Glanceable Verification Indicators for AI-Assisted Storytelling

Muhan Xu
Creative Computing Institute
University of the Arts London
London, United Kingdom
m.xu0920221@arts.ac.uk

Ziwei Zhao
Beijing Institute of Graphic Communication
School of New Media
Beijing, China
Creative Computing Institute
University of the Arts London
London, United Kingdom
z.zhao0720211@arts.ac.uk

Junyi Zeng
School of Design
Royal College of Art
London, United Kingdom
junyi.zeng@network.rca.ac.uk

Tianhong Wang
School of Psychological and Cognitive Sciences
Peking University
Beijing, China
zxhydlx@163.com

Francesca Panetta
AKO Storytelling Institute
University of the Arts London
London, United Kingdom
f.panetta@arts.ac.uk

Rebecca Fiebrink
Creative Computing Institute
University of the Arts London
London, United Kingdom
r.fiebrink@arts.ac.uk

Nick Bryan-Kinns
Creative Computing Institute
University of the Arts London
London, United Kingdom
n.bryankinns@arts.ac.uk

Mick Grierson
Creative Computing Institute
University of the Arts London
London, United Kingdom
m.grierson@arts.ac.uk

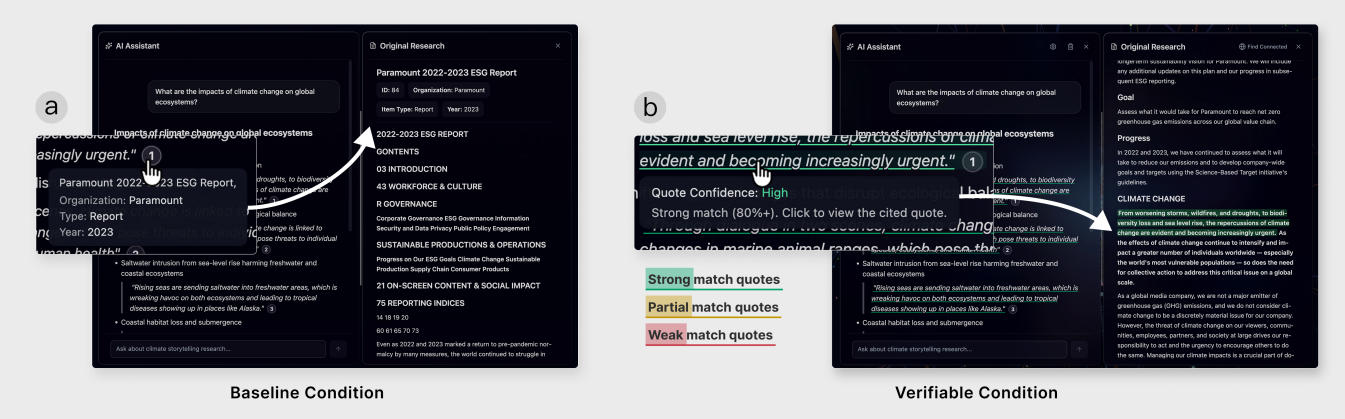

Figure 1: Screenshots of Baseline and Verifiable conditions. Both conditions include citation badges (a) that open source documents on click. Verifiable condition adds color-coded underlines (b) indicating quote matching confidence that scrolls to exact quotes on click.

Abstract

AI-assisted knowledge workers face challenges, including over-reliance on potentially incorrect outputs and LLM hallucinations. Manual verification requires time and domain expertise, and is particularly problematic under deadline pressure. We investigated whether visual confidence cues could serve as glanceable verification indicators in AI-assisted climate storytelling. In a within-subjects study (n = 26), creative professionals edited scripts with AI assistance. We compared baseline and verifiable conditions, both providing AI-generated inline quotes with numbered citation badges. The verifiable condition added color-coded confidence indicators, which matched quotes to sources (green for strong, yellow for partial, and red for weak match). Initial results showed that despite both conditions rating similarly for perceived verification

CHI EA '26, Barcelona, Spain

ACM ISBN 979-8-4007-2281-3/26/04
https://doi.org/10.1145/3772363.3798491

ease, the verifiable condition led to higher trust ratings, citation hovers, and source views. We contribute glanceable verification indicator designs with empirical evidence of increased verification engagement and perceived trust in AI-assisted storytelling, with design implications for verifiable AI systems.

CCS Concepts

• **Human-centered computing → Empirical studies in HCI**.

