

Next, like the content of references, their presentation in the user interface (UI) is equally consequential. For example, thoughtful interaction design has been shown to reduce cognitive load and improve user comprehension, engagement, and trust in digital learning [16, 38, 39, 50]. We aim to characterize the presentation and interaction for LLM-powered conversational AI systems. So we ask: *"How do different conversational AI systems present references in the user interface for users to interact with?"*.

Lastly, while important, inspecting interface designs alone cannot reveal real-world behavior. Only through empirical observation can we understand how users actually perceive and utilize these references. Therefore, we ask: *"How do users interact with references across different conversational AI systems?"*

Together, these three research questions span how references are generated, how they are surfaced in the interface, and how they are ultimately taken up (or ignored) by users, providing a holistic view of references in conversational AI systems.

To answer these questions, we conducted two system analyses and one content utilization user study across nine conversational AI systems [48]. We first examined the UI design choices associated with references, focusing on when they are included, their visual styles, placement, formatting, and interaction mechanisms. Next, we evaluated the quality of references using the CRAAP framework (Currency, Relevance, Authority, Accuracy, and Purpose) [5], while also measuring the quantity and consistency of references provided by each system. Lastly, we conducted a preliminary user study to understand how users interact with references across systems.

Our findings reveal notable differences across the nine systems, all summarized in Figure 1. ChatGPT and Claude achieved the highest overall quality, with CRAAP scores of 15.48/20, whereas Hunyuan-TurboS obtained the lowest score (11.65/20). Presentation strategies also varied, ranging from inline links to numbered footnotes and icons. Systems also exhibited divergent patterns in reference number and consistency: Claude maintained high and stable reference counts, whereas DeepSeek provided fewer references that fluctuated across repeated queries. Furthermore, our user study reveals that actual interaction–specifically hovering and clicking–remained generally low across all evaluated systems.

[1] https://google.com
[2] https://bing.com
[3] https://chatgpt.com/
[4] https://gemini.google.com/

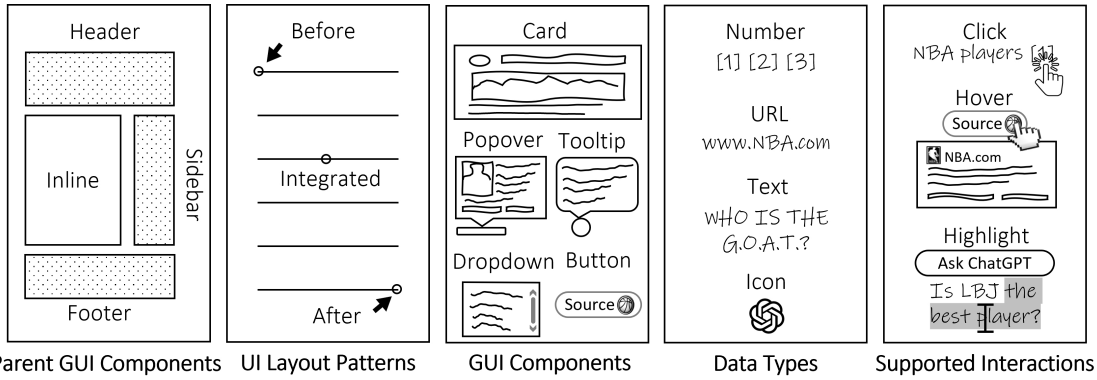

Figure 2: Five design dimensions for presenting references in conversational AI systems: (1) Parent GUI Components (header, inline, sidebar, footer), (2) Data Types (number, URL, text, icon), (3) UI Layout Patterns (before, integrated, after), (4) GUI Components (card, popover, tooltip, dropdown, button), and (5) Supported Interactions (click, hover, highlight). Each dimension represents a design decision that affects how users discover, access, and engage with references.

Collectively, these analyses provide insights into the quality and presentation of references in conversational AI systems. We posit that these differences may have implications for user trust and engagement with these systems [40]. Consequently, we call on the HCI community to study which UI layouts and interaction designs best support users, including the design of adaptive UIs that may tailor to specific or evolving user needs [36]. For example, systems can offer adaptive designs, such as inline references for students and concise answers for experts. Developers can also use CRAAP scores to ensure high-quality reference outputs. In doing so, we can collectively pave the way toward genuinely 'user-centric AI' [35]. The primary contributions of this work include:

(1) Findings from an analysis about when and how LLM-powered conversational AI systems present references for users to view and interact with.
(2) Findings from an analysis about the quantity, consistency, and quality of the references output in LLM-powered conversational AI systems.
(3) Findings from a user study about the engagement levels and interaction patterns of users with references in LLM-powered conversational AI systems.

