RESEARCH ARTICLE

Addressing Uncertainty in LLM Outputs for Trust Calibration Through Visualization and User Interface Design

Helen Armstrong [a], **Ashley L. Anderson** [a,b], **Rebecca Planchart** [a], **Kweku Baidoo** [a], **and Matthew Peterson** [a]

[a] Department of Graphic Design and Industrial Design, North Carolina State University, Raleigh, NC, USA; [b] School of Visual Arts, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA

Corresponding author: Helen Armstrong (hsarmstr[at]ncsu.edu)

Abstract: Large language models (LLMs) are becoming ubiquitous in knowledge work. However, the uncertainty inherent to LLM summary generation limits the efficacy of human-machine teaming, especially when users are unable to properly calibrate their trust in automation. Visual conventions for signifying uncertainty and interface design strategies for engaging users are needed to realize the full potential of LLMs. We report on an exploratory interdisciplinary project that resulted in four main contributions to explainable artificial intelligence in and beyond an intelligence analysis context. First, we provide and evaluate eight potential visual conventions for representing uncertainty in LLM summaries. Second, we describe a framework for uncertainty specific to LLM technology. Third, we specify 10 features for a proposed LLM validation system — the Multiple Agent Validation System (MAVS) — that utilizes the visual conventions, the framework, and three virtual agents to aid in language analysis. Fourth, we provide and describe four MAVS prototypes, one as an interactive simulation interface and the others as narrative interface videos. All four utilize a language analysis scenario to educate users on the potential of LLM technology in human-machine teams. To demonstrate applicability of the contributions beyond intelligence analysis, we also consider LLM-derived uncertainty in clinical decision-making in medicine and in climate forecasting. Ultimately, this investigation makes a case for the importance of visual and interface design in shaping the development of LLM technology.

Implications for practice: This article provides guidance on explainability and transparency for AI interface design through the consideration of uncertainty in LLM summaries. Our Uncertainty Framework for Explainable Summaries (UFES) can guide system design and help users interpret

@: ISSUE〉 ARTICLE〉

Cite this article:

Armstrong, H., Anderson, A. L., Planchart, R., Baidoo, K., & Peterson, M. (2025). Addressing uncertainty in LLM outputs for trust calibration through visualization and user interface design. *Visible Language, 59*(2), 176–217.

First published online August 15, 2025.

https://visible-language.org/journal/

Visible Language Consortium:
University of Leeds (UK)
University of Cincinnati (USA)
North Carolina State University (USA)

and act on LLM outputs. Our specifications for 10 features in a Multiple Agent Validation System (MAVS) can be implemented with current technology to aid user understanding, trust calibration, and decision-making. As an open resource, we provide eight visualization options with readable code that represent uncertainty within relevant passages of text. We also include four prototypes of MAVS to reference for implementation, or to educate stakeholders on the value of LLMs when carefully leveraged. While we situate this guidance in an intelligence analysis context, outcomes are relevant to any LLM systems that produce summaries of information.

Statement of applicability: This is a visual and interface design investigation in the context of intelligence analysis (specifically language analysis). It is relevant to any human-centered application of LLM technology that explicitly addresses uncertainty in outputs.

Application domains: intelligence analysis, knowledge work, computer science (LLM technology and RAG), psychology, and human-computer interaction; connections are also drawn to LLM use in medicine (diagnosis) and climate science (forecasting).

Keywords: explainable AI; human-machine teaming; intelligence analysis; large language models; trust calibration; uncertainty; user interface design; visual representation

1. Introduction

Researchers and practitioners increasingly rely on generative artificial intelligence (AI) systems as essential tools for navigating information-dense environments. When faced with vast quantities of information, humans can deploy large language models (LLMs) — AI systems trained to understand and generate human language — to efficiently summarize content. But these summaries are not foolproof. LLM summaries will always include a level of uncertainty. And we do not yet have a conventional means for understanding this uncertainty or for presenting it to users. Without this, users cannot properly calibrate their trust in AI systems, leaving the full potential of human-machine teaming unrealized.

This article addresses the need to visually signify the uncertainty specific to LLM summaries, to guide that signification with an uncertainty framework, and to situate the resulting signification and explanation using interface design strategies. We do this by reporting on a 12-month design investigation that examines uncertainty in LLM summaries for the intelligence community, funded by the United States Department of Defense and focused on language analysis — analysis where source information, or *intelligence traffic,* primarily takes the form of transcribed or written language. This investigation was a collaboration with the Laboratory for Analytic Sciences (LAS), which supports the advancement of technology and tradecraft relevant to the

mission of the United States intelligence community. Among other research areas, LAS focuses on human-machine teaming, particularly the optimal utilization of automated systems by intelligence analysts. This high-stakes, security-critical space demands exceptional precision, making it an ideal test environment for AI explainability and LLM uncertainty.

The investigation began as an exploration of uncertainty visualization and expanded to encompass a proposal for an interactive system with multiple LLM agents that assist language analysts in summary validation. The resulting validation process enables critical trust calibration between analysts and the AI system. It is directly relevant to LLM use by all knowledge workers beyond the intelligence context, and further, to any situation that is dependent upon LLM summarization for decision-making. The general contributions of this investigation, as articulated in this article, are:

1. An open resource of implementable visual conventions for representing uncertainty, with criteria for selecting among them (Section 4).
2. A framework for uncertainty in LLM summaries (Section 5).
3. Design specifications for a Multiple Agent Validation System (MAVS) to empower knowledge workers while helping them calibrate trust in AI, including a 10-element feature set (Section 6).
4. Rich prototypes for alternative versions of MAVS. The prototypes include a simulation interface that allows users to utilize MAVS in a scripted scenario (Section 7.1) and three narrative interfaces that enrich the MAVS concept by envisioning dynamically reconfigurable user interface implementations (Section 7.2).

Further special contributions for the design community are:

5. The investigation's overall process, which is a model that can readily be adapted for any design investigations that involve collaboration with non-design experts and that center visual exploration in their activities (Section 3).
6. Similarly adaptable processes for developing a framework (Section 5.1), producing a range of visual studies (Section 4), and designing speculative interfaces with a STEM focus on implementation rather than as a mode of criticism (Section 7).

Finally, special contributions for other communities of practice — medicine and climate science communities — are:

7. Explicit declarations regarding how this work may be applied in medical diagnosis and climate forecasting, as examples of application beyond the intelligence analysis space (Section 8).

2. Literature Review

Language analysts in the intelligence community have only recently begun exploring how they might utilize LLM summaries in their workflow, shifting from direct database queries to increasingly relying on more opaque LLM processes of retrieval, analysis, and summarization. This emergent human-machine teaming can augment human cognitive abilities and leverage human and AI capabilities. AI-assisted human decision-making has the potential to outperform full automation in the national security sector, as in other critical sectors such as medicine, law, financial services, and law enforcement (Tomsett et al., 2020; Zhang et al., 2020). However, the probabilistic nature of AI models — leading to fundamental levels of uncertainty — necessitates human oversight in critical scenarios. The uncertainty specific to LLMs makes it difficult for analysts to gauge information reliability. When users cannot fully grasp how automated systems work — particularly in complex scenarios where comprehensive technical knowledge is impractical — their willingness to rely on the systems depends heavily on trust. If uncertainty and vulnerability were not factors, trust would not be necessary (Lee & See, 2004).

Establishing and maintaining human trust of AI systems is quite challenging in a high-stakes environment. Inappropriate levels of trust between users and AI systems often lead to misuse (as overreliance) or disuse (as underreliance) of automation (Lee & See, 2004). Dietvorst et al. (2016) frame underreliance as *algorithm aversion*. Humans tend to trust algorithms until they detect that they are imperfect — and all algorithms are imperfect — at which point they may avoid the algorithms, bypassing them. Alternatively, if an interface gives users some control over an AI's predictions — and it can be very limited control — humans will have a greater tendency to use the AI, or to overcome algorithm aversion (Dietvorst et al., 2016). Whether due to overreliance or underreliance, miscalibrated trust diminishes the benefits of using an AI system. Overtrusting is particularly problematic when humans reinforce their own negative societal biases (Stevenson, 2018; Suresh et al., 2020). The AI system can become a convenient excuse for problematic recommendations rather than augmenting and improving human decision-making. Lee and See (2004) have noted that the diminishment of trust through system misuse and disuse is a closed-loop process: "If the system is not trusted, it is unlikely to be used; if it is not used, the operator may have limited information regarding its capabilities, and it is unlikely that trust will grow" (p. 68). So how might we interrupt the loop of diminishing trust?

