Glanceable Verification Indicators for AI-Assisted Storytelling: Signals and Evaluation for Verification Interventions

Muhan Xu[1], Ziwei Zhao[1,2], Junyi Zeng[3], Tianhong Wang[4], Francesca Panetta[1], Rebecca Fiebrink[1], Nick Bryan-Kinns[1] and Mick Grierson[1]

[1]University of the Arts London, London, United Kingdom
[2]Beijing Institute of Graphic Communication, Beijing, China
[3]Royal College of Art, London, United Kingdom
[4]Peking University, Beijing, China

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). Our initial results suggested a verification paradox: despite similar perceived ease of verification across conditions, the verifiable condition showed higher trust ratings, citation hovers, and source views. In this position paper, we contribute to the TimelyAI workshop discussion on two topics. Regarding *what signals matter*, we present evidence that source-match confidence scores can direct user attention toward uncertain content, while contextual factors, including time pressure, task stage, and individual differences, modulated signal uptake. Regarding *evaluating intervention effects*, the verification paradox suggested that self-report measures alone can miss behavioral effects of verification interventions, and that evaluation should capture both benefits and costs simultaneously.

Keywords

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

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 [1, 2, 3, 4]. Further, knowledge workers are known to over-rely on AI outputs even when incorrect [5, 6, 7, 8, 9, 10, 11]. 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 [12]. While Retrieval-Augmented Generation (RAG) can mitigate hallucinations by incorporating verified external information into the generation process [2, 13, 14, 15, 16, 17, 18, 19], RAG systems remain vulnerable to ambiguous queries and noisy sources [2, 20, 21], potentially generating unreliable outputs [16, 22]. These limitations underscore the need for verification support.

Current verification approaches face challenges across both automatic and manual methods. Automatic metrics such as ROUGE [23] show poor correlation with human judgments [24, 25, 26]. Further, LLM self-evaluation risks evaluation biases [27], and LLM-generated citations can produce incorrect

Proceedings of TimelyAI: When Should Generative AI Assistants Intervene? A Workshop Co-located with the 5th Symposium on Human-Computer Interaction for Work (CHIWORK 2026), Linz, Austria, June 22 to 25, 2026
m.xu0920221@arts.ac.uk (M. Xu)
0009-0004-9692-3148 (M. Xu); 0009-0006-5799-1951 (Z. Zhao); 0009-0009-0668-3088 (J. Zeng); 0000-0002-9008-1582 (T. Wang); 0009-0000-6868-9076 (F. Panetta); 0000-0002-7609-2234 (R. Fiebrink); 0000-0002-1382-2914 (N. Bryan-Kinns); 0000-0002-6981-5414 (M. Grierson)

references [28, 29, 30]. Zhang et al. [30] 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 [31], introduces different challenges. Lee et al. [32] 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 [32], requires time-intensive reading across multiple sources [33, 31, 32], and increases cognitive burden as users shift from task execution to accuracy checking [31, 32]. Under deadline pressure or when tasks are perceived as lower-stakes, users risk reduced critical engagement and verification effort [32, 33]. 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 [34, 35, 36, 37], such as clickable sources that support users in identifying inaccuracies [10, 38]. Design for imperfection research emphasizes strategically revealing system uncertainty to support trust calibration [39, 40]. 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.* While prior work has examined what verification support to provide, less attention has been paid to evaluating whether and how such support effectively supports users during task performance. In this position paper, we draw on empirical findings from a within-subjects study to address two questions relevant to the TimelyAI workshop agenda: *what* signals effectively direct user attention to uncertain AI outputs, and *how* should the effects of such verification interventions be evaluated.

We investigated how color-coded confidence indicators combined with clickable verbatim quotes can support verification in AI-assisted climate storytelling. In a within-subjects study ($n = 26$), creative professionals edited scripts with AI assistance. We compared 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. Based on these findings, we contribute to the workshop discussion in two ways. First, we present evidence that source-match confidence scores serve as an effective signal for directing user attention to uncertain content, while identifying contextual factors (e.g., time pressure, task stage, individual differences) can affect signal uptake. Second, we demonstrate a verification paradox, increased verification engagement without a corresponding decrease in perceived verification ease rating, suggesting that evaluating verification interventions requires behavioral measures alongside self-report to capture effects that users may not consciously recognize.

2. Methodology

2.1. System Design

We developed Story Assistant, a web-based system for evaluating 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 [41], a publicly available collection covering topics such as audience demand, industry insights, and the impact of climate narratives. The system employed LightRAG [19] for knowledge graph construction and dual-level retrieval. Retrieved entities and text excerpts are posed to the open-weights model GPT-OSS-120b [42] with fixed parameters ($Temperature = 0$, $Top\_P = 1$, $Seed = 42$) to ensure reproducibility. To support optional exploration of document relationships, the system also provides a 3D knowledge graph visualization [43].

We first prompt an LLM to generate responses with verbatim quotes and citations (see Appendix B). Story Assistant operates as a closed-domain verification system [34]: confidence scores measure consistency between generated quotes and provided source documents, treating content absent from documents as inconsistent regardless of its factual accuracy [44, 45]. This allows us to center user

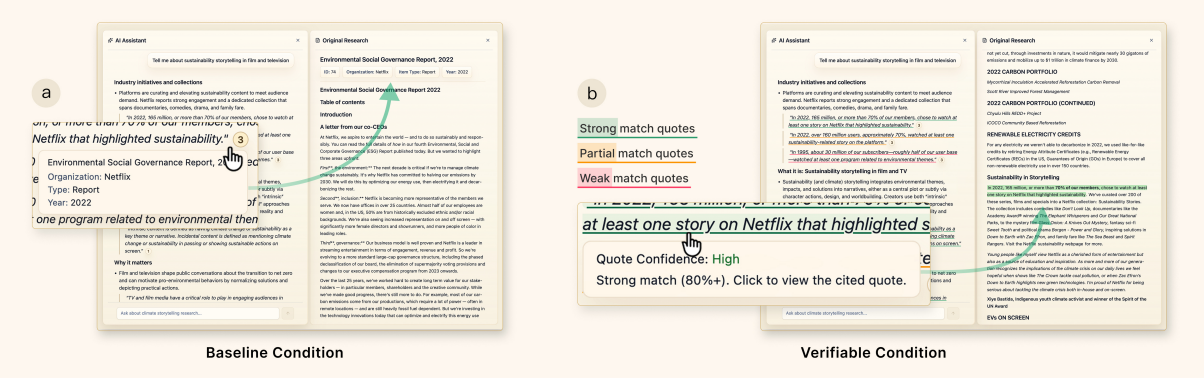

Figure 1: Screenshots of Baseline and Verifiable conditions. Both conditions include citation badges (a) that open source documents on click. Verifiable condition adds color-coded underlines (b) indicating quote matching confidence that scrolls to exact quotes on click.