2 System Analyses

2.1 Methodology

First, we selected nine conversational AI systems based on market visibility and user adoption [48]: *ChatGPT-4o, Claude 3.5 Sonnet, DeepSeek-V3, Gemini 2.5, Grok 3, Hunyuan-TurboS, Mistral Medium3, Perplexity AI*, and *Qwen3-235B-A22B-2507*. Next, to study AI systems' outputs on different types of questions, we curated a query set covering two dimensions: (1) six question types based on the taxonomy proposed by Mohasseb et al. [32]–*factoid, confirmation, hypothetical, choice, causal, and list* and (2) five domains–*sports, politics, geography, science, and literature* (Figure 1). Subsequently, we prompted each of the 9 systems once per query, extracting references and metadata through direct UI interaction[5]. This initial round yielded 270 system responses (30 questions × 9 systems), containing 1,517 unique references used for our primary UI and quality analyses. Lastly, to evaluate reference consistency, we used independent-session repeats: for each question, we opened a fresh chat, asked the query, and recorded the number of references. In total, we conducted 1,350 trials (30 questions × 9 systems × 5 iterations) to examine how reference counts fluctuate across repeated conversations. We evaluated the internal consistency of these reference counts across iterations for each system using Cronbach's α [12].

Following extraction, one author analyzed the data to perform two complementary analyses about: (1) *Presentation Design* to identify patterns in layout, interactivity, and data type; and (2) *Content Analysis* to evaluate reference credibility and quality. For the former, one author visually inspected the UI screenshots for each system response, and developed a codebook of reference attributes–such as UI layout and data types–which was finalized in consultation with a second author, resolving any discrepancies via consensus. Similarly, for the latter, one author employed the CRAAP framework [5]–a widely-used standard for evaluating source credibility–assessing each of its five dimensions on a 4-point scale (1 to 4, as illustrated in Figure 3): Currency (timeliness), Relevance (topic alignment), Authority (source expertise), Accuracy (factual correctness), and Purpose (objectivity)–finalized via consensus with the second author. Although CRAAP is designed for evaluating human-authored sources that actually exist [5], we adapted it for AI-generated references that include hallucinated or inaccessible references (e.g., 404 errors). We assigned such references the minimum score (1/5) for Accuracy and Relevance–the dimensions directly undermined by non-existence–while Currency, Authority, and Purpose were assessed from available metadata. We verified existence by accessing each URL and cross-checking details against known publications.

[5]All data was collected between July 1-7, 2025.

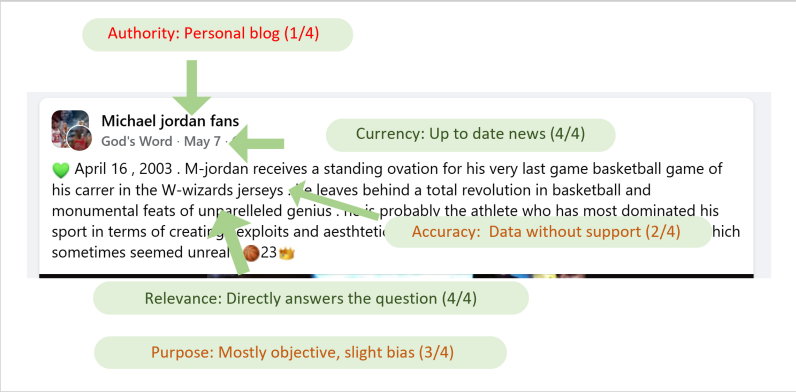

Figure 3: Quality assessment example using the CRAAP test. A social media post is evaluated across five criteria: Authority (1/4): a personal blog lacking formal credentials. Currency (4/4): up-to-date information. Accuracy (2/4): claims lack sufficient supporting evidence. Relevance (4/4): directly addresses the subject. Purpose (3/4): largely objective but shows slight bias.

2.2 Findings: Presentation Design

We identified five design dimensions to characterize how systems present references. Figure 2 summarizes them, and Figure 1 shows how systems align with each. Each dimension is described below.