Keywords

Verification, Trustworthy, LLM, Climate Storytelling, Script Editing, GraphRAG

ACM Reference Format:

Muhan Xu, Ziwei Zhao, Junyi Zeng, Tianhong Wang, Francesca Panetta, Rebecca Fiebrink, Nick Bryan-Kinns, and Mick Grierson. 2026. Glanceable Verification Indicators for AI-Assisted Storytelling. 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, 7 pages. https://doi.org/10.1145/3772363.3798491

1 Introduction

AI-assisted knowledge workers face challenges due to known problems with Large Language Models (LLMs). It is widely understood that LLMs hallucinate, generating plausible yet factually incorrect claims [5, 14, 21, 22]. Further, knowledge workers are known to over-rely on AI outputs even when incorrect [4, 6–8, 24, 41, 48]. Even in the context of professional storytelling, fear of misinformation can reduce user trust in AI systems for complex narrative tasks where strong accuracy is required in relation to background research [40]. While Retrieval-Augmented Generation (RAG) can mitigate hallucinations by incorporating verified external information into the generation process [13, 19, 21, 28, 29, 36, 38, 43], RAG systems remain vulnerable to ambiguous queries and noisy sources [9, 16, 21], potentially generating unreliable outputs [37, 43]. These limitations underscore the need for verification support.

Current verification approaches face challenges across both automatic and manual methods. Automatic metrics such as ROUGE [30] show poor correlation with human judgments [12, 15, 32]. Further, LLM self-evaluation risks evaluation biases [42], and LLM-generated citations can produce incorrect references [1, 31, 47]. Zhang et al. [47] propose simplifying verification by training models to quote verbatim from sources, yet incomplete verbatim quotes necessitate manual checking. Manual verification, where users check AI output correctness and decide on refinements [23], introduces different challenges. Lee et al. [26] documented knowledge workers manually verifying citations and cross-referencing AI-generated content against external sources. However, the verification process demands domain expertise to assess source credibility [26], requires time-intensive reading across multiple sources [23, 26, 46], and increases cognitive burden as users shift from task execution to accuracy checking [23, 26]. Under deadline pressure or when tasks are perceived as lower-stakes, users risk reduced critical engagement and verification effort [26, 46]. This tension between verification requirements and practical limitations suggests a need for interaction designs that support verification while reducing manual effort.

Recent research has integrated verification support into interactive interfaces [17, 18, 25, 27], such as clickable sources that support users in identifying inaccuracies [24, 45]. Design for imperfection research emphasizes strategically revealing system uncertainty to support trust calibration [11, 44]. We build on this work to propose **glanceable verification indicators**: low-attention visual cues that prompt users to verify sources without disrupting primary task focus.

We investigate how color-coded confidence indicators combined with clickable verbatim quotes can support verification in AI-assisted climate storytelling. We conducted a within-subjects study ($n = 26$) comparing a baseline condition to a verifiable condition that added color-coded underlines to inline quotes based on source match confidence. Despite similar perceived ease of verification across conditions, the verifiable condition showed higher trust ratings, more citation hovers, and more source views. Our research contributes to the field in three ways. First, we design glanceable verification indicators that translate confidence scores into color-coded cues for verifying inline quotes. Second, we provide empirical evidence of verification behaviors in AI-assisted storytelling contexts with domain-specific datasets. Third, we identify design tensions between verification support and information overload. Our findings suggest that glanceable indicators can increase verification engagement and perceived trust, extending current thinking in design for imperfection research [11, 44].

2 Methodology

2.1 System Design

We developed a web-based system, Story Assistant, to support user testing of glanceable verification indicators in the context of climate storytelling (Fig. 1). The system addresses the tension between verification requirements and time constraints through color-coded click-to-source verification. Story Assistant enables creative professionals to access 162 climate storytelling research documents curated by the University of Southern California [2], implementing LightRAG [19] for knowledge graph construction and dual-level retrieval. We feed retrieved entities and text excerpts to the open-weights model GPT-OSS-120b [34], using fixed parameters ($Temperature = 0$, $Top\_P = 1$, $Seed = 42$) for reproducibility. To support optional document relationship exploration, the system also provides a 3D knowledge graph visualization [3].

We first prompt an LLM to generate responses with verbatim quotes and citations (see Appendix A). Story Assistant operates as a closed-domain verification system [25]: confidence scores measure consistency between generated quotes and provided source documents, treating content absent from documents as inconsistent regardless of its factual accuracy [33, 39]. This allows us to center user verification interactions against known source materials, without introducing the retrieval complexity of open-domain fact-checking.

Our verifiable system then translates confidence scores (see Table 1 and Appendix B) into color-coded underlines (Fig. 1b). Green underlines (confidence ≥ 0.8) signal strong matches, yellow (0.4-0.8) indicate partial matches requiring verification, and red (<0.4) mark weak matches needing manual fact-checking. Clicking underlined quotes opens the source document and scrolls to the exact quoted location (Fig. 1b). Numbered citation badges ① show metadata on

Table 1: Confidence score calculation attempts exact match, falls back through fuzzy to keyword, to determine color-coded underlines.

Match Strategy	Quote Matching Method	Confidence
Exact match	Verbatim character-level or sequential word match after formatting normalization	1.0
Fuzzy match	Substantial word overlap with paraphrasing, omissions, or ellipsis-separated quotes	0.4–1.0
Keyword match	Only isolated keywords matched (≥3), indicating substantial deviation from source	<0.4

Note: Both quote and source text normalized (markdown formatting and punctuation) before matching

hover (file name, organization, document type, year) and open the source document on click (Fig. 1a).