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 C) 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. This design builds on prior evidence that color-coded confidence cues can support decision-making when AI outputs contain errors [46]. Clicking underlined quotes opens the corresponding source document and scrolls to the exact quoted location (Fig. 1b), enabling users to verify quotes directly against source materials [10, 38]. Numbered citation badges ① show metadata on hover (file name, organization, document type, year) and open the source document on click (Fig. 1a). The system proactively displays confidence signals for all generated quotes, rather than requiring users to initiate verification.

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 that displayed metadata on hover and opened source documents on click. The **Baseline** condition presented inline quotes as plain text, while the **Verifiable** condition rendered inline quotes with color-coded confidence underlines. In the Verifiable condition, clicking an underlined quote opened the source document at the exact quoted location (Fig. 1b), with an optional 3D knowledge graph also available for exploration. User studies were conducted remotely via Microsoft Teams from October to November 2025 with University Research Ethics Committee approval.

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 of the assigned condition, 20 minutes of script editing, and a 3-minute post-task survey. In each task, participants used AI assistance to develop at least two climate story suggestions based on a single-line concept. Two concepts were counterbalanced across conditions: "Europe is hit by a record-breaking heatwave during a detective thriller" and "A near-future story where Europe becomes the dominant climate solution leader." Task briefs provided possible exploration directions but allowed participants to pursue whatever direction they found most valuable (see Appendix D). Two professional storytellers from the research team refined the task briefs for domain relevance and clarity. This task design allowed participants to navigate documents as needed, enabling observation of verification during time-constrained exploration.

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,

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

and interactive experiences. Notably, 17/26 participants had over eight years of experience, and none reported color vision deficiency. Participants received a 30 GBP Amazon gift card as compensation.

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

2.3. Measures and Data Analysis

Our study employed questionnaires, interaction logs, and interviews. Questionnaires (see Appendix E) used 7-point Likert scales. For perceived trust, we used the Trust Scale for Explainable AI (TXAI) [47] with item 6 excluded to avoid negatively worded items [48]. For task experience, we measured perceived usefulness and ease of use [49]. For verification behaviors, we adapted measures from Kazemitabaar et al. [31], 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. Source view duration was computed as the interval between the source document's open and close events. We further categorized views as short (<30s) or long (≥30s) to distinguish between quick reference

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

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

checks and sustained reading. 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 (see Appendix F).

For quantitative data, we used Wilcoxon signed-rank tests to compare conditions using SPSS 29.0. This non-parametric approach was chosen given our sample size ($n = 26$), ordinal Likert-scale data, and count-based interaction logs. 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

We compared user ratings between the Verifiable and Baseline conditions (Table 3) using Wilcoxon signed-rank tests ($n = 26$). 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 the observed differences reflect verification feature effects rather than changes in query behavior. Knowledge graph engagement was limited: only 6/26 participants clicked graph nodes (28 total clicks).

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

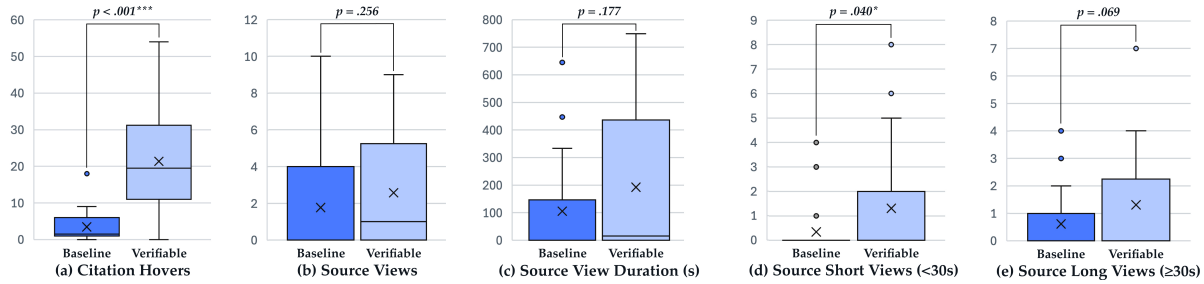

Figure 2: Per-user verification behavior counts from interaction logs ($n = 26$) comparing Baseline and Verifiable conditions: (a) Citation Hovers and Source documents: (b) Views, (c) View Duration, (d-e) Short and Long Views. Each box plot shows median (line), mean (x), interquartile range (box), and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

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.

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: $M = 3.31$, $SD = 1.91$, $p = .058$).

Our preliminary qualitative analysis revealed how trust developed through interaction with verification features. Eight participants reported that trust increased through actively checking sources, with P22 noting that trust developed gradually: "*after a while [...] I then became aware that the reference system was clear and of a high standard.*" P20 reported that high confidence scores made them trust "*the references more.*" However, P15 reported "*I felt very confident [...] whatever it's giving me is quite verified*" and "*didn't walk through it very thoroughly*," raising the possibility that confidence scores may reduce verification depth for some users.

3.2. Color-Coded Underline 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.

Interview data revealed how participants experienced color-coded underlines during verification. Half of the participants (13/26) reported that color-underlines saved time on source verification, and 6/26 noted that color-coding made confidence levels easy to see. The contrast with Baseline was notable: in the Baseline condition, 5/26 participants reported skipping verification entirely, with P25 noting that verification "*was going to take me a lot of work.*" In the Verifiable condition, the same participant P25 noted that color cues made it "*so easy off the bat of what things were red flagging for me*," and P9 described color-underlines as an "*easy traffic light system, which I'm slightly obsessed with now.*"

Beyond reducing verification effort, color-underlines shaped how participants directed their verification behavior. Six participants reported using color cues to guide selective source filtering: assessing source robustness (P3, P4, P5), deprioritizing low-confidence content (P9, P25), and balancing source reliability against content relevance (P26). P9 found that color-underlines helped determine "*what the next question was going to be and what path I was going to follow.*" Additionally, two participants (P5, P8) reported that color-underlines served as reminders to fact-check AI outputs. P5 described color-coding as "*a really important cognitive move to remind me that I am reading AI information, I'm not just reading definite facts.*"

When participants proceeded to verify source content, they described features that reduced the attentional cost of navigation as particularly useful. Nine participants found click-through access to sources helpful, and 10/26 valued direct links to exact source locations. P13 valued accessing sources "*without closing my first window [...] side-by-side information I found really useful.*" P1 reported that "*direct linkage and highlighting of the actual phrase or sentence in the source material*" increased verification confidence.