User interface design is one possible answer to the question of trust calibration. Appropriate trust calibration occurs when a user's trust in an automated system corresponds accurately with the system's capabilities (Lee & See, 2004). Achieving appropriate trust calibration can produce superior human-machine performance

(Sorkin & Woods, 1985; Wickens et al., 2000). Specific interface features have been proposed that might support the appropriate calibration of trust (e.g., Corritore et al., 2003; Cummings, 2006). Borgo et al. (2024) synthesized 40 related papers to provide nine claims about the impact of interface design choices on perceived trustworthiness. Four of these claims are particularly relevant here:

1. "AI transparency, intelligibility, or explainability fosters trust,"
2. "Communicating uncertainty fosters trust,"
3. "Adding interactivity fosters trust," and
4. "Social factors influence trust" (Borgo et al., 2024, pp. 23–24).

AI transparency, intelligibility, or explainability fosters trust. Interface features that empower users to verify results through access to and investigation of original sources produce transparency for otherwise opaque AI models (Borgo et al., 2024; Sultanum et al., 2019), positively impacting trust (Dasgupta et al., 2017; Krueger et al., 2020; Sperrle et al., 2021). Sultanum et al. (2019) explain that linking back to original sources enables users to analyze and compare source material with the LLM output. The resulting analysis offers insight into the LLM's process, which helps the user more clearly delineate the boundaries of system capabilities and understand its outputs. Methods of delineation have frequently been addressed in the literature, although researchers disagree on the details (Bansal et al., 2019; Bellotti et al., 2001; Borgo et al., 2024; Doshi-Velez & Kim, 2017; Sultanum et al., 2019; Tintarev & Masthoff, 2007; Weisz et al., 2024).

Communicating uncertainty fosters trust. Many researchers point to the role of uncertainty awareness in trust calibration (Amershi et al., 2019; Bansal et al., 2019; Kocielnik et al., 2019; Tomsett et al., 2020). A user's trust in a system correlates directly with how well the user understands its underlying uncertainties (Sacha et al., p. 76). An important factor for such understanding is user comprehension of what a system does not or cannot know (Tomsett et al., 2020). Once aware of output uncertainties the user can more quickly form an accurate mental model of the system's true capabilities (Tomsett et al., 2020, p. 2). Borgo et al. (2024) suggest that the user interface should clearly display the uncertainties and limitations inherent in a system's data and results. Essential information about uncertainty should be prioritized to address human cognitive limitations (Alhadad et al., 2018; Baldassi et al., 2006), and the design of the interface should carefully direct users' attention (Shneiderman, 1996; Rosenholtz et al., 2007).

Padilla et al. (2018) consider practical strategies for reducing cognitive load during decision-making with visualizations. They recommend that designers focus on prioritizing and hierarchically structuring information: "Identify the critical information needed for a task and use a visual encoding technique that directs participants' attention to this information" (p. 22). Alhadad et al. (2018) suggest several strategies for directing attention and reducing cognitive strain through visualizations, including coherence,

chunking, contiguity, segmenting, and signaling. These recommendations point to the role visual and interaction design play in focusing the user on vital information needed to make decisions.

Dual process theory (Evans & Stanovich, 2013; Padilla et al., 2018) breaks decision-making into two types of processing: humans first make simple, lightweight decisions (Type 1 processing) before moving on to complex, demanding, laborious decisions (Type 2 processing). The dual process approach aligns with established user experience (UX) design principles, such as progressive disclosure through layered interfaces, the principle that UX should reveal increasingly complex data to users in stages or layers (Forsey et al., 2024; Joshi et al., 2017). Designers should leverage such strategies to support both Type 1 and Type 2 processing to impact trust calibration through sustained interaction. Simply visualizing uncertainty is not enough.

Adding interactivity fosters trust. Borgo et al. (2024) also emphasize that interactive features can build trust by enabling users to test and verify system behavior; to customize outputs to better serve their needs; and to contribute domain expertise to improve performance. Hands-on interaction helps users understand a system's capabilities and limitations, allowing them to better predict its behavior across different scenarios (p. 26). Accurately predicting system behavior across scenarios is key to successful trust calibration. A user's ability to predict such behavior affects their own tendencies to either engage or disengage with AI.

Social factors influence trust. Personifying an AI system as a virtual agent can foster trust (Weisz et al., 2024), particularly when the interface combines modalities such as speech, voice, and visual presence (Rheu et al., 2021). Nass and Brave (2005) argue that humans instinctively process artificial voices like human ones — a natural, social response that makes voice interfaces effective tools for building trust when designed to mimic human interaction patterns. Graaf and Malle (2017) showed that virtual agents provide an effective avenue for fostering user trust because users attribute human-like intentions and reasoning to AI systems. To fulfill their potential, virtual agents must be able to explain system actions, or else systems will remain opaque to users and mistrust is likely to develop (Williams et al., 2015). If virtual agents offer insightful explanations and exhibit human behavioral and linguistic patterns, users of AI systems can more easily form accurate mental models to guide system use and decision-making.

The literature thus provides guidance for addressing trust calibration. It supports these three key points:

1. Human-machine teaming can produce better results in critical decision-making spaces than human or AI working alone (Tomsett et al., 2020; Zhang et al., 2020; Zhao et al., 2023).

2. Appropriate trust calibration is key to successful human-machine teaming (Lee & See, 2004; Sorkin & Woods, 1985; Stevenson, 2018; Suresh et al., 2020; Wickens et al., 2000).
3. Interface design can be used to effectively communicate AI capabilities and thus support trust calibration (Borgo et al., 2024; Corritore et al., 2003; Cummings, 2006).

Furthermore, the literature suggests fundamental interface design strategies for facilitating trust calibration (TC):

- *TC1: Transparency.* Support trust calibration by enabling users to verify and interrogate LLM outputs with interface features such as direct source investigation (Bellotti et al, 2001; Borgo et al., 2024; Dasgupta et al., 2017; Doshi-Velez & Kim, 2017; Krueger et al., 2020; Sperrle et al., 2021; Sultanum et al., 2019; Tintarev & Masthoff, 2007; Weisz et al., 2024).
- *TC2: Visualization.* Support trust calibration by visualizing uncertainty to communicate limitations inherent to the LLM's data and results (Amershi et al., 2019; Banshal et al., 2019; Borgo et al., 2024; Kocielnik et al., 2019; Sacha et al., 2015; Tomsett et al., 2020).
- *TC3: Alignment.* Support trust calibration by enabling users to interact with visualizations in alignment with human decision-making processes (Alhadad et al., 2018; Baldassi et al., 2006; Evans & Stanovich, 2013; Kirschner et al., 2011; Padilla et al., 2018; Rosenholtz et al., 2007; Shneiderman, 1996).
- *TC4: Interactivity.* Support trust calibration by enabling users to affect system results through a range of explicit user interactions, including but not limited to user settings (Borgo et al., 2024; Dietvorst, 2016; Lee & See, 2004).
- *TC5: Virtual Agents.* Support trust calibration by enabling users to conceptualize model functionality and seek explanation through multiple AI agents (Borgo et al., 2024; Graaf & Malle, 2017; Nass & Brave, 2005; Rheu et al., 2021; Weisz et al., 2025; Williams et al., 2015).

We will revisit TC1–TC5 as we consider specific user interface features for representing and explaining uncertainty in LLM summaries.