2.2 User Study Design

We implemented a within-subjects design with two conditions using identical retrieval and generation settings. Both conditions provided numbered citation badges. The **Baseline** condition presented quotes as plain, non-interactive text. The **Verifiable** condition additionally enabled three features: (1) color-coded confidence indicators for inline quotes, (2) clickable quotes opening source documents at exact locations, and (3) 3D knowledge graph exploration.

Each 60-minute session began with an introduction, followed by two counterbalanced tasks (one per condition), and concluded with a 10-minute semi-structured interview. Each task consisted of a 2-minute walkthrough, 20 minutes of script editing, and a 3-minute survey. In each task, participants used AI assistance to develop at least two climate story suggestions based on a single-line concept (e.g., "Europe is hit by a record-breaking heatwave during a detective thriller") with open-ended exploration guidance (see Appendix C). Two professional storytellers refined the task briefs for domain relevance and clarity. This task design allowed participants to navigate documents as needed, enabling observation of verification behaviors emerging during time-constrained exploration. User studies were conducted remotely via Microsoft Teams from October to November 2025 with University Research Ethics Committee approval.

We recruited 26 creative professionals (Table 2) through social media. All participants had experience with climate-related storytelling across television, film, journalism, games, documentaries, theatre, and interactive experiences. Notably, 17/26 participants had over eight years of experience, and none reported color vision deficiency.

2.3 Measures and Data Analysis

Our study employed questionnaires, interaction logs, and interviews. Questionnaires (see Appendix D) used 7-point Likert scales. We measured perceived usefulness and ease of use [10], and the Trust Scale for Explainable AI (TXAI) [20] with item 6 excluded to avoid negatively worded items [35]. We measured verification behaviors adapted from Kazemitabaar et al. [23], as we found no validated scales for AI verification behaviors. Interaction logs captured query and response content, citation hovers, and source document opens and closes across both conditions. The Verifiable condition additionally tracked color-underline hovers and knowledge graph exploration. Semi-structured interviews explored query strategies, trust assessment, verification behavior, system usability, and future-adoption intentions.

Table 2: Demographics of the participants in the user study.

ID	Age	Gender	Professional role	Experience
P01	35-44	Woman	Podcast Producer	4-7 years
P02	45-54	Woman	Story Consultant	8+ years
P03	55-64	Woman	Story Consultant	8+ years
P04	18-24	Woman	Documentary Filmmaker	1-3 years
P05	35-44	Woman	Screenwriter/Impact Producer	4-7 years
P06	25-34	Woman	Story Consultant	4-7 years
P07	35-44	Non-binary	Story Consultant	14 years
P08	45-54	Woman	Journalist/Desk Editor	8+ years
P09	35-44	Woman	Story Consultant	8+ years
P10	25-34	Man	Journalist/Desk Editor	1-3 years
P11	35-44	Woman	Game Designer/Writer	8+ years
P12	35-44	Man	Documentary Filmmaker	8+ years
P13	35-44	Man	Interactive Media Creator	8+ years
P14	35-44	Man	Journalist/Desk Editor	8+ years
P15	25-34	Man	Documentary Filmmaker	1-3 years
P16	35-44	Woman	Story Consultant	8+ years
P17	25-34	Man	Story Consultant	4-7 years
P18	45-54	Woman	Funder/Executive Producer	8+ years
P19	35-44	Man	Documentary Filmmaker	8+ years
P20	35-44	Man	Film Producer	8+ years
P21	35-44	Man	Funder/Innovation Partner	0-1 year
P22	55-64	Man	Film production educator	17 years
P23	25-34	Non-binary	Theater Maker/Producer/Dramaturg	1-3 years
P24	35-44	Woman	Documentary Filmmaker	8+ years
P25	35-44	Woman	Story Consultant	8+ years
P26	35-44	Woman	Game Designer/Writer	8+ years

For quantitative data, we used Wilcoxon signed-rank tests to compare conditions using SPSS 29.0. We computed effect sizes $r = |Z|/\sqrt{N}$, where r = .10, .30, and .50 represent small, medium, and large effects, respectively. For qualitative data, we extracted illustrative quotes from transcripts to contextualize quantitative findings. Detailed thematic analysis is underway for future work.

3 Initial Results

3.1 Trust, Control, and Task Experience Ratings

We compared user ratings between the Verifiable and Baseline conditions (Table 3) using Wilcoxon signed-rank tests ($n = 26$). The Verifiable condition showed significantly higher ($p = .036$) TXAI trust ratings ($M = 5.11$, $SD = 0.84$) compared to Baseline ($M = 4.58$, $SD = 1.33$). Participants reported that in the Verifiable condition, it was significantly easier to intervene (Verifiable: $M = 4.50$, $SD = 0.71$; Baseline: $M = 3.65$, $SD = 1.23$; $p = .004$) and steer the AI (Verifiable: $M = 4.77$, $SD = 1.53$; Baseline: $M = 4.08$, $SD = 1.94$; $p = .025$). Task ease ratings (Verifiable: $M = 5.35$, $SD = 1.41$; Baseline: $M = 4.50$, $SD = 1.97$; $p = .025$) and perceived success (Verifiable: $M = 4.69$, $SD = 1.62$; Baseline: $M = 4.00$, $SD = 1.79$; $p = .033$) both increased significantly.