3.3. Source Document Engagement

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: $M = 1.31$, $SD = 2.13$).

Interview data provided context for these source verification patterns. Given the 20-minute task duration, 11/26 participants reported skipping or deferring verification due to time pressure. P23 described feeling "*a bit rushed [...] if I just had more time, I would have been able to verify it properly*," while P7 reported prioritizing the storytelling task over verification, being "*more focused on how to tell the story instead of verifying the information.*" When sources were checked, 7/26 participants reported scanning rather than reading fully, with P22 noting scanning sources "*looking for keywords*" given limited time. Verification depth also varied by intended use (7/26). P18 noted that fictional brainstorming was "*not looking at factual accuracy at this point*," while P8 noted that production-level work "*would actually go to the original source.*"

3.4. Usability and Dataset Limitations

Our preliminary qualitative analysis documented two main usability concerns that participants reported. First, 12/26 participants found response density visually overwhelming, with P23 describing quotes as "*a big wall of text.*" Second, 14/26 participants found the system insufficiently conversational for creative collaboration, with P13 describing the system as "*more an index than an LLM that I can use to discuss ideas.*"

Dataset constraints emerged across coverage, recency, and accuracy. 14/26 participants noted that specific queries exceeded dataset coverage, with geographic underrepresentation concern were mentioned by 7/26 participants. 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.*" Content quality raised concerns, including factually misleading content (P13, P18, P24) and outdated sources (P8, P14). P18 flagged hydrogen-related content as "*misleading and not factually correct*," noting that source reports themselves may be inaccurate.

4. Discussion

4.1. What Signals Direct Verification Attention

Our results showed that source-match confidence scores served as an effective signal for directing user attention toward uncertain AI outputs. Interaction logs indicated that color-coded underlines accounted for 94.6% of citation hovers in the Verifiable condition, rather than citation badges, suggesting that color-underlines were the primary driver of verification engagement. Moreover, low-confidence underlines received significantly more hovers per underline than high-confidence ones, indicating that color-coding actively directed user attention toward uncertain content rather than simply adding clickable elements. Post-hoc validation confirmed that 94.0% of medium- and low-confidence quotes were indeed incorrect, suggesting that text matching provides a meaningful signal for source provenance, warranting user attention. Beyond directing attention, two participants reported that color-underlines served as reminders to fact-check AI outputs: P5 described color-coding as sustaining awareness that they were "*reading AI information*" rather than "*definite facts*." These results suggest that confidence-based signals served two roles in our study: guiding verification effort toward uncertain content, and maintaining critical engagement with AI outputs. These observations extend design for imperfection research [39, 40, 10] and contribute to ongoing work on addressing over-reliance concerns [6, 7, 8, 9, 5, 10, 11].

Beyond directing attention within individual responses, our results suggest that confidence signals influenced how some participants navigated the system more broadly. Participants reported using color cues to filter sources within responses and deprioritize low-confidence content. For some participants, this filtering extended to subsequent query strategies: P9 reported that color-underlines helped determine "*what the next question was going to be and what path I was going to follow*." When participants did proceed to verify sources under time constraints, click-through access supported rapid source inspection. Prior research documents that knowledge workers manually cross-reference AI-generated content against multiple external sources, consuming considerable time [32]. In our study, 9/26 participants found click-through access to sources helpful, and 10/26 valued direct links to exact source locations. P13 described side-by-side source presentation as reducing attention breaks, which occur when opening separate browser tabs for external source checking. These observations are consistent with prior findings that clickable sources can reduce over-reliance and help users identify inaccuracies [10, 38]. The significant increase in brief source inspections under 30 seconds in the Verifiable condition further suggests that glanceable cues triggered verification, and click-to-exact-location then supported source inspection without requiring users to leave their primary task.

However, user responses to confidence signals varied across individuals, task contexts, and time constraints, suggesting that the effectiveness of a single always-on signal design depends on contextual factors. At the individual level, while color-coded underlines converted some participants into active verifiers (e.g., P25 shifted from skipping verification in Baseline to actively using color cues), nine participants never opened a source document even when click-through access was available. This pattern may reflect what Vasconcelos et al. [6] characterize as a strategic decision, where users weigh perceived verification effort against potential accuracy gains. Time pressure further affects signal uptake: 11/26 participants reported skipping or deferring verification due to the 20-minute task duration, consistent with documented barriers in which users reduce verification effort under time constraints [31, 32, 33]. At the task level, participants reported less concern for factual accuracy during brainstorming than for production-level work, aligning with prior work documenting that creative professionals report greater concern about AI accuracy for tasks requiring factual consistency than for early-stage ideation [12].

These patterns have implications for the workshop question of *what* signals matter for GenAI intervention. Our system displays confidence signals for all generated quotes regardless of user state, operating as an always-on intervention rather than adapting to user needs. User-side signals observed in our study, including time pressure, task stage, and individual verification tendencies, suggest that future systems could move beyond always-on designs by using these contextual signals to modulate intervention strategy. For example, systems might increase signal salience when behavioral indicators suggest reduced verification engagement under time pressure, or reduce signal density during early-

stage brainstorming, where participants reported lower accuracy demands. However, our study was not designed to test adaptive intervention, and these directions remain speculative. Future work should examine how combining system-side confidence signals with user-side behavioral and contextual signals could support more timely verification interventions.

4.2. Evaluating the Effects of Verification Intervention

Our findings suggest a verification paradox that carries implications for how verification interventions are evaluated. Participants reported similar ease of verification across the two conditions, yet interaction logs showed significantly more verification engagement in the Verifiable condition, including a six-fold increase in citation hovers. This is paradoxical because we anticipated that increased verification engagement would correspond to higher perceived ease of verification, yet participants reported no significant difference. Interview data further show this disconnect: half of the participants (13/26) reported that color-underlines saved time on verification, with 6/26 noting that color-coding made confidence levels easy to see. As P9 noted, the traffic light color scheme leveraged culturally familiar visual conventions, allowing reliability assessments through quick visual scanning. Together, these results suggest that color-underlines prompted source checking that participants may not consciously recognize as verification, reflecting the core aim of glanceable design: supporting verification without disrupting primary task focus. This pattern aligns with prior work showing that color-coded highlighting can prompt users to seek additional information on flagged content, improving decision accuracy without affecting perceived reliability or experience ratings [46], and differs from cognitive forcing interventions that require independent judgment at the cost of less favorable subjective ratings [5]. For the workshop question of *how* to evaluate intervention timing, this paradox demonstrates that self-report measures alone can miss the behavioral effects of verification interventions, and future studies should incorporate behavioral measures alongside subjective ratings.