- **Parent GUI Components**: This dimension describes the placement of references within the interface. All nine systems used *Inline* references within the text. To provide additional details without clutter, several systems (e.g., Grok, Gemini) also employed *Sidebar* (66.7%, $N = 6$) or *Footer* (55.6%, $N = 5$) placements. Only two systems positioned references in a *Header* (22.2%, $N = 2$).
- **UI Layout Patterns**: This dimension characterizes the arrangement of references relative to the main content. Most systems (77.8%, $N = 7$), including ChatGPT, adopted an *Integrated* layout where references are embedded within responses. Two systems placed references *Before* the answer (22.2%, $N = 2$), as observed in Claude, while another two positioned them *After* the answer (22.2%, $N = 2$), as in Grok.
- **GUI Components**: This dimension concerns the interactive elements used to display references. Most systems implemented *Button* components (66.7%, $N = 6$) to facilitate interaction with references. *Dropdown* (55.6%, $N = 5$) and *Tooltip* (55.6%, $N = 5$) were also common for revealing additional details, while *Popover* (44.4%, $N = 4$) was used by some systems (e.g., Grok, Qwen) for richer previews. Only three systems employed *Card* components (33.3%, $N = 3$).
- **Data Types**: This dimension identifies the kinds of information shown for each reference. Most systems, including ChatGPT and Gemini, prioritized clickable *URL* links (88.9%, $N = 8$) for direct access to references. *Icon* (77.8%, $N = 7$) and *Text* (77.8%, $N = 7$) representations were also prevalent to balance recognition and description. Five systems (e.g., DeepSeek) used *Number* indicators (55.6%, $N = 5$) as compact inline markers.
- **Supported Interactions**: This dimension specifies the user interactions supported for referencing. All systems supported basic *Click* interactions for navigation. Most (77.8%, $N = 7$) also supported *Hover* actions for previews. Only three systems (33.3%, $N = 3$), such as Qwen, enabled *Highlight* interactions for operations like citing or quoting selected text.

These findings reveal diverse design strategies across the nine systems, reflecting distinct approaches to balancing reference visibility with reading flow. Some prioritize integrated inline references that enhance discoverability but risk cluttering the main content, while others favor sidebar or footer placements that preserve response readability at the cost of reduced prominence. Ultimately, no single strategy is optimal, highlighting the need for adaptive designs that dynamically adjust to user and task demands.

2.3 Findings: Quality Analysis

We assessed the reference outputs of the nine selected systems using the CRAAP framework (Currency, Relevance, Authority, Accuracy, and Purpose). We found notable differences in reference quality across systems, summarized in Figure 4, and described below:

- **Currency**: ChatGPT scored highest (3.39 ± 1.10), followed by Grok (3.25 ± 1.17) and Gemini (3.19 ± 1.25). Hunyuan-TurboS had the lowest score (1.76 ± 1.29), indicating inconsistency in the timeliness of references.
- **Relevance**: Claude achieved the highest relevance score (3.52 ± 0.60), followed by Mistral Medium3 (3.48 ± 0.67) and ChatGPT (3.43 ± 0.68). Qwen had the lowest score (2.03 ± 1.05), suggesting weaker alignment between content and question.
- **Authority**: DeepSeek obtained the highest authority score (3.27 ± 0.83), followed by Qwen (3.22 ± 0.80) and Hunyuan-TurboS (3.02 ± 0.97). Mistral Medium3 scored lowest (2.36 ± 1.03), reflecting reliance on less credible sources.
- **Accuracy**: ChatGPT achieved the highest accuracy (3.25 ± 0.69), followed by Claude (3.23 ± 0.55) and Gemini (3.19 ± 0.63). Qwen had the lowest accuracy (2.27 ± 1.18), indicating a higher incidence of factual errors.
- **Purpose**: Claude and Gemini scored highest (3.12 ± 0.48 and 3.12 ± 0.55), followed by Perplexity (3.07 ± 0.59). Qwen had the lowest score (2.21 ± 1.14), indicating less objective information presentation.