Table 3: Within-subjects comparison of quantitative results ($n = 26$) between Baseline and Verifiable conditions. We calculated the Mean (M) and Standard Deviation (SD) for each measure and condition, along with the Wilcoxon signed-rank test results and effect sizes. All measures use 7-point Likert scales. "LIB" stands for "Lower Is Better", with best results in bold.

Measure	Baseline (*SD*)	Verifiable (*SD*)	*Z*	*p*	*r*
Verification Behavior					
Citation Hovers	3.46 (4.22)	**21.35 (13.45)**	−4.28	<.001***	.84
Source Views	1.77 (2.54)	**2.58 (2.93)**	−1.14	.256	.22
Source View Duration (s)	105.22 (197.40)	**192.31 (267.01)**	−1.35	.177	.26
Source Short Views (<30s)	0.35 (0.98)	**1.31 (2.13)**	−2.05	.040*	.40
Source Long Views (≥30s)	0.62 (1.10)	**1.31 (1.87)**	−1.82	.069	.36
Usability and Trust					
Trust Scale for Explainable AI	4.58 (1.33)	**5.11 (0.84)**	−2.10	.036*	.41
Perceived Usefulness	4.83 (2.00)	**5.32 (1.35)**	−1.45	.148	.28
Perceived Ease of Use	5.36 (1.11)	**5.47 (0.89)**	−0.41	.680	.08
Task Ease	4.50 (1.97)	**5.35 (1.41)**	−2.25	.025*	.44
Verification Ease	5.19 (1.44)	**5.38 (1.20)**	−0.42	.673	.08
Perceived Success	4.00 (1.79)	**4.69 (1.62)**	−2.13	.033*	.42
Controllability					
Intervention Ease	3.65 (1.23)	**4.50 (0.71)**	−2.84	.004**	.56
Steering Ease	4.08 (1.94)	**4.77 (1.53)**	−2.24	.025*	.44
Perceived Control	3.88 (1.45)	**4.50 (1.39)**	−1.82	.069	.36
Trade-offs					
Frustration (LIB)	3.31 (1.91)	**2.62 (1.39)**	−1.90	.058	.37
Information Overload (LIB)	**2.15 (1.67)**	2.88 (1.82)	−1.84	.066	.36

*$p < .05$, **$p < .01$, ***$p < .001$

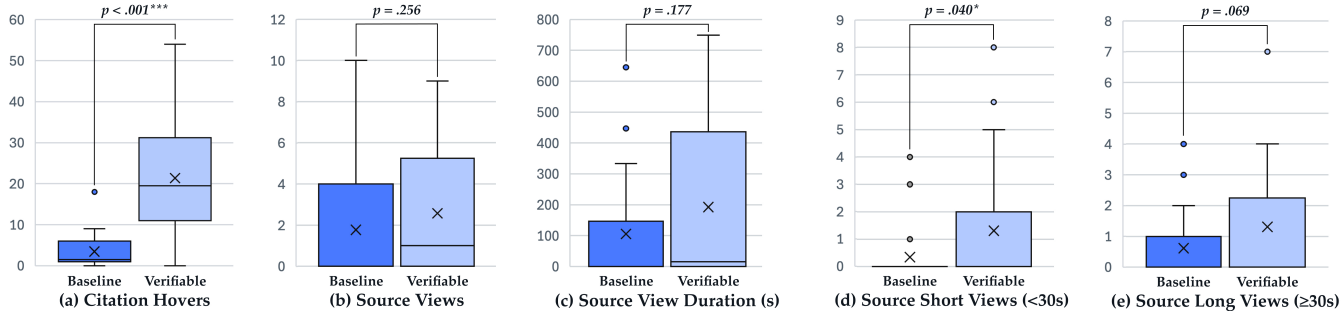

Figure 2: Per-user verification behavior counts from interaction logs ($n = 26$) comparing Baseline and Verifiable conditions: (a) Citation Hovers and Source documents: (b) Views, (c) View Duration, (d-e) Short and Long Views. Each box plot shows median (line), mean (x), interquartile range (box), and outliers (points).

However, perceived ease of verification did not reach significance (Verifiable: $M = 5.38$, $SD = 1.20$; Baseline: $M = 5.19$, $SD = 1.44$; $p = .673$). Examining potential trade-offs, information overload marginally increased (Verifiable: $M = 2.88$, $SD = 1.82$; Baseline: $M = 2.15$, $SD = 1.67$; $p = .066$), while frustration marginally decreased approaching significance (Verifiable: $M = 2.62$, $SD = 1.39$; Baseline: $M = 3.31$, $SD = 1.91$, $p = .058$).