Evaluating verification interventions also requires capturing both benefits and costs simultaneously. In our study, the Verifiable condition showed higher trust ratings with increased verification engagement. However, information overload marginally increased ($p = .066$), and 12/26 participants found response density visually overwhelming. While color-underlines supported rapid reliability scanning, inline verbatim quotes can increase visual density, risking reintroduction of the verification burden that glanceability aims to reduce. Prior meta-analysis work also underscores that added AI transparency can increase cognitive load and complicate decision-making [50]. These trade-offs suggest that evaluation of verification interventions should not focus solely on effectiveness metrics (e.g., engagement, trust) but should also assess intervention costs (e.g., information overload, task disruption). For future design iterations, following the "design for imperfection" principle [39], verification support should make uncertainty visible to caution users about potential flaws, while providing on-demand access for source provenance evaluation, rather than presenting all verification information by default. One direction worth investigating involves applying color codes only to citation badges (e.g., ① ② ③) while revealing verbatim quotes only upon hover, reducing default visual complexity while preserving verification access.

A further evaluation nuance concerns whether verification interventions may inadvertently reduce verification depth. Eight participants reported that their trust in the system developed through actively checking sources, with P22 noting that repeated source verification progressively built confidence in the system's reference quality. However, P15 reported treating confidence scores as sufficient validation: "*I felt very confident [...] whatever it's giving me is quite verified*" and "*didn't walk through it very thoroughly*." While this observation comes from a single participant and should be interpreted cautiously, it aligns with a broader pattern documented in prior work: Bo et al. [11] observed that users can paradoxically increase their confidence following incorrect reliance choices, and Kim et al. [10] found that explanations can increase user reliance on both correct and incorrect responses. This suggests that evaluation for future verification interventions should assess not only whether users engage with signals, but also whether interventions may substitute for rather than prompt deeper checking

4.3. Limitations and Future Directions

Three limitations should be noted. First, our Verifiable condition combined color-coded click-to-source indicators with a 3D knowledge graph, making it difficult to isolate individual feature effects. However, engagement with the knowledge graph was limited (6/26 participants; 28 total clicks), as it was provided as an optional exploration feature without specific task guidance. Meanwhile, color-coded underlines accounted for 94.6% of citation hovers in the Verifiable condition, suggesting that the observed effects are more likely associated with the color-coded underlines. Future studies should isolate each feature to clarify individual contributions. Second, our 20-minute task reflected real-world time constraints but may have limited verification depth. The 2-minute walkthrough may also have been insufficient for participants to fully learn the verification features. Longer-term studies are needed to observe how verification behaviors evolve with sustained use. 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.

These limitations point to future directions relevant to the broader workshop agenda. Regarding *how to combine signals into a timing judgement*, our system uses a single system-side signal (confidence scores) displayed uniformly. Future work could explore combining confidence signals with user behavioral indicators (e.g., hover frequency, source view duration, query patterns) to determine when and how intensively to surface verification cues. Regarding *intervention timing*, our always-on design does not adapt to user or task state. The contextual factors identified in our study, including time pressure reducing verification engagement, task stage shifting accuracy demands, and individual differences in verification tendencies, suggest that adaptive timing of verification interventions could improve their effectiveness. However, these directions are speculative based on our current evidence, and require future empirical investigation to determine whether adaptive approaches outperform always-on designs in practice.

5. Conclusion

We investigated how glanceable verification indicators can support user verification in AI-assisted climate storytelling. In a within-subjects study ($n = 26$) with creative professionals, we compared a baseline condition to a verifiable condition that added color-coded confidence underlines. Our findings reveal a verification paradox: despite similar perceived ease of verification across conditions, the verifiable condition showed higher perceived trust ratings and a six-fold increase in citation hovers, suggesting that color-coded underlines prompted verification behaviors that participants may not consciously recognize as verification.

We contribute to the TimelyAI workshop discussion in two ways. First, regarding *what signals matter*, source-match confidence scores effectively directed user attention toward uncertain content, while user-side factors including time pressure, task stage, and individual verification tendencies affect signal uptake. These patterns suggest that future systems could move beyond always-on designs by combining system-side confidence signals with user behavioral and contextual signals to support more adaptive verification interventions. Second, regarding *evaluating intervention effects*, the verification paradox suggests that self-report measures alone may miss behavioral effects of verification interventions, and future evaluations should capture both benefits and costs.

However, our results are limited to a single domain with a 20-minute task, and we demonstrate increased verification engagement without confirming improved verification accuracy. Future work should isolate individual indicator effects, examine how verification behaviors evolve with sustained use, and explore whether adaptive intervention timing outperforms always-on designs in practice.

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.

Declaration on Generative AI

During the preparation of this work, the authors did not use any generative AI or AI-assisted technologies.

References

[1] S. Farquhar, J. Kossen, L. Kuhn, Y. Gal, Detecting hallucinations in large language models using semantic entropy, Nature 630 (2024) 625–630. URL: https://www.nature.com/articles/s41586-024-07421-0. doi:10.1038/s41586-024-07421-0.

[2] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, T. Liu, A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions, ACM Transactions on Information Systems 43 (2025) 1–55. URL: https://dl.acm.org/doi/10.1145/3703155. doi:10.1145/3703155.

[3] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, Survey of Hallucination in Natural Language Generation, ACM Computing Surveys 55 (2023) 1–38. URL: https://dl.acm.org/doi/10.1145/3571730. doi:10.1145/3571730.

[4] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, 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, 2021, pp. 610–623. URL: https://dl.acm.org/doi/10.1145/3442188.3445922. doi:10.1145/3442188.3445922.

[5] Z. Buçinca, M. B. Malaya, K. Z. Gajos, 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 (2021) 1–21. URL: https://dl.acm.org/doi/10.1145/3449287. doi:10.1145/3449287.

[6] H. Vasconcelos, M. Jörke, M. Grunde-McLaughlin, T. Gerstenberg, M. S. Bernstein, R. Krishna, Explanations Can Reduce Overreliance on AI Systems During Decision-Making, Proceedings of the ACM on Human-Computer Interaction 7 (2023) 1–38. URL: https://dl.acm.org/doi/10.1145/3579605. doi:10.1145/3579605.