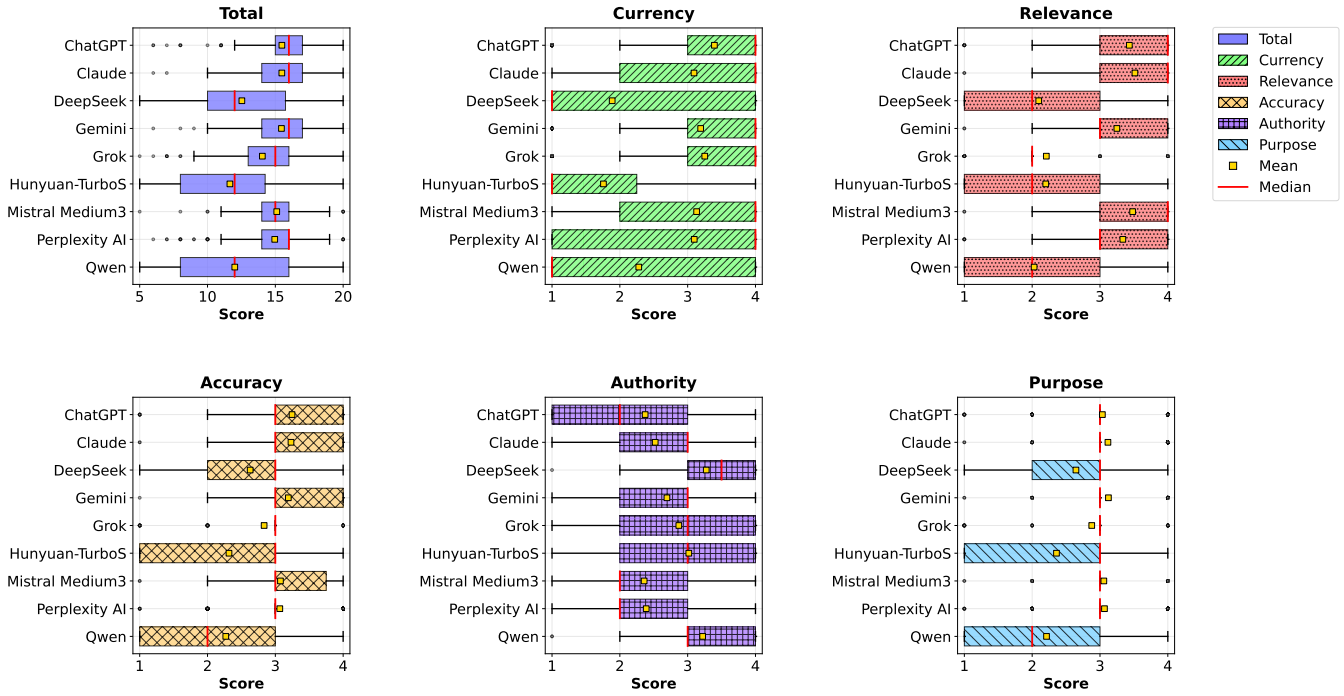

Figure 4: Quality of references from nine conversational AI systems across the CRAAP Framework. This figure presents six horizontal box plots, visualizing the score distribution for each conversational AI system on the CRAAP criteria (Currency, Relevance, Authority, Accuracy, Purpose) and a Total score. Each box plot shows the median (red line), mean (yellow square), interquartile range (box), and outliers (black dots). For metrics where the box appears as a single line, it indicates that the 25th and 75th percentiles coincide, meaning a high consistency in scores for the middle 50% of observations. Observations show that Western models generally outperform others, with ChatGPT and Claude showing high consistency across all criteria.

In addition to the CRAAP **quality** dimensions, we also evaluated the **quantity** and **consistency** of references. Quantity captures the average number of citations a system produces per response, while consistency reflects the stability of that count across multiple iterations. Consistency was assessed using *Cronbach's* α, a reliability coefficient ranging from 0 to 1 that indicates the internal stability of repeated outcomes [1, 42]. Values above 0.70 are generally considered acceptable.

We found that Claude produced exactly 10 references in every iteration, demonstrating perfect mathematical consistency (α = 1.000). However, this lack of variation likely reflects a rigid, hard-coded constraint rather than adaptable behavior. In comparison, ChatGPT generated an average of 9.5 references with excellent consistency (α = 0.961), suggesting stable yet dynamic reference generation. Interestingly, while Hunyuan-TurboS provided notably fewer references (4.0 on average), it maintained a similarly high level of consistency (α = 0.960). Conversely, DeepSeek provided the fewest references (2.1 on average) and exhibited the lowest consistency (α = 0.811), indicating significant fluctuations across runs. Overall, there were notable differences in terms of quality, quantity, and consistency of referenecs across systems. ChatGPT emerged as the most robust performer, offering high CRAAP scores alongside stable and numerous reference counts. Claude also performed highly but exhibited some rigidity in the quantity of references. In contrast, DeepSeek and Qwen produced fewer, lower-quality references with greater variability.

3 Content Utilization Analysis

Having analyzed how systems present references in the UI and their actual quality, we next investigated how users interact with these references during a series of information-seeking tasks.

3.1 Methodology

We recruited 12 undergraduate students (8 male, 4 female) via university email. Participants were asked to use each of the nine AI systems to answer a curated set of questions spanning across multiple domains and question types (as illustrated in Figure 1). For example, participants were asked specific questions such as, *"Why do leaves change color in the fall across different tree species and regions?"* and *"Which is more effective in fighting climate change: renewable energy or carbon capture technology?"* Note that the order of questions was randomized to mitigate presentation bias.