3. Investigation Process

Early in this investigation we focused on developing a wide range of visual conventions for representing uncertainty in LLM summaries. This early exploratory work soon shifted to a more convergent, evaluative phase in which the extended multidisciplinary collaborative team — language analysts, computer scientists, and psychologists informing design researchers — narrowed the options down from approximately 150 to eight. In

parallel with these visual explorations, we began adapting an existing framework for uncertainty (Skeels et al., 2010) but ultimately developed a new framework for classifying uncertainty specific to LLM summaries.

In a parallel phase of the project, we considered how language analysts might interact with visualizations of uncertainty. We realized quickly that if analysts did not appropriately trust the information represented by visualizations, they would not use that information — in which case representational quality would be irrelevant. To address appropriate trust calibration, we drew from UX findings established in an earlier project with the Laboratory for Analytic Sciences (LAS, 2024), as well as the trust in automation literature. Through these efforts we gleaned five trust calibration interface strategies appropriate to intelligence analysis (Section 2).

Using a persona, scenarios, task flows, and interface strategies as a starting point, we decided to develop an interactive simulation interface, both as a concept generator for ourselves and as an educational tool for language analysts who use AI in their tradecraft. We pinpointed 10 core system features to address trust calibration within this simulation and cohered them into an LLM validation system concept. While the interactivity of the simulation enables analysts to directly experience the proposed system in full interaction fidelity, the requisite development time limited our own formative design exploration of the proposed system's potential. To overcome this limitation, we pivoted to additional scenario video prototyping. The three resulting narrative interfaces involved no backend development, resulting in a more nimble iterative process that reflected our evolving understanding of uncertainty and trust. These narrative interfaces envision future possibilities for trust calibration in dynamically reconfigurable user interfaces. This dual method of developing key interface features within both current and future interface structures allowed us to pivot in response to expert assessments, while permitting the lateral movements typical of design exploration.

The orderliness of this description and this section title's implicit suggestion of a singular process are both potentially misleading. As shown in Figure 1, we engaged in a 12-month discovery process that, through sustained interaction with our partners outside of design and through our own sensemaking-through-design, was neither orderly nor predictable. The project coordination strip in Figure 1 lists four key planning moments for our project. It was not until we were past the quarter mark that it became possible to create a "plan-out," the first plan that envisioned the project through completion. We knew from experience that early efforts would need to be fulfilled before later efforts could be specified. These plans are visualized in Figure 1 as ripples that affected the main investigation outcomes because each plan synthesized collective sensemaking and suggested adjustments to all ongoing tasks.

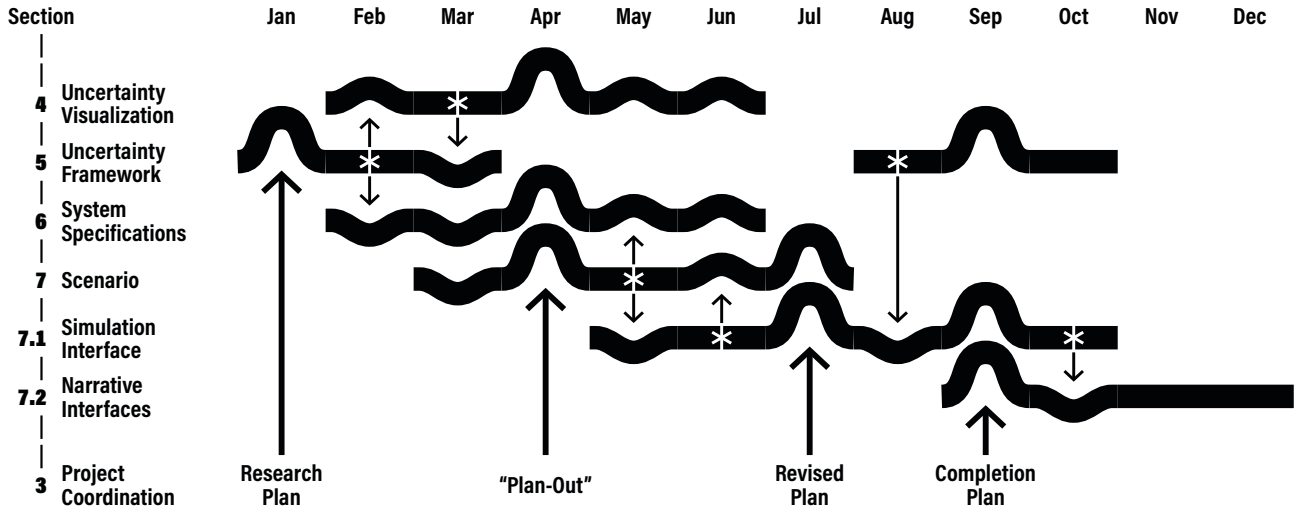

Figure 1. Investigation process for the project "Developing Visual Conventions for Explainable LLM Outputs in Intelligence Analysis Summaries," conducted January ("Jan") through December. Section numbers are references internal to this article. Ripples represent impacts spreading from one investigation component to others.

Likewise, our work toward individual outcomes frequently caused ripples across parallel tasks. Figure 1 depicts these ripples as smaller than those created by the comprehensive plans, but they were crucial nonetheless. As depicted, five formative investigation contributions caused ripples:

1. Uncertainty visualization (Section 4)
2. The uncertainty framework (Section 5)
3. LLM validation system specifications (Section 6)
4. A scenario encapsulated in the prototypes (Section 7)
5. The simulation interface (Section 7.1)

A sixth contribution was summative, informed by all previous work:

6. The narrative interfaces (Section 7.2)

It was thus through an immersive and messy process that this investigation took form. We now address the main investigation contributions in turn.

4. Uncertainty Visualization

To discover visual conventions for uncertainty in LLM summaries, we utilized design exploration and co-design, sharing progress in biweekly sessions with the extended multidisciplinary collaborative team. This meeting structure encouraged team participation regarding the formal qualities of the visual studies, their effectiveness in

conveying uncertainty, and the broader implications for intelligence analysis workflows and decision-making processes. We gathered feedback through interviews, surveys, and informal user testing. To ensure flexibility for incorporating visual conventions into summarized LLM outputs, we organized exploratory studies into three uncertainty cue locations: *inline,* embedded directly within the summary text itself; *interstitial,* positioned in the spaces between lines of text; and *adjacent,* appearing outside of the summary. To maximize variation, we did not initially concern ourselves with practicality, but we later removed all impractical studies from consideration. We created what we called "concept zero," a simplified interface prototype for situating visual conventions. In the midst of our exploration we adopted an accessible color palette from the IBM Design Language to address Web Content Accessibility Guidelines. Ultimately, we generated approximately 150 visual studies for conveying uncertainty in LLM summaries.

Discussions with our collaborators also produced criteria to guide future work that seeks to establish a singular visual convention for representing uncertainty. Using these criteria, researchers could stage empirical studies relating visual conventions to mental models of AI and to user preferences.

- *Experiential.* Visualizations of uncertainty should provide a sense of severity at a glance, with text subjectively *feeling* as uncertain as it has been deemed to be.
- *Reflective.* Visualizations of uncertainty should make sense upon reflection, ideally aligning with an accurate or at least useful mental model of uncertainty. This accuracy or utility will increase educational impact.
- *Legible.* Uncertain text should be readable so that it can be validated by users, though legibility can be dynamically variable if users are permitted to inspect text passages, in which case a responsive system can make text clearer.
- *Implementable.* The display of visualizations needs to be technically feasible. We have favored common web display technology as a gauge of implementation feasibility, including considerations of which visual conventions utilize prescribed display functions (e.g., blur in CSS) and which require workarounds (e.g., background images to approximate inline display functions that do not exist in CSS).

Using these four criteria, we derived 12 options from the collection of visual studies, each of which we implemented in CSS (in the simulation interface, Section 7.1). In consultation with our collaborators, we deactivated four of these options for a total of eight potential visual conventions for uncertainty. These are reproduced in Figure 2. We evaluated these visual conventions according to the semiotic modes by which they operate, the conceptual implications of those modes, and the patterns of criteria

fulfillment across the set. Semiotic modes and conceptual implications are documented in Table 1.