[7] Y. Zhang, Q. V. Liao, R. K. E. Bellamy, 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, 2020, pp. 295–305. URL: https://dl.acm.org/doi/10.1145/3351095.3372852. doi:10.1145/3351095.3372852.

[8] G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, D. Weld, 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, 2021, pp. 1–16. URL: https://dl.acm.org/doi/10.1145/3411764.3445717. doi:10.1145/3411764.3445717.

[9] Z. Buçinca, P. Lin, K. Z. Gajos, E. L. Glassman, 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, 2020, pp. 454–464. URL: https://dl.acm.org/doi/10.1145/3377325.3377498. doi:10.1145/3377325.3377498.

[10] S. S. Y. Kim, J. W. Vaughan, Q. V. Liao, T. Lombrozo, O. Russakovsky, 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, 2025, pp. 1–19. URL: https://dl.acm.org/doi/10.1145/3706598.3714020. doi:10.1145/3706598.3714020.

[11] J. Y. Bo, S. Wan, A. Anderson, 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, 2025, pp. 1–23. URL: https://dl.acm.org/doi/10.1145/3706598.3714097. doi:10.1145/3706598.3714097.

[12] Y. Tang, H. Li, M. Lan, X. Ma, H. Qu, 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, 2025, pp. 1–18. URL: https://dl.acm.org/doi/10.1145/3706598.3714120. doi:10.1145/3706598.3714120.

[13] O. Ram, Y. Levine, I. Dalmedigos, D. Muhlgay, A. Shashua, K. Leyton-Brown, Y. Shoham, In-Context Retrieval-Augmented Language Models, Transactions of the Association for Computational Linguistics 11 (2023) 1316–1331. URL: https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00605/118118/In-Context-Retrieval-Augmented-Language-Models. doi:10.1162/tacl_a_00605.

[14] W. Fan, Y. Ding, L. Ning, S. Wang, H. Li, D. Yin, T.-S. Chua, Q. Li, 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, 2024, pp. 6491–6501. URL: https://dl.acm.org/doi/10.1145/3637528.3671470. doi:10.1145/3637528.3671470.

[15] K. Shuster, S. Poff, M. Chen, D. Kiela, J. Weston, Retrieval Augmentation Reduces Hallucination in Conversation, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 3784–3803. URL: https://aclanthology.org/2021.findings-emnlp.320/. doi:10.18653/v1/2021.findings-emnlp.320.

[16] X. Wang, Z. Wang, X. Gao, F. Zhang, Y. Wu, Z. Xu, T. Shi, Z. Wang, S. Li, Q. Qian, R. Yin, C. Lv, X. Zheng, X. Huang, 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, 2024, pp. 17716–17736. URL: https://aclanthology.org/2024.emnlp-main.981. doi:10.18653/v1/2024.emnlp-main.981.

[17] X. Li, J. Jin, Y. Zhou, Y. Zhang, P. Zhang, Y. Zhu, Z. Dou, From Matching to Generation: A Survey on Generative Information Retrieval, ACM Transactions on Information Systems 43 (2025) 1–62. URL: https://dl.acm.org/doi/10.1145/3722552. doi:10.1145/3722552.

[18] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, in: Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., Virtual, 2020, pp. 9459–9474. URL: https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

[19] Z. Guo, L. Xia, Y. Yu, T. Ao, C. Huang, LightRAG: Simple and Fast Retrieval-Augmented Generation, in: C. Christodoulopoulos, T. Chakraborty, C. Rose, V. Peng (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2025, Association for Computational Linguistics, Suzhou, China, 2025, pp. 10746–10761. URL: https://aclanthology.org/2025.findings-emnlp.568/. doi:10.18653/v1/2025.findings-emnlp.568.

[20] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, H. Wang, Retrieval-Augmented Generation for Large Language Models: A Survey, 2024. URL: http://arxiv.org/abs/2312.10997. doi:10.48550/arXiv.2312.10997.

[21] F. Cuconasu, G. Trappolini, F. Siciliano, S. Filice, C. Campagnano, Y. Maarek, N. Tonellotto, F. Silvestri, 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, 2024, pp. 719–729. URL: https://dl.acm.org/doi/10.1145/3626772.3657834. doi:10.1145/3626772.3657834.

[22] Y. Shao, Y. Jiang, T. Kanell, P. Xu, O. Khattab, M. Lam, Assisting in Writing Wikipedia-like Articles From Scratch with Large Language Models, in: K. Duh, H. Gomez, S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 6252–6278. URL: https://aclanthology.org/2024.naacl-long.347/. doi:10.18653/v1/2024.naacl-long.347.

[23] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text summarization branches out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013.pdf.

[24] T. Falke, L. F. R. Ribeiro, P. A. Utama, I. Dagan, I. Gurevych, Ranking Generated Summaries by Correctness: An Interesting but Challenging Application for Natural Language Inference, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 2214–2220. URL: https://aclanthology.org/P19-1213/. doi:10.18653/v1/P19-1213.

[25] M. Gao, X. Wan, DialSummEval: Revisiting Summarization Evaluation for Dialogues, in: M. Carpuat, M.-C. de Marneffe, I. V. Meza Ruiz (Eds.), Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 5693–5709. URL: https://aclanthology.org/2022.naacl-main.418/. doi:10.18653/v1/2022.naacl-main.418.

[26] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, C. Zhu, G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 2511–2522. URL: https://aclanthology.org/2023.emnlp-main.153/. doi:10.18653/v1/2023.emnlp-main.153.

[27] P. Wang, L. Li, L. Chen, Z. Cai, D. Zhu, B. Lin, Y. Cao, L. Kong, Q. Liu, T. Liu, Z. Sui, 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, 2024, pp. 9440–9450. URL: https://aclanthology.org/2024.acl-long.511. doi:10.18653/v1/2024.acl-long.511.

[28] N. Liu, T. Zhang, P. Liang, Evaluating Verifiability in Generative Search Engines, in: Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 7001–7025. URL: https://aclanthology.org/2023.findings-emnlp.467. doi:10.18653/v1/2023.findings-emnlp.467.

[29] A. Agrawal, M. Suzgun, L. Mackey, A. Kalai, Do Language Models Know When They're Hallucinating References?, in: Y. Graham, M. Purver (Eds.), Findings of the Association for Computational Linguistics: EACL 2024, Association for Computational Linguistics, St. Julian's, Malta, 2024, pp. 912–928. URL: https://aclanthology.org/2024.findings-eacl.62/.