For each question-answer pair, we instructed participants to study the latter and evaluate its accuracy. We provided no instructions about the presence or position of references; we also did not ask participants to ask follow-up questions to fetch missing references, ensuring that all engagement–including the decision to forgo verification–was natural and entirely self-directed. To help

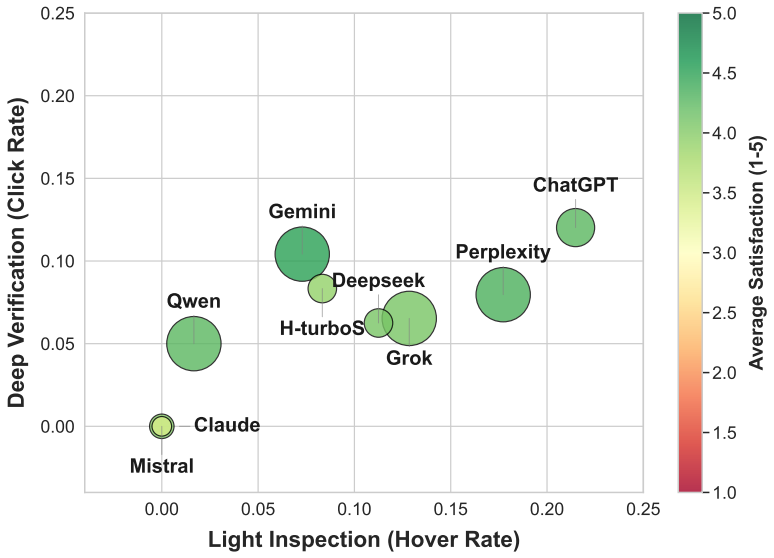

Figure 5: User Interaction Behavior. The bubble chart shows Light Inspection (Hover Rate) against Deep Verification (Click Rate) across nine AI systems. Bubble size represents reference counts, while color indicates average user satisfaction (1-5).

us analyze user behavior, we recorded hover and click frequencies (proxies for visual attention and active verification), reference counts, and 5-point Likert satisfaction scores. Detailed questions and findings are available in supplemental material.

3.2 Findings

Figure 5 summarizes the user study results. Verification rates were consistently low across systems, with hover and click frequencies remaining below 25% for all platforms. ChatGPT achieved the highest engagement (~22% hover, ~12% click). Hover interactions consistently exceeded clicks, suggesting a stronger preference for quick tooltips or popovers over direct source access [19, 26].

Despite limited verification activity, satisfaction scores remained high (mostly >4.0; Figure 5), revealing a potential *trust gap* in which users accepted AI-generated content without validation [25, 34, 37]. Participants often agreed with the system's answer without viewing references to confirm its claims. This trend persisted even for systems with lower reference accuracy (e.g., Qwen; Section 2.3), indicating that plausibility often outweighed verification.

Overall, these results suggest that users relied more on perceived plausibility than on explicit source checking [7, 15, 30]. This behavior reflects a tendency to treat conversational AIs as authoritative experts rather than informational aids, risking over-reliance [41, 44]–though this requires further study.

4 Limitations

Our content analysis and user study had many limitations; first, given the rapid evolution of conversational AI systems, our findings represent a snapshot of the landscape as of July 2025; future model updates may cause specific reference behaviors to shift [11]. Next, while our methodology penalized hallucinated references with the lowest *Accuracy* and *Relevance* scores, we recognize that conversational AI systems inherently produce fabricated content. Traditional frameworks like CRAAP are designed with the assumption that the evaluated sources actually exist. Therefore, because relying on frameworks built for human-authored content is not a long-term solution, we emphasize the critical need to establish specialized evaluation metrics and rules specifically designed for references provided by AI systems.

Next, although our investigation relies on a fixed and focused set of questions, tools, and participants, our findings provide preliminary evidence that the disparity between reference presentation and quality is a critical issue and establishing this as an important area for future large-scale investigation. Additionally, interaction behavior is currently modeled solely based on mouse cursor inputs, which may not be a complete proxy for attention [21]. Future work may consider integrating eye-tracking technology to more accurately characterize user attention [33]. Future research should also explore how specific UI interventions, such as credibility warning labels [23] or interactive visualization of source provenance [28], can encourage users toward more active verification behaviors.