3.2 Citation and Source Engagement

Interaction logs (Fig. 2) captured citation engagement, source document views, and knowledge graph exploration. Query submission remained stable across conditions (Baseline: $M = 9.46$, $SD = 4.57$; Verifiable: $M = 8.31$, $SD = 3.80$; $Z = -1.15$, $p = .251$), suggesting isolating verification feature effects. Meanwhile, only 6/26 participants clicked graph nodes (28 total clicks), indicating limited 3D knowledge graph engagement.

The Verifiable condition showed a substantial increase in verification engagement. Hovers on citations significantly increased ($p < .001$) from Baseline ($M = 3.46$, $SD = 4.22$, 21/26 users) to Verifiable ($M = 21.35$, $SD = 13.45$, 25/26 users), a six-fold increase with a large effect size ($r = .84$). Notably, in the Verifiable condition, color-coded underlines account for 94.6% of citation hovers (525/555) rather than citation badges. Among participants who hovered on both types ($n = 18$), low-confidence underlines received

significantly more ($Z = -2.72$, $p = .006$, $r = .64$) hovers per underline ($n = 40$, $M = 3.24$, $SD = 4.78$) than high-confidence underlines ($n = 486$, $M = 0.86$, $SD = 0.49$), indicating repeated verification of uncertain quotes. To validate confidence score accuracy, we manually post-hoc reviewed all 84 medium- and low-confidence quotes, representing 6.3% of the 1,340 total generated quotes. Of the 84 quotes reviewed, 79 (94.0%) were confirmed incorrect: 37 fabricated, 36 misattributed, 6 paraphrases, with 5 false positives.

Source views increased from Baseline ($M = 1.77$, $SD = 2.54$, 11/26 users) to Verifiable ($M = 2.58$, $SD = 2.93$, 17/26 users; $p = .256$), with increased source view duration (Baseline: $M = 105.22$s, $SD = 197.40$; Verifiable: $M = 192.31$s, $SD = 267.01$; $p = .177$). Participants made significantly more brief inspections under 30 seconds in the Verifiable condition ($p = .040$) compared to the Baseline condition (Baseline: $M = 0.35$, $SD = 0.98$; Verifiable: $M = 1.31$, $SD = 2.13$).

Participants' qualitative feedback reflected engagement with confidence underlines. P9 described color-underline as an "*easy traffic light system, which I'm slightly obsessed with now*," while P25 noted "*humans are so adapted to seeing the red, yellow, green...so easy off the bat of what things were red flagging for me*." P20 found confidence scores increased trust "*when I see 100%...confidence score*."

3.3 Usability and Dataset Limitations

Our preliminary qualitative analysis documented limitations participants encountered. P23 noted visual density issues, describing quotes as "*a big wall of text*", requiring less overwhelming presentation. Time constraints limited further verification (P4, P15, P21, P23). Given more time, participants would read source documents (P4, P23), Google cross-checking (P4, P15), and manually verify specific facts (P23). Moreover, participants identified dataset coverage gaps. P16 noted missing Global South content despite climate storytelling being "*a global issue*." P17 could not get a response to region-specific data. P20 termed "*not enough information*" responses as "*hard stops*," expecting partial answers rather than complete refusals since typical AI systems "*don't stop unless ... too confidential or private*." P18 flagged hydrogen-related responses as "*misleading and not factually correct*," and P23 noted "*outdated sources*."

4 Discussion

4.1 The Verification Paradox

Our initial findings suggest a verification paradox: participants reported similar ease of verification across conditions, yet interaction logs showed significantly more verification engagement in the Verifiable condition. These results suggest that color-underline functioned as glanceable verification indicators, prompting source checking that participants did not consciously recognize as verification. This reflects the core aim of glanceable design: supporting verification without disrupting primary task focus. As P9 suggested, the traffic light color scheme leveraged culturally familiar visual conventions, allowing reliability assessments through quick visual scanning.

This design directly addresses documented barriers: users reduce verification effort under time pressure [23, 26, 46]. By integrating glanceable indicators into the reading flow, the design enables quick assessments despite time constraints that participants report can limit deeper verification. This paradox suggests a rich design space for future research: exploring a broader vocabulary of glanceable cues (e.g., color, shape, and position) that support verification engagement without increasing perceived burden.

4.2 Towards Trust Calibration Through Glanceable Indicators

The dual increase in trust ratings and verification engagement suggests glanceable verification indicators may support trust calibration. Interaction logs show that low-confidence underlines received significantly more hovers than high-confidence ones, suggesting that color-coding actively directed attention to uncertain content rather than simply adding clickable elements. These findings extend design for imperfection research [11, 24, 44] and contribute to ongoing work on addressing over-reliance concerns [4, 6–8, 24, 41, 48]. Given the potential of glanceable verification indicators, future research can extend glanceable verification indicators to other AI-assisted contexts, such as legal and medical, where accuracy is critical yet time for verification is limited. Future studies should explore how varied visual cue designs drive verification across different expertise levels and task constraints.