[30] J. Zhang, M. Marone, T. Li, B. Van Durme, D. Khashabi, Verifiable by Design: Aligning Language Models to Quote from Pre-Training Data, in: L. Chiruzzo, A. Ritter, L. Wang (Eds.), 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), Association for Computational Linguistics, Albuquerque, New Mexico, 2025, pp. 3748–3768. URL: https://aclanthology.org/2025.naacl-long.191/. doi:10.18653/v1/2025.naacl-long.191.

[31] M. Kazemitabaar, J. Williams, I. Drosos, T. Grossman, A. Z. Henley, C. Negreanu, A. Sarkar, 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, 2024, pp. 1–19. URL: https://dl.acm.org/doi/10.1145/3654777.3676345. doi:10.1145/3654777.3676345.

[32] H.-P. H. Lee, A. Sarkar, L. Tankelevitch, I. Drosos, S. Rintel, R. Banks, N. Wilson, 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, 2025, pp. 1–22. URL: https://dl.acm.org/doi/10.1145/3706598.3713778. doi:10.1145/3706598.3713778.

[33] B. Yun, D. Feng, A. S. Chen, A. Nikzad, N. Salehi, 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, 2025, pp. 1–19. URL: https://dl.acm.org/doi/10.1145/3706598.3713337. doi:10.1145/3706598.3713337.

[34] P. Laban, J. Vig, M. Hearst, C. Xiong, C.-S. Wu, 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, 2024, pp. 1–23. URL: https://dl.acm.org/doi/10.1145/3654777.3676419. doi:10.1145/3654777.3676419.

[35] K. Gu, R. Shang, T. Althoff, C. Wang, S. M. Drucker, 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, 2024, pp. 1–22. URL: https://dl.acm.org/doi/10.1145/3613904.3642497. doi:10.1145/3613904.3642497.

[36] F. Leiser, S. Eckhardt, V. Leuthe, M. Knaeble, A. Mädche, G. Schwabe, A. Sunyaev, HILL: A Hallucination Identifier for Large Language Models, in: Proceedings of the CHI Conference on Human Factors in Computing Systems, ACM, Honolulu HI USA, 2024, pp. 1–13. URL: https://dl.acm.org/doi/10.1145/3613904.3642428. doi:10.1145/3613904.3642428.

[37] Z. Gu, I. Arawjo, K. Li, J. K. Kummerfeld, E. L. Glassman, 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, 2024, pp. 1–22. URL: https://dl.acm.org/doi/10.1145/3613904.3642699. doi:10.1145/3613904.3642699.

[38] M. Xu, X. Li, T. Lan, X. Zeng, T. Wang, S. Sheng, X. Deng, J. Sun, R. Fiebrink, N. Bryan-Kinns, M. Grierson, Comparing structural and conversational AI summarisation for time-constrained academic reading support among graduate students 0 (????) 1–23. URL: https://doi.org/10.1080/10494820.2025.2604649. doi:10.1080/10494820.2025.2604649.

[39] J. D. Weisz, J. He, M. Muller, G. Hoefer, R. Miles, W. Geyer, Design Principles for Generative AI Applications, in: Proceedings of the CHI Conference on Human Factors in Computing Systems, ACM, Honolulu HI USA, 2024, pp. 1–22. URL: https://dl.acm.org/doi/10.1145/3613904.3642466. doi:10.1145/3613904.3642466.

[40] U. Ehsan, Q. V. Liao, S. Passi, M. O. Riedl, H. Daumé, Seamful XAI: Operationalizing Seamful Design in Explainable AI, Proceedings of the ACM on Human-Computer Interaction 8 (2024) 1–29. URL: https://dl.acm.org/doi/10.1145/3637396. doi:10.1145/3637396.

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

[42] OpenAI, S. Agarwal, L. Ahmad, J. Ai, S. Altman, A. Applebaum, E. Arbus, R. K. Arora, Y. Bai, B. Baker, H. Bao, B. Barak, A. Bennett, T. Bertao, N. Brett, E. Brevdo, G. Brockman, S. Bubeck, C. Chang, K. Chen, M. Chen, E. Cheung, A. Clark, D. Cook, M. Dukhan, C. Dvorak, K. Fives, V. Fomenko, T. Garipov, K. Georgiev, M. Glaese, T. Gogineni, A. Goucher, L. Gross, K. G. Guzman, J. Hallman, J. Hehir, J. Heidecke, A. Helyar, H. Hu, R. Huet, J. Huh, S. Jain, Z. Johnson, C. Koch, I. Kofman, D. Kundel, J. Kwon, V. Kyrylov, E. Y. Le, G. Leclerc, J. P. Lennon, S. Lessans, M. Lezcano-Casado, Y. Li, Z. Li, J. Lin, J. Liss, Lily, Liu, J. Liu, K. Lu, C. Lu, Z. Martinovic, L. McCallum, J. McGrath, S. McKinney, A. McLaughlin, S. Mei, S. Mostovoy, T. Mu, G. Myles, A. Neitz, A. Nichol, J. Pachocki, A. Paino, D. Palmie, A. Pantuliano, G. Parascandolo, J. Park, L. Pathak, C. Paz, L. Peran, D. Pimenov, M. Pokrass, E. Proehl, H. Qiu, G. Raila, F. Raso, H. Ren, K. Richardson, D. Robinson, B. Rotsted, H. Salman, S. Sanjeev, M. Schwarzer, D. Sculley, H. Sikchi, K. Simon, K. Singhal, Y. Song, D. Stuckey, Z. Sun, P. Tillet, S. Toizer, F. Tsimpourlas, N. Vyas, E. Wallace, X. Wang, M. Wang, O. Watkins, K. Weil, A. Wendling, K. Whinnery, C. Whitney, H. Wong, L. Yang, Y. Yang, M. Yasunaga, K. Ying, W. Zaremba, W. Zhan, C. Zhang, B. Zhang, E. Zhang, S. Zhao, gpt-oss-120b & gpt-oss-20b Model Card, 2025. URL: http://arxiv.org/abs/2508.10925. doi:10.48550/arXiv.2508.10925.

[43] V. Asturiano, vasturiano/3d-force-graph, 2026. URL: https://github.com/vasturiano/3d-force-graph.

[44] L. Tang, P. Laban, G. Durrett, 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, 2024, pp. 8818–8847. URL: https://aclanthology.org/2024.emnlp-main.499. doi:10.18653/v1/2024.emnlp-main.499.