5 Conclusion

We analyzed 1,517 references across nine LLM-based conversational AI systems to characterize the presentation, quality, quantity, and consistency of references in the user interface. We found notable differences between systems across all metrics; however, overall, current designs often relegate references to static decorations rather than first-class interactive elements necessary for active verification. These results advocate for adaptive UIs [36] that dynamically tailor reference granularity to user needs and expertise.

References

[1] Katie Adamson and Susan Prion. 2013. Reliability: Measuring Internal Consistency Using Cronbach's α. *Clinical Simulation in Nursing* 9 (05 2013), e179–e180.

doi:10.1016/j.ecns.2012.12.001

[2] Andres Algaba, Carmen Mazijn, Vincent Holst, Floriano Tori, Sylvia Wenmackers, and Vincent Ginis. 2024. Large language models reflect human citation patterns with a heightened citation bias. *arXiv preprint arXiv:2405.15739* (2024). doi:10.48550/arXiv.2405.15739

[3] Hongye An, Arpit Narechania, Emily Wall, and Kai Xu. 2024. vitaLITy 2: Reviewing Academic Literature Using Large Language Models. Presented at the NLVIZ Workshop, IEEE VIS 2024. doi:10.48550/arXiv.2408.13450 arXiv:2408.13450 [cs.HC]

[4] Gal Bakal, Ali Dasdan, Yaniv Katz, Michael Kaufman, and Guy Levin. 2025. Experience with GitHub Copilot for Developer Productivity at Zoominfo. *arXiv preprint arXiv:2501.13282* (2025). doi:10.48550/arXiv.2501.13282

[5] Sarah Blakeslee. 2004. The CRAAP test. *Loex Quarterly* 31, 3 (2004), 4.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[7] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 188 (April 2021), 21 pages. doi:10.1145/3449287

[8] 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* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 489, 14 pages. doi:10.1145/3491102.3517471

[9] Shawn Carolan, Amy Wu Martin, C.C. Gong, and Sam Borja. 2025. 2025: The State of Consumer AI. https://menlovc.com/perspective/2025-the-state-of-consumer-ai/.

[10] Cheng Chen and S. Shyam Sundar. 2023. Is this AI trained on Credible Data? The Effects of Labeling Quality and Performance Bias on User Trust. 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 816, 11 pages. doi:10.1145/3544548.3580805

[11] Lingjiao Chen, Matei Zaharia, and James Zou. 2023. How is ChatGPT's behavior changing over time? doi:10.48550/arXiv.2307.09009 arXiv:2307.09009 [cs.CL]

[12] Lee J Cronbach. 1951. Coefficient alpha and the internal structure of tests. *psychometrika* 16, 3 (1951), 297–334.

[13] Smit Desai, Christina Ziying Wei, Jaisie Sin, Mateusz Dubiel, Nima Zargham, Shashank Ahire, Martin Porcheron, Anastasia Kuzminykh, Minha Lee, Heloisa Candello, Joel E Fischer, Cosmin Munteanu, and Benjamin R. Cowan. 2024. CUI@CHI 2024: Building Trust in CUIs—From Design to Deployment. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 460, 7 pages. doi:10.1145/3613905.3636287

[14] Yifan Ding, Matthew Facciani, Ellen Joyce, Amrit Poudel, Sanmitra Bhattacharya, Balaji Veeramani, Sal Aguinaga, and Tim Weninger. 2025. Citations and trust in LLM generated responses. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'25/IAAI'25/EAAI'25)*. AAAI Press, Article 2652, 9 pages. doi:10.1609/aaai.v39i22.34550

[15] Shutong Fan, Lan Zhang, and Xiaoyong Yuan. 2026. When AI Persuades: Adversarial Explanation Attacks on Human Trust in AI-Assisted Decision Making. doi:10.48550/arXiv.2602.04003 arXiv:2602.04003 [cs.AI]

[16] Masyura Ahmad Faudzi, Zaihisma Che Cob, Sharul Azim Sharudin, Ridha Omar, and Masitah Ghazali. 2023. The Effects of User Interface Design for Mobile Learning Application on Learner's Extraneous Cognitive Load: A Conceptual Framework. In *Proceedings of the Asian HCI Symposium 2023* (Online, Indonesia) *(Asian CHI '23)*. Association for Computing Machinery, New York, NY, USA, 51–57. doi:10.1145/3604571.3604579

[17] Massimo Franceschet. 2011. PageRank: standing on the shoulders of giants. *Commun. ACM* 54, 6 (June 2011), 92–101. doi:10.1145/1953122.1953146

[18] Jiangen He. 2025. Who Gets Cited? Gender- and Majority-Bias in LLM-Driven Reference Selection. doi:10.48550/arXiv.2508.02740 arXiv:2508.02740 [cs.DL]