However, the marginal increase in perceived information overload suggests that maintaining verification support without task disruption requires careful design balance. As participants indicated, while color-underlines support rapid reliability scanning, inline verbatim quotes increase visual density, risk reintroducing the verification burden glanceability aims to reduce. Future work should balance glanceability against visual density. One possible approach might involve applying color codes only to citation badges (e.g., ① ② ③) while showing verbatim quotes only upon hover.

4.3 Dataset Boundaries and User Expectation Calibration

Dataset limitations emerged across coverage, scale, recency, and accuracy dimensions. Participants noted a lack of Global South perspectives, frequent empty responses from the 162-document dataset for out-of-scope queries, and concerns about outdated or inaccurate information. These limitations reflect known RAG challenges of insufficient retrieved documents [9, 13, 16] and noisy sources [9, 21]. Our strict prompt design, requiring refusal for out-of-scope queries, amplified user frustration. Participants encountered "*hard stops*" conflicting with expectations shaped by general-purpose AI systems, even after reading task introductions specifying dataset scope. This suggests domain-specific datasets with curated sources address retrieval quality [37, 43] but simultaneously constrain response scope, creating friction when systems prioritize accuracy over responsiveness.

To foster appropriate reliance [8, 24], future work could empirically examine whether users prefer plausible responses over explicit refusals, and how user tolerance for uncertainty varies across task contexts. Based on our preliminary results, we suggest two approaches. First, expand dataset coverage through expert review panels to ensure controlled diversity and quality. Alternatively, augment the internal dataset with tiered web search, prioritizing whitelisted reputable sources over general web results. Second, calibrate user expectations by explicitly acknowledging system

limitations without frustrating users. This includes providing partial responses with explicit knowledge gap acknowledgments or suggesting query refinements aligned with system scope. Future domain-specific tools should balance dataset coverage constraints with communication strategies that calibrate user expectations.

4.4 Limitations

Three limitations should be noted. First, our Verifiable condition combined color-coded indicators, clickable quotes, and a 3D knowledge graph, making it difficult to isolate individual feature effects. However, given the limited engagement with the knowledge graph (6/26 participants; 28 total clicks), the observed effects are more likely associated with the color-coded quotes. Future studies should isolate each feature to clarify individual effects. Second, our 20-minute task reflected real-world time constraints but may have limited verification depth. Longer-term studies are needed to observe how verification behaviors evolve. Third, while post-hoc analysis revealed hallucinated citations in the AI output, we did not measure participant error detection. This limits our claims: we demonstrate increased verification engagement, but cannot confirm improved verification accuracy. To understand whether increased verification engagement translates to successful error detection, future studies should evaluate how varying glanceable indicator designs affect user verification accuracy against different levels of system hallucination.

5 Conclusion

We contribute a glanceable verification indicator design that directs attention to uncertain content within AI-assisted climate storytelling. In a within-subjects study ($n = 26$), color-underlines showed increased trust ratings, citation hovers, and source views. These initial findings suggest that revealing system uncertainty through familiar visual conventions can support increased verification behaviors and perceived trust. Future work should explore more glanceable verification indicator strategies to support appropriate trust calibration and extend the design for imperfection research.

Acknowledgments

The authors thank Josh Cockcroft, Sean Marshall, Olga Sutskova, Yazhuo Zhou, and Mengzhi He for their valuable support and discussions on this work. They also want to thank all participants for their valuable time and all reviewers for their constructive feedback.

References

[1] Ayush Agrawal, Mirac Suzgun, Lester Mackey, and Adam Kalai. 2024. Do Language Models Know When They're Hallucinating References?. In *Findings of the Association for Computational Linguistics: EACL 2024*, Yvette Graham and Matthew Purver (Eds.). Association for Computational Linguistics, St. Julian's, Malta, 912–928. https://aclanthology.org/2024.findings-eacl.62/

[2] Entertainment Alliance. n.d.. Sustainability On Screen: A Research Database. https://sustainable-entertainment.com/sustainability-on-screen-a-research-database

[3] Vasco Asturiano. 2026. vasturiano/3d-force-graph. https://github.com/vasturiano/3d-force-graph

[4] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–16. doi:10.1145/3411764.3445717

[5] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Virtual Event Canada, 610–623. doi:10.1145/3442188.3445922

[6] Jessica Y Bo, Sophia Wan, and Ashton Anderson. 2025. To Rely or Not to Rely? Evaluating Interventions for Appropriate Reliance on Large Language Models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–23. doi:10.1145/3706598.3714097

[7] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. ACM, Cagliari Italy, 454–464. doi:10.1145/3377325.3377498

[8] 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. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 1–21. doi:10.1145/3449287

[9] Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The Power of Noise: Redefining Retrieval for RAG Systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Washington DC USA, 719–729. doi:10.1145/3626772.3657834

[10] Fred D. Davis. 1989. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly* 13, 3 (Sept. 1989), 319. doi:10.2307/249008

[11] Upol Ehsan, Q. Vera Liao, Samir Passi, Mark O. Riedl, and Hal Daumé. 2024. Seamful XAI: Operationalizing Seamful Design in Explainable AI. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (April 2024), 1–29. doi:10.1145/3637396