- *Semiotic modes.* Four visual conventions (VCs) utilize analogies exclusively (VC1) or primarily (VC6–CV8). This is potentially powerful because analogies reveal structural similarities between two domains that reflect true characteristics (Gentner & Smith, 2012). Three VCs utilize metaphors (VC2–VC4), which can be coherent and improve understanding, while being dependent upon familiarity with a suggested source domain (Johnson, 1987; Lakoff & Johnson, 1980). But VC5

Figure 2. Eight potential visual conventions for representing uncertainty severity in LLM summaries.

Table 1. Semiotic modes and conceptual implications for the eight uncertainty-signifying visual conventions shown in Figure 2.

Visual convention	Evaluation: mode and implication
VC1: Strikethrough	**Mode:** response analogy, uncertain passages should be editorially rejected (thin line) or redacted (thick line). **Implication:** uncertainty is a mistake.
VC2: Transparency	**Mode:** disappearance metaphor, uncertain passages are fading out of existence. **Implication:** certainty is tangibility.
VC3: Static	**Mode:** television-static metaphor, uncertainty makes passages difficult to resolve. **Implication:** uncertainty is interference in a passage-signal.
VC4: Fill	**Mode:** fluid-volume metaphor based on an up-is-more orientational metaphor. **Implication:** uncertainty is a quantity in passages.
VC5: Pattern	**Mode:** primarily arbitrary symbolic representation (pattern style); secondarily density analogy (pattern repetition interval). **Implication:** uncertainty is a pattern — this is meaning-poor — or uncertainty is a quantity in passages.
VC6: Text Blur	**Mode:** visual perceptual analogy, certain passages are focused on (attended to) instead of uncertain passages. **Implication:** certainty is in focus or is comfortably accessible.
VC7: Zig-Zag	**Mode:** primarily editorial markup analogy; secondarily wave frequency metaphor. **Implication:** uncertainty passages are unfinished, or uncertainty is a force.
VC8: Weight	**Mode:** primarily mass analogy; secondarily arbitrary relative representation (color). **Implication:** uncertainty is conspicuous — this is meaning-poor.

uses arbitrary distinctions of patterns, which likely limits its efficacy, requiring viewers to learn symbols.

- *Conceptual implications.* Two VCs conceptualize *certainty as a thing* and *uncertainty as a reduction of that thing* (VC2, VC6). We suspect that this is a healthy way to view uncertainty. The remainder conceptualize *uncertainty as a thing,* which may be useful because uncertainty is what language analysts must interrogate. Two VCs are listed as being "meaning-poor" in Table 1 because their semiotic modes as executed do not suggest an obvious mental model, in our estimation (VC5, VC8).

- *Experiential criterion.* The two VCs that use graphic mark variability to differentiate degrees of severity — background patterns (VC5) and wavy underlines (VC7) — do not appear to give an immediate impression of uncertainty. The VC that depends on font weight variation (VC8) is possibly too subtle for viewers without graphic design expertise. The remaining VCs all appear to have some claim to the experiential.
- *Reflective criterion.* Our difficulty in describing the conceptual implications of the VCs we labeled as "meaning-poor" (VC5, VC8) would likely equate with minimal contributions to viewers' understanding of uncertainty. The remaining VCs appear to have some claim to the reflective.
- *Legible criterion.* Legibility was largely assured through hover states that offer a relatively unobstructed view of otherwise obscured text. Three VCs render passages of "obvious" severity as entirely or nearly unreadable before hovering (VC1, VC3, VC6), while two VCs effectively leave all text unobscured at all times (VC7, VC8). The VCs tend to interact with qualities that may typically be adjustable through accessibility settings, which is a complicating factor.
- *Implementable criterion.* Three VCs utilize normal web display settings without the need for background images or pseudo-class workarounds that are not compatible across browsers (VC2, VC6, VC8), while the remainder require background images or workarounds to implement. VC1 is a special case. Though strikethrough is readily available in HTML and in more rudimentary applications, it is not a customizable property in CSS. VC1 would be the most implementable if only one level of severity was to be signaled, but indicating multiple levels of severity with strikethrough currently requires a workaround.

The eight visual conventions have been implemented in a web-based simulation with readable code. This open resource enables others to reevaluate and even modify them, thus discovering additional strengths and weaknesses in the visual conventions.

5. Uncertainty Framework

5.1. Uncertainty Framework Development

One of the first tasks we undertook in our investigation was to define uncertainty in relation to LLM summaries for intelligence analysis. In our search of existing literature in early 2024, we found no uncertainty frameworks specific to LLM summaries for intelligence analysis. However, we did locate relevant research on trust in AI and automated decision aids (Zerilli et al., 2022; Fell et al., 2020; Heger et al., 2016; Manzey et al., 2012; Okamura & Yamada, 2020; Prabhudesai et al., 2023; Vaswani et al., 2017), as well as on uncertainty information and explanation visualization in other contexts (Karran et al., 2005; Skeels et al., 2010; Thomson et al., 2005).

Among these, Skeels et al. (2010) provided the most useful initial framework, having structured classification of uncertainty with what we believed to be sufficient range and granularity for visual exploration guidance. The framework identifies five types of uncertainty: completeness, credibility, disagreement, inference, and measurement precision. Skeels et al. (2010) isolated these types through an analysis of uncertainty across many scientific fields, including ecology, computational biology, and medicine, with a focus on improving information visualization — as in diagrams, not in the annotation of text as is required in language analysis.

Our revision of, and ultimate departure from, Skeels et al.'s (2010) framework was informed through biweekly conversations with collaborating experts in language analysis, computer science, and psychology. We asked them to speculate on what kinds of uncertainty might be associated with LLMs, and we tried to map their suggestions to Skeels et al.'s (2010) types. We began to remove types, add types, and rename types. We repeatedly needed to pull back to determine *where* exactly we were attempting to identify uncertainty — out in the world, in the sources, or in the summary itself? We ultimately decided to limit our investigation to the uncertainty types that might appear in the summary itself due to the probabilistic nature of LLM technology.

5.2. Overview of Uncertainty Types

Our LLM-oriented uncertainty framework is a contribution to research in the context of intelligence analysis, knowledge work, and interface design. Though it is not the result of a systematic study like Skeels et al.'s (2010) general uncertainty framework, a provisional framework provides the requisite *a priori* structure for subsequent empirical studies that could validate it or suggest adjustments. It is a necessary first step.

We call our framework the Uncertainty Framework for Explainable Summaries (UFES). We do not specify the intelligence context or language analysis in the title as we have adopted definitions that we believe more fundamentally address LLM summaries. There are five types of uncertainty in UFES, and they are defined as follows.

1. *Meaning uncertainty:* misinterpreting word sense for technical, cultural, or uncommon terms, or for jargon.
2. *Reference uncertainty:* mistaking associations from demonstratives ("those"), adverbs ("there"), definite articles ("the"), or pronouns ("they").
3. *Conjecture uncertainty:* jumping to conclusions, incorrectly completing partial information, or making assumptions.
4. *Credibility uncertainty:* trusting a statement that was unserious, humorous, incongruous, a *non sequitur,* a manipulation, or an apparent lie.
5. *Evidence uncertainty:* making a claim without supporting evidence, either drawing from opaque training data or by hallucination.

Table 2. An imagined surreptitious recording between fictional characters André Silva and Baaba Owusu, and concerning fictional countries Avalon and Oceania. Line 4 includes a break in recording or transcription.