[19] Jiangen He and Jiqun Liu. 2025. Not All Transparency Is Equal: Source Presentation Effects on Attention, Interaction, and Persuasion in Conversational Search. doi:10.48550/arXiv.2512.12207 arXiv:2512.12207 [cs.HC]

[20] Jie Huang and Kevin Chang. 2024. Citation: A Key to Building Responsible and Accountable Large Language Models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 464–473. doi:10.18653/v1/2024.findings-naacl.31

[21] Jeff Huang, Ryen W. White, and Susan Dumais. 2011. No clicks, no problem: using cursor movements to understand and improve search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 1225–1234. doi:10.1145/1978942.1979125

[22] Yanwei Huang and Arpit Narechania. 2026. WebSeek: Facilitating Proactive and Reactive Guidance for Decision Making on the Web. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3772318.3791945

[23] Farnaz Jahanbakhsh and David R Karger. 2024. A Browser Extension for in-place Signaling and Assessment of Misinformation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 946, 21 pages. doi:10.1145/3613904.3642473

[24] Anjali Khurana, Hariharan Subramonyam, and Parmit K Chilana. 2024. Why and When LLM-Based Assistants Can Go Wrong: Investigating the Effectiveness of Prompt-Based Interactions for Software Help-Seeking. In *Proceedings of the 29th International Conference on Intelligent User Interfaces* (Greenville, SC, USA) *(IUI '24)*. Association for Computing Machinery, New York, NY, USA, 288–303. doi:10.1145/3640543.3645200

[25] Artur Klingbeil, Cassandra Grützner, and Philipp Schreck. 2024. Trust and reliance on AI — An experimental study on the extent and costs of overreliance on AI. *Comput. Hum. Behav.* 160, C (Nov. 2024), 10 pages. doi:10.1016/j.chb.2024.108352

[26] Jane Li, Scott Huffman, and Akihito Tokuda. 2009. Good abandonment in mobile and PC internet search. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Boston, MA, USA) *(SIGIR '09)*. Association for Computing Machinery, New York, NY, USA, 43–50. doi:10.1145/1571941.1571951

[27] Jiachen Li, Elizabeth D Mynatt, Varun Mishra, and Jonathan Bell. 2025. 'Always Nice and Confident, Sometimes Wrong': Developer's Experiences Engaging Generative AI Chatbots Versus Human-Powered Q&A Platforms. *Proceedings of the ACM on Human-Computer Interaction* 9, 2 (2025), 1–22. doi:10.48550/arXiv.2309.13684

[28] Q Vera Liao and Jennifer Wortman Vaughan. 2023. Ai transparency in the age of llms: A human-centered research roadmap. *arXiv preprint arXiv:2306.01941* 10 (2023). doi:10.48550/arXiv.2306.01941

[29] Nelson Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating Verifiability in Generative Search Engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 7001–7025. doi:10.18653/v1/2023.findings-emnlp.467

[30] Edoardo Loru, Jacopo Nudo, Niccolò Di Marco, Alessandro Santirocchi, Roberto Atzeni, Matteo Cinelli, Vincenzo Cestari, Clelia Rossi-Arnaud, and Walter Quattrociocchi. 2025. The simulation of judgment in LLMs. *Proceedings of the National Academy of Sciences* 122, 42 (2025), e2518443122. doi:10.1073/pnas.2518443122 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2518443122

[31] Luise Metzger, Linda Miller, Martin Baumann, and Johannes Kraus. 2024. Empowering Calibrated (Dis-)Trust in Conversational Agents: A User Study on the Persuasive Power of Limitation Disclaimers vs. Authoritative Style. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 481, 19 pages. doi:10.1145/3613904.3642122

[32] Alaa Mohasseb, Mohamed Bader-El-Den, and Mihaela Cocea. 2018. Question categorization and classification using grammar based approach. *Information Processing & Management* 54, 6 (2018), 1228–1243. doi:10.1016/j.ipm.2018.05.001

[33] Arpit Narechania, Adam Coscia, Emily Wall, and Alex Endert. 2022. Lumos: Increasing Awareness of Analytic Behavior during Visual Data Analysis. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (Jan. 2022), 1009–1018. doi:10.1109/TVCG.2021.3114827

[34] Arpit Narechania, Alex Endert, and Atanu Sinha. 2025. Guidance Source Matters: How Guidance from AI, Expert, or a Group of Analysts Impacts Visual Data Preparation and Analysis. *ACM IUI* (2025). doi:10.1145/3708359.3712166

[35] Arpit Narechania, Alex Endert, and Atanu R Sinha. 2025. Agentic Enterprise: AI-Centric User to User-Centric AI. *arXiv preprint arXiv:2506.22893* (2025). doi:10.48550/arXiv.2506.22893

[36] Arpit Ajay Narechania. 2024. *Designing, Developing, and Democratizing Guidance for Visual Analytics.* Ph. D. Dissertation. Georgia Institute of Technology.