[12] Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking Generated Summaries by Correctness: An Interesting but Challenging Application for Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 2214–2220. doi:10.18653/v1/P19-1213

[13] Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, Barcelona Spain, 6491–6501. doi:10.1145/3637528.3671470

[14] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature* 630, 8017 (June 2024), 625–630. doi:10.1038/s41586-024-07421-0

[15] Mingqi Gao and Xiaojun Wan. 2022. DialSummEval: Revisiting Summarization Evaluation for Dialogues. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 5693–5709. doi:10.18653/v1/2022.naacl-main.418

[16] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. doi:10.48550/arXiv.2312.10997

[17] Ken Gu, Ruoxi Shang, Tim Althoff, Chenglong Wang, and Steven M. Drucker. 2024. How Do Analysts Understand and Verify AI-Assisted Data Analyses?. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–22. doi:10.1145/3613904.3642497

[18] Ziwei Gu, Ian Arawjo, Kenneth Li, Jonathan K. Kummerfeld, and Elena L. Glassman. 2024. An AI-Resilient Text Rendering Technique for Reading and Skimming Documents. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–22. doi:10.1145/3613904.3642699

[19] Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2025. LightRAG: Simple and Fast Retrieval-Augmented Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 10746–10761. doi:10.18653/v1/2025.findings-emnlp.568

[20] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2023. Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science* 5 (Feb. 2023), 1096257. doi:10.3389/fcomp.2023.1096257

[21] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems* 43, 2 (March 2025), 1–55. doi:10.1145/3703155

[22] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *Comput. Surveys* 55, 12 (Dec. 2023), 1–38. doi:10.1145/3571730

[23] Majeed Kazemitabaar, Jack Williams, Ian Drosos, Tovi Grossman, Austin Zachary Henley, Carina Negreanu, and Advait Sarkar. 2024. Improving Steering and Verification in AI-Assisted Data Analysis with Interactive Task Decomposition. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and*

Technology. ACM, Pittsburgh PA USA, 1–19. doi:10.1145/3654777.3676345

[24] Sunnie S. Y. Kim, Jennifer Wortman Vaughan, Q. Vera Liao, Tania Lombrozo, and Olga Russakovsky. 2025. Fostering Appropriate Reliance on Large Language Models: The Role of Explanations, Sources, and Inconsistencies. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–19. doi:10.1145/3706598.3714020

[25] Philippe Laban, Jesse Vig, Marti Hearst, Caiming Xiong, and Chien-Sheng Wu. 2024. Beyond the Chat: Executable and Verifiable Text-Editing with LLMs. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. ACM, Pittsburgh PA USA, 1–23. doi:10.1145/3654777.3676419

[26] Hao-Ping (Hank) Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. 2025. The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–22. doi:10.1145/3706598.3713778

[27] Florian Leiser, Sven Eckhardt, Valentin Leuthe, Merlin Knaeble, Alexander Mädche, Gerhard Schwabe, and Ali Sunyaev. 2024. HILL: A Hallucination Identifier for Large Language Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–13. doi:10.1145/3613904.3642428

[28] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc., Virtual, 9459–9474. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[29] Xiaoxi Li, Jiajie Jin, Yujia Zhou, Yuyao Zhang, Peitian Zhang, Yutao Zhu, and Zhicheng Dou. 2025. From Matching to Generation: A Survey on Generative Information Retrieval. *ACM Transactions on Information Systems* 43, 3 (May 2025), 1–62. doi:10.1145/3722552

[30] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013.pdf

[31] Nelson Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating Verifiability in Generative Search Engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Singapore, 7001–7025. doi:10.18653/v1/2023.findings-emnlp.467

[32] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 2511–2522. doi:10.18653/v1/2023.emnlp-main.153

[33] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 1906–1919. doi:10.18653/v1/2020.acl-main.173