Line	Transcription
1	**Silva:** He didn't like having to give that speech as [indecipherable] those rural teachers. They aren't going to see the big picture
2	**Owusu:** There is no big picture. It's just what you say to them, and what you say to the big public schools.
3	**Silva:** Look, it was all about timing. And that's gonna be on the cycle for a couple of days, he gets to duck out. Look at what we have coming up. It's manufacturing on the border, customs, coordination, the union bosses on both sides. The Avalon secretary
4	[indecipherable audio, 5′23″]
5	[unidentified]: and he plans to make an ultimatum, quietly.
6	**Owusu:** But that secret won't keep.
7	**Silva:** Sure but
8	**Owusu:** But Oceania will know what it means. They're ready to act on it. It's only days

As suggested earlier, UFES primarily refers to uncertainty *in summaries themselves,* not uncertainty in source information, which is irreducible when that information is isolated. For instance, if a query addresses the possibility of life on Mars, and a summary states that there is no *evidence* of life on Mars, it is not a case of evidence uncertainty — the stated lack of evidence is an accurate accounting of a knowledge reality. Essentially, UFES is concerned with misleading claims in summaries, and it serves to identify the manner in which a claim may be misleading.

Our thinking here was crystalized in a summary example that utilizes three fictional countries — *Kobian* administration officials discuss *Avalon* and *Oceania.* This example is provided in Table 2. The following summary claim is an accurate representation of the source material displayed in that table, in which "[indecipherable audio, 5′23″]" separates a pronoun from its thus unknown referent. An imagined summary provided by an LLM that consults the Table 2 source includes:

> There is mention of an "ultimatum," but an apparent gap in the source recording makes it impossible to determine who is making an ultimatum that will serve as a communication to Oceania, or if this is a serious ultimatum and the degree to which it is mission relevant.

Though there is obviously irreducible uncertainty in the imagined source material — whatever was said in the missing five minutes — the summary's claim itself is not uncertain.

UFES could inform development of an LLM validation system by providing a starter language from which to articulate soft prompts. But the more immediate use of UFES is to improve human understanding of uncertainty in the LLM outputs. We now provide more detail on the five types of uncertainty we have defined here.

5.3. Additional Detail on Uncertainty Types in UFES

Meaning and reference uncertainty. Meaning and reference uncertainty emerged late in our development process, and only when we consciously confronted the mismatch between Skeels et al.'s (2010) numeric focus and our language focus. As such, these types have no corollary in their framework. We intend meaning uncertainty to be more localized, at the level of individual terms, and reference uncertainty to be distributed, as arising from relationships between statements. We considered "denotative" and "indexical" as alternative names for these types of uncertainty, respectively, because both deal with meaning derivation. But we opted for a more colloquial or less technical terminology.

Reference uncertainty begs further explanation than its definition alone. Returning to the example of Table 2, consider a different LLM summary claim to the one used above:

> …The Secretary of Labor appears to be preparing to make an ultimatum regarding the manufacturing conflict…

Cross-referencing with Table 2, the gap in transcription noted in line 4 occurs between mention of "the Avalon secretary" and the "he" who is making an ultimatum. It is a questionable assumption that "he" is the secretary when there is a gap of over five minutes. Even attribution of "labor" to the secretary is questionable. It appears to be an assumption following "union bosses" in the sentence immediately preceding "the Avalon secretary." These represent two degrees of reference uncertainty embedded in the claim that *the Secretary of Labor* is *making an ultimatum.*

Conjecture uncertainty. Conjecture uncertainty bears some similarity to Skeels et al.'s (2010) inference uncertainty. In early stages of developing UFES, we had a broader definition of inference uncertainty. The difference is apparent in our notes:

> An assertion is in some manner ambiguous, with more than one possible meaning available to complete it. An error occurs when a claim about the assertion relies on a misinterpretation (in cases of logically resolvable ambiguity), or otherwise fails to acknowledge the ambiguity inherent to the assertion (in cases of irreducible ambiguity). Misinterpretations and ambiguity can be rooted in: idiomatic

> expressions; cultural nuances; context-specific phrases; polysemy; unconventional sentence structure; conjugation; pronouns; rare or uncommon terms; technical jargon; tone; humor; or sarcasm.

This transitional definition is so general that it embodies our revised sense of three types of uncertainty: meaning ("rare or uncommon terms"), reference ("pronouns"), and credibility ("sarcasm"). Basically, the transitional definition is too broad to be useful, and it is emblematic of the verbal gymnastics that were necessary to conform linguistic considerations to Skeels et al.'s (2010) numeric considerations. Ultimately, in addition to refining our own definition of a related kind of uncertainty, we opted to designate it *conjecture* to avoid a mismatched comparison with Skeels et al.'s *inference.*

Nevertheless, our conjecture uncertainty is admittedly difficult to isolate. To retain the focus on LLM summary generation, we emphasize that it refers to completing partial information without adequately acknowledging the completion act. Conjecture uncertainty occurs when an alternative summary claim could have been drawn from the same partial information. It is an assumption. This does embody aspects of Skeels et al.'s (2010) completeness uncertainty, but missing numeric values in a data set are far more conspicuous, and far less ambiguous, than partial linguistic information. The following two examples may help to clarify conjecture uncertainty further.

First, imagine a scenario where there exists copious intelligence traffic about a series of meetings between an adversarial country's president and a group of legislators on a particular issue, yet none of that traffic offers specific details about the meetings themselves. Instead, what is available are conversations among those legislators that occurred after the meetings, in which they complain about interpersonal dynamics and personal agendas (e.g., getting the chief of staff to admit that he is wrong about *anything*). If a summary characterized these meetings as the president strategizing against the legislators based solely on such traffic, it would have conjecture uncertainty. It is an assumption that the expressed feelings following the meetings transfer fully to the meeting's agenda.

Second, imagine another scenario where available traffic inconsistently presents a fictional president's views on anti-ballistic missile deployment. In two sources, he appears strongly and moderately for increased deployment, and in two other sources he appears strongly and moderately against it. Taken at face value, it could appear that the president has no strategy, or that he is obfuscating, and thus a summary may characterize his views on the matter as "suggesting an absent or clandestine strategy." However, if a closer look at the sources reveals that his statements (or insider statements about his views) were made in confidence and with conviction, the "absent" or "clandestine" claim has conjecture uncertainty — it is a step too far in assumption. (We have played this scenario out further, where ordering the sources by date reveals that the president's

views evolved over time — as in, the sources were all accurately capturing moments in a sequence.)

Credibility uncertainty. For credibility uncertainty alone we retain Skeels et al.'s (2010) name for a type. We did so because, though the original type does indeed focus on numerics, its concept of credibility transfers to linguistics in ways we do not find distorting. While we do expand the type's definition, we do not feel the need to replace any major aspect of it.

The most straightforward interpretation of credibility uncertainty in LLM outputs is trusting the words of somebody inherently untrustworthy, and our definition does account for this. Skeels et al. (2010) note that a "human source may be considered untrustworthy based on past behavior or associations" (p. 76). Credibility uncertainty can also be more contextual. Some people may be inherently more credible than others based on their expertise and believing certain statements from inherently nonexpert sources will carry a degree of credibility uncertainty. Skeels et al. (2010) account for this as well: "…information from a specialist may lead to less uncertainty than information from a generalist…" (p. 76).

We additionally consider situational factors. The situation in which a person makes a statement is an everpresent complication, whether that person is generally trustworthy or untrustworthy. If an expert or an insider is making a joke, they are not necessarily leveraging their beliefs, and thus their access to relevant knowledge does not validate the joke's implications. And much speech is rhetorical, aimed at convincing others through argumentation more than through the explication of truth, even in casual conversation. In practice, there are not purely trustworthy or untrustworthy people. Instead there are situations in which individual statements may or may not be credible. Thus, credibility uncertainty refers to assertions themselves and not the people who make them. An assertion may be less credible due to its speaker's identity, but it is the assertion itself that has credibility value in our framework — the statement is the thing.

Evidence uncertainty. When objective truth and complete understanding are not realistic goals, what does it mean to have or to lack "evidence"? The other four types of uncertainty in UFES all concern the interpretation of available information. Clearly a summary can include claims that are more thoroughly unfounded than they are uncertain in meaning, reference, conjecture, or credibility.