[37] Samir Passi and Mihaela Vorvoreanu. 2022. *Overreliance on AI: Literature Review.* Technical Report MSR-TR-2022-12. Microsoft. https://www.microsoft.com/en-us/research/publication/overreliance-on-ai-literature-review/

[38] Ar Poorva Priyadarshini. 2024. The impact of user interface design on user engagement. *International Journal of Engineering Research & Technology (IJERT)* 13, 3 (2024). doi:10.48550/arXiv.2508.02740

[39] Abdul Razaque, Salim Hariri, and Joon Yoo. 2025. Ai-Driven User Interface Design: Enhancing Digital Learning and Skill Development. (01 2025). doi:10.2139/ssrn.5114814

[40] Daniel M. Russell, Chinmay Kulkarni, Elena L. Glassman, Hariharan Subramonyam, and Nikolas Martelaro. 2024. Human-Computer Interaction and AI: What Practitioners Need to Know to Design and Build Effective AI systems from a Human Perspective. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 600, 3 pages. doi:10.1145/3613905.3636270

[41] Vera Schmitt, Isabel Bezzaoui, Charlott Jakob, Premtim Sahitaj, Qianli Wang, Arthur Hilbert, Max Upravitelev, Jonas Fegert, Sebastian Möller, and Veronika Solopova. 2025. Beyond Transparency: Evaluating Explainability in AI-Supported Fact-Checking. In *Proceedings of the 4th ACM International Workshop on Multimedia AI against Disinformation (MAD' 25)*. Association for Computing Machinery, New York, NY, USA, 63–72. doi:10.1145/3733567.3735566

[42] Kayla Schroeder and Zach Wood-Doughty. 2025. Can You Trust LLM Judgments? Reliability of LLM-as-a-Judge. doi:10.48550/arXiv.2412.12509 arXiv:2412.12509 [cs.CL]

[43] Nikhil Sharma, Q. Vera Liao, and Ziang Xiao. 2024. Generative Echo Chamber? Effect of LLM-Powered Search Systems on Diverse Information Seeking. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1033, 17 pages. doi:10.1145/3613904.3642459

[44] Sofia Eleni Spatharioti, David Rothschild, Daniel G Goldstein, and Jake M Hofman. 2025. Effects of LLM-based Search on Decision Making: Speed, Accuracy, and Overreliance. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1025, 15 pages. doi:10.1145/3706598.3714082

[45] Ivan Srba, Olesya Razuvayevskaya, João A. Leite, Robert Moro, Ipek Baris Schlicht, Sara Tonelli, Francisco Moreno García, Santiago Barrio Lottmann, Denis Teyssou, Valentin Porcellini, Carolina Scarton, Kalina Bontcheva, and Maria Bielikova. 2026. A Survey on Automatic Credibility Assessment Using Textual Credibility Signals in the Era of Large Language Models. *ACM Trans. Intell. Syst. Technol.* 17, 2, Article 26 (Jan. 2026), 80 pages. doi:10.1145/3770077

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 6000–6010.

[47] William H Walters and Esther Isabelle Wilder. 2023. Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports* 13, 1 (2023), 14045. doi:10.1038/s41598-023-41032-5

[48] Thomas Wolf. 2025. open-llm-leaderboard (Open LLM Leaderboard). https://huggingface.co/open-llm-leaderboard

[49] Vinzenz Wolf and Christian Maier. 2024. ChatGPT usage in everyday life: A motivation-theoretic mixed-methods study. *International Journal of Information Management* 79 (2024), 102821. doi:10.1016/j.ijinfomgt.2024.102821

[50] Waheeb Yaqub, Otari Kakhidze, Morgan L. Brockman, Nasir Memon, and Sameer Patil. 2020. Effects of Credibility Indicators on Social Media News Sharing Intent. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3313831.3376213

[51] Enaam Youssef, Mervat Medhat, Soumaya Abdellatif, and Mahra Al Malek. 2024. Examining the effect of ChatGPT usage on students' academic learning and achievement: A survey-based study in Ajman, UAE. *Computers and Education: Artificial Intelligence* 7 (2024), 100316. doi:10.1016/j.caeai.2024.100316