[34] OpenAI, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, Kai Chen, Mark Chen, Enoch Cheung, Aidan Clark, Dan Cook, Marat Dukhan, Casey Dvorak, Kevin Fives, Vlad Fomenko, Timur Garipov, Kristian Georgiev, Mia Glaese, Tarun Gogineni, Adam Goucher, Lukas Gross, Katia Gil Guzman, John Hallman, Jackie Hehir, Johannes Heidecke, Alec Helyar, Haitang Hu, Romain Huet, Jacob Huh, Saachi Jain, Zach Johnson, Chris Koch, Irina Kofman, Dominik Kundel, Jason Kwon, Volodymyr Kyrylov, Elaine Ya Le, Guillaume Leclerc, James Park Lennon, Scott Lessans, Mario Lezcano-Casado, Yuanzhi Li, Zhuohan Li, Ji Lin, Jordan Liss, Lily, Liu, Jiancheng Liu, Kevin Lu, Chris Lu, Zoran Martinovic, Lindsay McCallum, Josh McGrath, Scott McKinney, Aidan McLaughlin, Song Mei, Steve Mostovoy, Tong Mu, Gideon Myles, Alexander Neitz, Alex Nichol, Jakub Pachocki, Alex Paino, Dana Palmie, Ashley Pantuliano, Giambattista Parascandolo, Jongsoo Park, Leher Pathak, Carolina Paz, Ludovic Peran, Dmitry Pimenov, Michelle Pokrass, Elizabeth Proehl, Huida Qiu, Gaby Raila, Filippo Raso, Hongyu Ren, Kimmy Richardson, David Robinson, Bob Rotsted, Hadi Salman, Suvansh Sanjeev, Max Schwarzer, D. Sculley, Harshit Sikchi, Kendal Simon, Karan Singhal, Yang Song, Dane Stuckey, Zhiqing Sun, Philippe Tillet, Sam Toizer, Foivos Tsimpourlas, Nikhil Vyas, Eric Wallace, Xin Wang, Miles Wang, Olivia Watkins, Kevin Weil, Amy Wendling, Kevin Whinnery, Cedric Whitney, Hannah Wong, Lin Yang, Yu Yang, Michihiro Yasunaga, Kristen Ying, Wojciech Zaremba, Wenting Zhan, Cyril Zhang, Brian Zhang, Eddie Zhang, and Shengjia Zhao. 2025. gpt-oss-120b & gpt-oss-20b Model Card. doi:10.48550/arXiv.2508.10925

[35] Sebastian A. C. Perrig, Nicolas Scharowski, and Florian Brühlmann. 2023. Trust Issues with Trust Scales: Examining the Psychometric Quality of Trust Measures in the Context of AI. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–7. doi:10.1145/3544549.3585808

[36] Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-Context Retrieval-Augmented Language Models. *Transactions of the Association for Computational Linguistics* 11 (Nov. 2023), 1316–1331. doi:10.1162/tacl_a_00605

[37] Yijia Shao, Yucheng Jiang, Theodore Kanell, Peter Xu, Omar Khattab, and Monica Lam. 2024. Assisting in Writing Wikipedia-like Articles From Scratch with Large Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 6252–6278. doi:10.18653/v1/2024.naacl-long.347

[38] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval Augmentation Reduces Hallucination in Conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Punta Cana, Dominican Republic, 3784–3803. doi:10.18653/v1/2021.findings-emnlp.320

[39] Liyan Tang, Philippe Laban, and Greg Durrett. 2024. MiniCheck: Efficient Fact-Checking of LLMs on Grounding Documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Miami, Florida, USA, 8818–8847. doi:10.18653/v1/2024.emnlp-main.499

[40] Yuying Tang, Haotian Li, Minghe Lan, Xiaojuan Ma, and Huamin Qu. 2025. Understanding Screenwriters' Practices, Attitudes, and Future Expectations in Human-AI Co-Creation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–18. doi:10.1145/3706598.3714120

[41] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. 2023. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (April 2023), 1–38. doi:10.1145/3579605

[42] Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large Language Models are not Fair Evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Bangkok, Thailand, 9440–9450. doi:10.18653/v1/2024.acl-long.511

[43] Xiaohua Wang, Zhenghua Wang, Xuan Gao, Feiran Zhang, Yixin Wu, Zhibo Xu, Tianyuan Shi, Zhengyuan Wang, Shizheng Li, Qi Qian, Ruicheng Yin, Changze Lv, Xiaoqing Zheng, and Xuanjing Huang. 2024. Searching for Best Practices in Retrieval-Augmented Generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Miami, Florida, USA, 17716–17736. doi:10.18653/v1/2024.emnlp-main.981

[44] Justin D. Weisz, Jessica He, Michael Muller, Gabriela Hoefer, Rachel Miles, and Werner Geyer. 2024. Design Principles for Generative AI Applications. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–22. doi:10.1145/3613904.3642466

[45] Muhan Xu, Xiaomei Li, Te Lan, Xiaoke Zeng, Tianhong Wang, Sixiong Sheng, Xiaolin Deng, Jingjing Sun, Rebecca Fiebrink, Nick Bryan-Kinns, and Mick Grierson. 2026. Comparing Structural and Conversational AI Summarisation for Time-Constrained Academic Reading Support among Graduate Students. 0, 0 (2026), 1–23. doi:10.1080/10494820.2025.2604649

[46] Bhada Yun, Dana Feng, Ace S. Chen, Afshin Nikzad, and Niloufar Salehi. 2025. Generative AI in Knowledge Work: Design Implications for Data Navigation and Decision-Making. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–19. doi:10.1145/3706598.3713337

[47] Jingyu Zhang, Marc Marone, Tianjian Li, Benjamin Van Durme, and Daniel Khashabi. 2025. Verifiable by Design: Aligning Language Models to Quote from Pre-Training Data. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Luis Chiruzzo, Alan Ritter, and Lu Wang (Eds.). Association for Computational Linguistics, Albuquerque, New Mexico, 3748–3768. doi:10.18653/v1/2025.naacl-long.191

[48] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, Barcelona Spain, 295–305. doi:10.1145/3351095.3372852