[45] J. Maynez, S. Narayan, B. Bohnet, R. McDonald, On Faithfulness and Factuality in Abstractive Summarization, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 1906–1919. URL: https://aclanthology.org/2020.acl-main.173/. doi:10.18653/v1/2020.acl-main.173.

[46] S. E. Spatharioti, D. Rothschild, D. G. Goldstein, J. M. Hofman, 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, ACM, Yokohama Japan, 2025, pp. 1–15. URL: https://dl.acm.org/doi/10.1145/3706598.3714082. doi:10.1145/3706598.3714082.
[47] R. R. Hoffman, S. T. Mueller, G. Klein, J. Litman, Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance, Frontiers in Computer Science 5 (2023) 1096257. URL: https://www.frontiersin.org/articles/10.3389/fcomp.2023.1096257/full. doi:10.3389/fcomp.2023.1096257.
[48] S. A. C. Perrig, N. Scharowski, F. Brühlmann, 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, 2023, pp. 1–7. URL: https://dl.acm.org/doi/10.1145/3544549.3585808. doi:10.1145/3544549.3585808.
[49] F. D. Davis, Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology, MIS Quarterly 13 (1989) 319. URL: https://www.jstor.org/stable/249008?origin=crossref. doi:10.2307/249008.
[50] D. Göndöcs, S. Horváth, V. Dörfler, Uncovering the dynamics of human-AI hybrid performance: A qualitative meta-analysis of empirical studies, International Journal of Human-Computer Studies 205 (2025) 103622. URL: https://www.sciencedirect.com/science/article/pii/S107158192500179X. doi:10.1016/j.ijhcs.2025.103622, gSCC: 0000019 2026-04-04T21:51:50.021Z 0.83.

APPENDIX

A. Formative Interview Questions

Opening Script: *Thank you for joining us today. This interview will explore how you currently use research literature in your script editing process and gather your thoughts on how AI might help this process. Your insights will directly inform the design of our system.*

A.1. Background Questions

- How many years have you been a script editor?
- How many years have you worked on climate storytelling projects?

A.2. Current Workflows and Literature Use

- What does your current process for script editing on climate elements in TV scripts look like?
- Do you currently use research literature in your script editing work? If so, how do you use it?
- Do you already use any tools to help you navigate research literature?
- What are the most common issues you encounter when using research literature in script editing on climate elements?
- How frequently do these issues arise—for example, how many times did they occur during your last experience on using research literature in script editing on climate elements?
- What is the impact of these issues—in terms of additional time or costs for you or your team?

A.3. Future Thinking of Scenarios

Note: At this point, participants were shown a reference slide introducing a research database [41], covering topics such as Business Rewards, Business Risks, Audience Insights, Industry Insights, and Effects/Impact on Audiences.

A.3.1. System Vision & Use Cases

Imagine an AI system designed to help you navigate climate storytelling research literature. How might such a system support your script editing process?

- **Data Sources:** What kind of research literature would you want to access through such a system (e.g., academic articles, news articles, reports, blog posts)? What data sources are important to you?
- **Usage Scenarios:** How would you envision using this tool (e.g., what questions or prompts might you ask)?
- **Workflow Integration:** Would such an AI tool beneficially fit into your workflows? If so, how?
- **Other Beneficiaries:** Who else might benefit from this tool, and why?
- **Evaluate:** How would you evaluate the quality of an AI-generated response (e.g., helpfulness, relevance, accuracy, depth, and level of detail)?
- **Concerns:** What barriers or concerns would you have around trusting such AI tools?

A.3.2. Trust & Transparency

What would make such an AI tool trustworthy/understandable for you?

- **System Understanding:** What would you want to know about how the system works overall?
- **Source Selection:** Why does the AI system select specific research literature sources over others?
- **Reliability:** How important is knowing the system's accuracy, reliability, or typical mistakes?
- **Key Questions:** What questions does the tool need to answer to be trustworthy?

A.3.3. Interaction & Output

- What would you expect the output or feedback from such a system to look like?
- Similar to ChatGPT, or is there another tool you like?

A.3.4. Bigger Picture AI Discussion

- **Current AI Integration:** Where / how are you already using Generative AI in your everyday work?
- **AI Concerns:** What are your concerns about the use of Generative AI?

B. Story Assistant System Prompt

Modified from LightRAG's official prompt template [19]. We restricted citations to Document Chunks only and enforced blockquote (>) formatting to enable automated quote extraction and text matching for confidence score calculation.

```
---Role---

You are a helpful assistant providing detailed, comprehensive responses. Use ONLY the "content" field from
----Document Chunks(DC)----- for quotes and citations. NEVER quote from "description" fields in
---Entities(KG)-----. Ignore Knowledge Graph for citation purposes.

---Goal---

Generate a comprehensive response based on the provided Document Chunks. Each bullet point MUST contain
2-4 sentences of detailed explanation before the supporting quote. Do not invent information not present
in the Document Chunks.

---Conversation History---
{history}
```

```
---Knowledge Graph and Document Chunks---
{context_data}

---RESPONSE GUIDELINES---
**1. Content & Adherence:**
- Provide comprehensive responses with 2-4 sentences of explanation per bullet point. Include context,
implications, and connections from the Document Chunks.
- CRITICAL: Quote ONLY from "content" text in ----Document Chunks(DC)-----. NEVER quote from "description"
in ---Entities(KG)-----.
- Use ONLY Document Chunks for quotations and citations. Do NOT cite Knowledge Graph entries.
- If you ONLY partially relevant or indirect information in the Document Chunks that could contribute to
an answer, use it to provide the most helpful response possible. Then acknowledge the limitation while
still being informative.
- ONLY when there is absolutely NO relevant or related information in the Document Chunks that could
contribute to an answer, state "Sorry, the dataset does not have enough information to answer this
question". Then explain why the dataset are insufficient for the user's query.

**2. Formatting & Language:**
- Structure: Use markdown headings (### Section Title) followed by bullet points (-).
- Each bullet point TWO-LINE format:
  - **Line 1**: 2-4 explanation sentences providing context, significance, and connections.
  - **Line 2**: MUST start with ONE > blockquote symbol, followed by verbatim quoted text in double quotes
(""), ending with citation number ONLY in format [n].

Example:
### Section Title
- Your detailed explanation (2-4 sentences) explaining the concept, providing context, and highlighting
significance or connections to broader themes.
  > "exact quoted text from document" [1]

- CRITICAL: Every quote MUST begin with > symbol. Copy quotes exactly without edits. Use sequential
citations [1], [2], [3].
- Write in the same language as the user's question.

---CRITICAL SOURCE MATCHING RULE---
When you quote text from a Document Chunk, you MUST:
1. Locate the exact "content" field containing that verbatim text
2. Find the "file_path" metadata associated with that specific Document Chunk
3. Use that exact file_path for the citation reference
Each quote [n] MUST link to the file_path of the Document Chunk where that quote actually appears.

**3. Citations / References:**
- CRITICAL: For each quote [n], trace back to the specific Document Chunk containing that exact text, then
use its file_path.
- VERIFICATION STEP: Before finalizing, confirm each quoted text appears in the Document Chunk whose
file_path you are citing.
- End with a "References" section listing all cited sources:
  - [1] 001_2022_filename.md
  - [2] 102_2023_filename.md
  - [3] 243_2020_filename.md
- Copy file_path character-by-character from the Document Chunk containing the quote (ID_YEAR_TITLE
format).

---USER CONTEXT---
- Additional user prompt: {user_prompt}

Response:
```