In LLM summary generation powered by retrieval-augmented generation (RAG), an LLM has documentable access to its summary (of course) and the RAG sources, but not to the training data that constitutes its underlying foundational model. Evidence uncertainty occurs when there is a discrepancy between the *summary* and *documentable sources.* A claim is made in the summary for which no form of evidence is available.

It is impossible to know why an LLM made a claim if no evidence is given. If there is no basis for the claim (besides perhaps a biasing query), it is a case of hallucination. But this is indistinguishable from two other possibilities: that the claim came from the black-box training data; or that there was some kind of failure in documenting the normally documentable sources. What is most immediate in an intelligence analysis context is that a claim has no supporting evidence. This is what determines a course of action for a language analyst or for other users.

Pulling out from an intelligence analysis context, this view of evidence uncertainty still has efficacy. Whatever the mechanism is — hallucination, training data, source disclosure error, or something else — a claim in a summary that cannot in any way be supported cannot be validated.

6. Specifications for an LLM Validation System

This section outlines the design specifications for a Multiple Agent Validation System, contributing to the intelligence community and any other cases involving decision-making and sensemaking with LLM summaries. These specifications are the product of all threads of the investigation and of sustained interaction with our multidisciplinary collaborators. As our exploration of uncertainty visualization increasingly raised issues of interface design, we asked language analysts what actions they might take when they encountered uncertainty in an LLM summary. They identified seven specific actions: (1) viewing source files, (2) asking the LLM about its sources, (3) assessing relevance to the query, (4) asking the LLM about its summary, (5) submitting a new query, (6) modifying their existing query, and (7) searching for more information elsewhere. This and other feedback eventually coalesced into an LLM validation system concept.

6.1. The MAVS Concept

We propose a Multiple Agent Validation System (MAVS) to make knowledge workers more efficient while mitigating the limitations of LLM technology, and to facilitate healthy trust calibration by addressing common user struggles with automation. Our MAVS specifications include 10 discrete features, conceptually distributed among three virtual agents: a Query Agent, an Analytic Agent, and an Evaluative Agent. Whether or not these virtual agents are implemented in separate LLMs or as roles within a single LLM, they are instantiated in the feature set as distinct entities to aid the user in developing an accurate mental model — both of MAVS's underlying processes and of LLM technology more generally. Figure 3 is a process diagram of MAVS. Note that validation — the V in MAVS — is completed by the user.

We organize the feature set here according to the virtual agents. Table 3 lists how MAVS features address the trust calibration interface design strategies that emerged in the

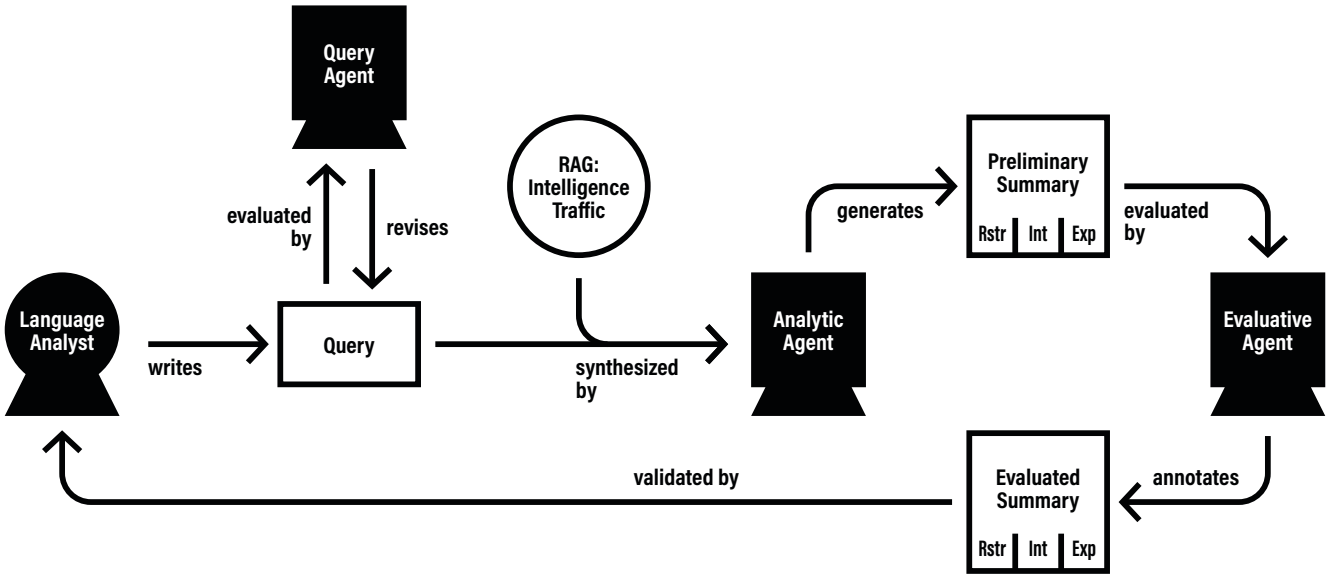

Figure 3. MAVS process diagram for the intelligence analysis context. Retrieval-augmented generation (RAG) focuses the system on intelligence traffic (i.e., collected sources). Summaries are available to users in three versions per the analytic sensitivity setting: restrictive ("rstr"), intermediate ("int"), and expansive ("exp").

Table 3. How the 10 core MAVS features address the trust calibration interface strategies (Section 2). The virtual agents are identified as QA (Query Agent), AA (Analytic Agent), and EA (Evaluative Agent).

Feature	Agent	Trust calibration interface strategies
Query History	QA	TC1: Transparency
Query Reshuffle	QA	TC4: Interactivity
Analytic Sensitivity	AA	TC4: Interactivity; TC5: Virtual Agents
Summary Sources	AA	TC1: Transparency
Visualization Sensitivity	EA	TC2: Visualization; TC3: Alignment; TC4: Interactivity; TC5: Virtual Agents
Uncertainty Visualization	EA	TC2: Visualization; TC3: Alignment
Uncertainty Alert Type Identification	EA	TC1: Transparency; TC2: Visualization; TC3: Alignment
Flagged Sources	EA	TC1: Transparency
Evaluative Agent Chat	EA	TC1: Transparency; TC4: Interactivity; TC5: Virtual Agents
Evaluation Export	EA	TC1: Transparency

literature review (Section 2, TC1–TC5). MAVS utilizes these key strategies to help users develop appropriate trust calibration with the goal of improving overall performance in human-machine collaboration. Figure 4 delineates areas within the simulation interface that are dedicated to the three virtual agents — the simulation interface is described in some detail in Section 7.1.

6.2. Query Agent Features

The Query Agent assists the user in the querying process. It has the lowest instantiation profile of the MAVS virtual agents.

Feature 1: Query history. The formality of the querying process, in which the user's investigation is comprehensively represented as text inputs and text outputs, affords a remarkably complete and accurate record of that investigation. The query history feature documents all user queries in a listing to which the user can return. The Query Agent dynamically generates short titles for queries, as commercial LLM products currently do. This listing indicates adjustments to queries resulting in distinct summaries with index counts of two and greater.

Feature 2: Query reshuffle. Prompting — writing queries — is a special skill and it can be done poorly, reducing or reversing the effectiveness of LLM summaries. With the query reshuffle feature, the user can request that the Query Agent analyze and improve their query, producing a new summary. This is an established capability for AI given the right training. A frequent outcome of query revision is debiasing, removing elements of queries that can push results in an inappropriate direction. For instance, the following query is a directive, not a question, which could effectively coax an LLM into confirming the query premise irrespective of the evidence: "Explain Nicolau's plan for anti-ballistic missile development and expansion." (Rysz Nicolau is president of the fictional country Kobia.) The directive assumes that Nicolau indeed has plans for such development and expansion. A debiased version of this prompt might be: "Does Nicolau have any plans for anti-ballistic missiles?" Query reshuffle modifies the query, which causes the Analytic Agent to generate a new summary, adding to the index count of the pre-shuffled query's listing in the query history.