C. Confidence Score Calculation

The system computes confidence scores through cascading text matching (Algorithm 1).

Algorithm 1 Confidence Score Calculation

Require: quote Q, source document S

Normalize both texts (remove markdown, standardize whitespace)
Exact match: if Q found verbatim in $S \rightarrow$ return 1.0
Fuzzy match: tokenize into words, compute overlap ratio
  $confidence \leftarrow$ |matched words|/|quote words|
  if $confidence \geq 0.4 \rightarrow$ return confidence
Keyword match: count keywords (>4 chars) found in S
  if $\geq$3 keywords match $\rightarrow$ return ratio (typically <0.4)
return 0.0 (no match)

D. User Study Task Briefs

D.1. Brief A: Extreme Heat in Europe

Background: You are developing a climate story with the starting concept: "**Europe is hit by a record-breaking heatwave during a detective thriller.**"

You have [Interface A/B] to assist you in developing this concept further. [Interface A/B] is an AI storytelling editor assistant that accesses over 150 pieces of climate communications research.

Your task (20 minutes): Use the AI tool as a resource to explore how you could develop this story concept further. How you use the tool is up to you. Some possible directions you might consider:

- You could explore creative ideas and story lines to see how they align with research on effective climate narratives.
- See what research and case studies there are
- Ask more general questions around industry approaches to handling environmental issues
- Explore how climate themes have been successfully integrated into existing entertainment content
- Learn how audiences' perceptions and behaviours change when they encounter climate storytelling in media and entertainment

Feel free to take the exploration in whatever direction seems most valuable to you.

Deliverable: Explore and play with [Interface A/B]. Then provide 2 or more story suggestions on how you could develop this story further based on what you have learned from [Interface A/B].

D.2. Brief B: Europe Leads Climate Solutions

Identical to Brief A, with the alternative starting concept: "**A near-future story where Europe becomes the dominant climate solution leader.**"

E. Post-Task Questionnaire

Participants completed the following questionnaire after using each interface condition (Interface A and Interface B). All items used Likert scales as specified below. Before presenting the questionnaire items, participants received the following instructions:

Considering your recent experience with [Interface A/B], please respond to the following items.

E.1. Perceived Usefulness and Ease of Use

Response scale: 1 (Unlikely) to 7 (Likely).

Perceived Usefulness:

- Using [Interface A/B] in my work-related task would enable me to accomplish tasks **more quickly**.

- Using [Interface A/B] would **improve my work-related task performance**.
- Using [Interface A/B] in my work-related task would **increase my productivity**.
- Using [Interface A/B] would **enhance my effectiveness** on the work-related task.
- Using [Interface A/B] would **make it easier** to do my work-related task.
- I would find [Interface A/B] **useful** in my work-related tasks.

Perceived Ease of Use:

- **Learning to operate** [Interface A/B] would be easy for me.
- I would find it easy to get [Interface A/B] **to do what I want it to do**.
- My interaction with [Interface A/B] would be **clear and understandable**.
- I would find [Interface A/B] to be **flexible** to interact with.
- It would be easy for me to **become skillful** at using [Interface A/B].
- I would find [Interface A/B] **easy to use**.

E.2. Trust Scale for Explainable AI

Response scale: 1 (Strongly Disagree) to 7 (Strongly Agree).

- I am **confident** in the tool. I feel that it **works well**.
- The outputs of the tool are very **predictable**.
- The tool is very **reliable**. I can count on it to be **correct** all the time.
- I feel **safe** that when I rely on the tool, I will get the **right answers**.
- The tool is **efficient** in that it works very quickly.
- The tool can perform the task **better than a novice** human user.
- I like using the system for **decision making**.

E.3. Verification and Control

Response scale: 1 (Very Difficult) to 7 (Very Easy).

- How **easy** was the AI tool for solving the tasks?
- How easy was it to **verify** and **validate** the AI's responses?
- How easy was it to **intervene** and **fix** the AI whenever it was doing something wrong?
- How easy was it to **steer** the AI towards your desired solution?

Response scale: 1 (Not at All) to 7 (Extremely).

- To what extent did you feel in **control** of the AI's analysis process?
- How **successful** do you feel you were in achieving accurate and satisfactory results using the AI?
- When working on the tasks, how **frustrated** did you feel?
- To what extent did you feel **overwhelmed** by the amount of **information** displayed in the UI and options that you had?

F. User Study Interview Questions

Opening Script: *Thank you for joining us today. This interview will gather your thoughts on your experience with [Interface A/B] during the task. Your insights will directly inform the design of our system.*

F.1. Opening

- Could you walk me through how you used [Interface A/B] during the task?
- What were you hoping to accomplish when you started?

F.2. System Usage and Workflow

- How did you decide what questions to ask [Interface A/B]?
- Can you describe how you used the AI's responses in your work?
- Which features did you find yourself using most? What drew you to those?

F.3. Verification and Trust

- When you saw information from [Interface A/B], how did you decide whether to trust it?
- Could you walk me through what you actually did when checking sources?
- How confident did you feel about verifying [Interface A/B]'s answers? What influenced that feeling?

F.4. System Experience

- What aspects worked well for you?
- What felt frustrating or unclear?
- If you could improve one thing, what would make the biggest difference?

F.5. Closing

- Can you see yourself using [Interface A/B] in the future? What would make you more or less likely to?

Participants were then asked if they had any additional comments before the interview concluded.