6.3. Analytic Agent Features

The Analytic Agent responds to the user's query by drawing from a vast quantity of information and returning a summary. This is probably the most familiar role for an LLM.

Feature 3: Analytic sensitivity. The analytic sensitivity feature permits the user to influence how the Analytic Agent generates summaries with settings of *expansive, intermediate,* and *restrictive.* The expansive setting increases discovery by presenting

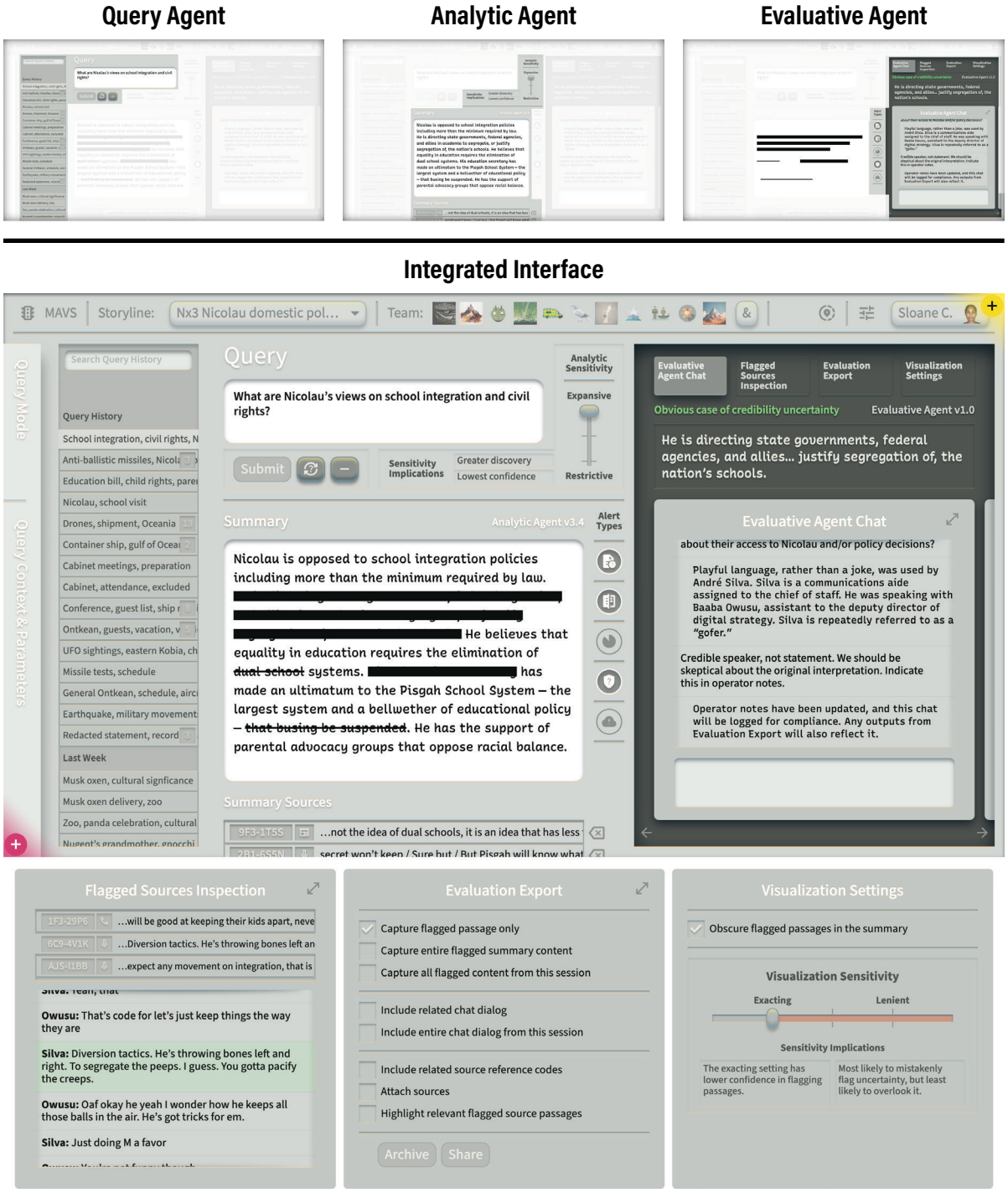

Figure 4. The three virtual agents in MAVS (top) as segmented in the full simulation interface (see Section 7.1). The three panels at bottom cycle through the area occupied by the Evaluative Agent chat at center right.

more possibilities to the user, but with a corresponding decrease in confidence. This may be appropriate in exploratory fact-finding. The restrictive setting results in greater confidence, but with lesser discovery that may overlook low-likelihood but potentially high-impact possibilities. This may be appropriate in crisis situations. As a setting that can be toggled, analytic sensitivity allows the user to see three versions of the Analytic Agent's summary for each query.

The analytic sensitivity feature could evolve iteratively through training and soft prompts. One possible method of implementation is for the Analytic Agent to independently generate *n* summaries, and then to compare those summaries. Claims that are shared across the highest proportion of independent summaries could be emphasized in a single common-claim summary — what is displayed with the restrictive setting. Reasonably strong claims that appear less frequently could be prioritized for a single uncommon-claim summary — some of what is displayed with the expansive setting. This method does beg the question as to how to keep the expansive summary to a reasonable length while still embodying some of the common claims, which should certainly not be ignored. Restrictive summaries may prove less useful due to the care taken to maximize reliability.

Feature 4: Summary sources. The ability to evaluate an LLM's claims is contingent upon access to the source material from which it reproduces patterns of language. Current LLM technology permits an accounting of this source material through retrieval-augmented generation (RAG). RAG enables an LLM to access a data set and to explicitly cite the sources of its claims in that data set — in contrast to the black-box behavior normally associated with LLMs. The summary sources feature discloses the specific sources the Analytic Agent used to generate a given summary. A relevant excerpt of the source is paired with metadata — in our case, a file code and an indication of recording medium — and the user can open the source directly to inspect it. The user can also remove a source from consideration, in which case the summary updates and the query history listing index count increases. Crucially in MAVS, the Evaluative Agent also has access to the disclosed summary sources, which is the basis of its validation process.

6.4. Evaluative Agent Features

The Evaluative Agent, the defining factor in MAVS, embodies an atypical role for an LLM agent. It utilizes specialized training and RAG to enact UFES (the uncertainty framework). The Evaluative Agent helps the user understand the Analytic Agent and validate its outputs. Ultimately, it evaluates congruence between the Analytic Agent's summary and its disclosed sources.

Feature 5: Visualization sensitivity. The visualization sensitivity feature permits the user to control when the Evaluative Agent bothers to provide markup, based on the

severity of uncertainty. Settings — *exacting, intermediate,* or *lenient* — characterize how the analyst asks the Evaluative Agent to behave. In attempting to recognize all instances of uncertainty in the summary, the exacting setting is the most likely to lead to mistakenly flagged passages. However, it is the least likely to overlook uncertainty — i.e., the most prone to false positives. The lenient setting results in more reliable uncertainty alerts. It is least likely to mistakenly flag instances of uncertainty, but most likely to overlook uncertainty — i.e., the most prone to false negatives. Unlike analytic sensitivity, toggling visualization sensitivity does not change the content of the summary. Instead, it flags more or fewer passages. The settings are operationalized according to plain language visible to the user: the lenient setting only flags *obvious* cases of uncertainty; the intermediate setting additionally flags *likely* cases; and the exacting setting additionally flags *conceivable* cases. There is no user option to bypass flagging obvious cases of uncertainty, as this would offer no initial means to validate the work of the Analytic Agent, hindering trust calibration.

Feature 6: Uncertainty visualization. There is an emotional component to uncertainty when there are professional stakes involved, and especially when there are security implications. In the flow of knowledge work, it is desirable that the user's emotional or gut sense of information is positively correlated with its certainty. The uncertainty visualization feature obscures passages in the summary commensurate with their assessed uncertainty severity. The levels of uncertainty controlled by visualization sensitivity — conceivable, likely, and obvious — are represented by increasing degrees of masking. A visual convention for achieving this may fully obscure obvious cases of uncertainty. However, cursor hover states for the summary itself permit the user to read flagged passages by responsively improving legibility.

Feature 7: Uncertainty alert type identification. To make accurate assessments when validating statements that have some degree of irreducible uncertainty, knowledge workers need to understand the basis of the uncertainty. We initially considered visualizing types of uncertainty instead of only severity level, but decided that this gives the user too much to learn and is distracting. Instead, the uncertainty alert type identification feature verbally identifies the type of uncertainty adjacent to the summary and only upon inspection. The Evaluative Agent identifies which of the five types of uncertainty is the reason a given passage was flagged, and concise definitions are provided for the types within the interface.

Feature 8: Flagged sources. An Evaluative Agent will be no more perfect than an Analytic Agent. Therefore a knowledge worker must be able to leverage their expertise to assess the Evaluative Agent's outcomes when validating the Analytic Agent's outcomes. The Evaluative Agent reports which of the Analytic Agent's disclosed sources led to an uncertainty alert. The flagged sources feature mimics the summary sources feature,

allowing the user to directly inspect sources in relation to an alert. An excerpt and metadata are immediately available, and the user can jump to a highlighted portion of the source information to begin the validation process, or they can open up the entire source.

Feature 9: Evaluative Agent chat. Providing users with a natural language mode of inquiry reduces the need to learn technically peculiar or unnatural interactions. A chat feature allows the user to engage in a conversation with the Evaluative Agent. When the user selects an individual flagged passage, the Evaluative Agent proactively explains why it was flagged. The user can engage the Evaluative Agent in conversation about the types of uncertainty, even if the agent does not tend to offer this information upon its initial description. This enables users to exert some control over LLM outputs, thus facilitating trust calibration.

Feature 10: Evaluation export. The more efficient augmented knowledge work becomes, the more difficult it will be for users to keep track of their investigations. The evaluation export feature documents the querying process for the user, along with all uncertainty alerts and interactions with the Evaluative Agent. In an intelligence analysis context, this is doubly important for compliance (i.e., aligning with strict regulations for reporting and documentation).

A robust implementation of MAVS would include these 10 features. Many of them can be experienced in the simulation interface (discussed in Section 7.1 and available in Peterson & Armstrong, 2024).

7. LLM Validation Prototypes

The LLM validation prototypes presented in this section contribute to knowledge work and the intelligence community in two important ways: they explicitly illustrate specifications that could be operationalized for practice (or for research in advance of practice); and they serve as educational tools that can help language analysts understand the potential of AI and set the stage for healthy trust calibration.

Scenarios situate design investigations within real-world contexts. In the scenario we developed for this investigation, a United States language analyst Sloane has been assigned to monitor and investigate the fictional country of Kobia and its president Rysz Nicolau. Our team employed this scenario within a simulation interface (in interactive demonstration form) and in three additional narrative interfaces (in video form). These prototypes are available in Peterson and Armstrong (2024). To most effectively demonstrate key MAVS functionality, we selected specific sections of the scenario to highlight within each prototype. The scenario content includes summaries, source excerpts, and chat scripts.

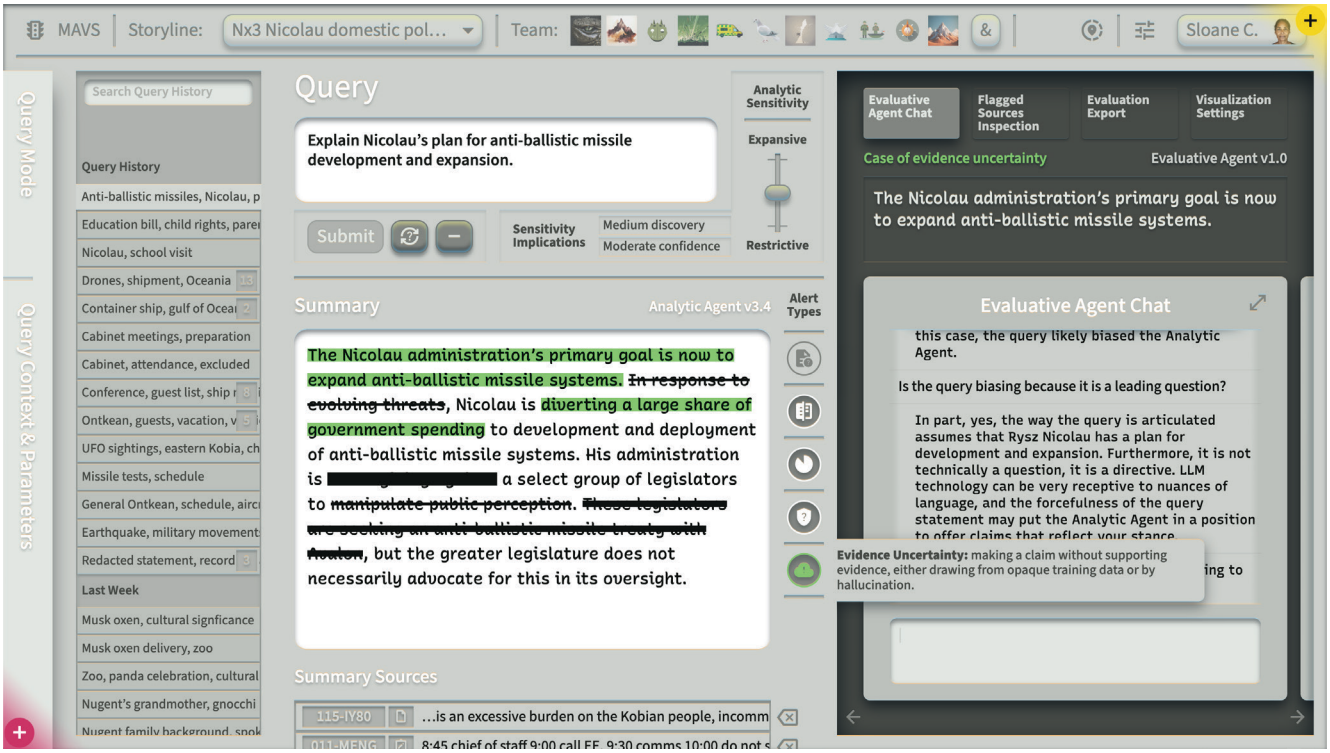

Figure 5. Simulation interface. In the pictured state, the user is hovering over the evidence uncertainty alert icon, which automatically highlights all evidence uncertainty passages in the summary.

7.1. Simulation Interface

The first Multiple Agent Validation System prototype is a simulation interface populated with scenario content. This prototype provides a realistic first-person experience of numerous MAVS features. Users play the part of Sloane as she engages in analysis of President Nicolau and his administration, with a narrative that permits significant lateral exploration through optional content. The web-based simulation interface was built in HTML, CSS, and JavaScript. While it does not incorporate actual AI — instead simulating AI — it is based on a tool developed by LAS (the collaborating lab) that utilizes an LLM and RAG with representative intelligence traffic.

We utilized a familiar control panel metaphor for the visual design of the interface to reduce cognitive load. Because simulation users are being asked to learn about an unfamiliar system (MAVS), they are not additionally asked to learn new interface conventions (Figure 5). The incomplete nature of the embodied scenario complicates the educational aspect of the prototype. Interface elements that are available at one time (i.e., that are scripted) are not available at others. To guide users through the scenario and to make sense of what is and is not interactable along the way, an instructional panel overlays one corner of the interface (Figure 6). The instructional panel suggests next steps with check boxes for completed tasks.

Another panel overlay permits users to select among eight uncertainty visualization styles to be utilized in the simulated LLM summaries (Figure 7). Like the instructional panel, the visualization panel would not be included in MAVS, for which a single visual convention for representing uncertainty would have been adopted. As such, the simulation interface allows practicing language analysts to experience different visual

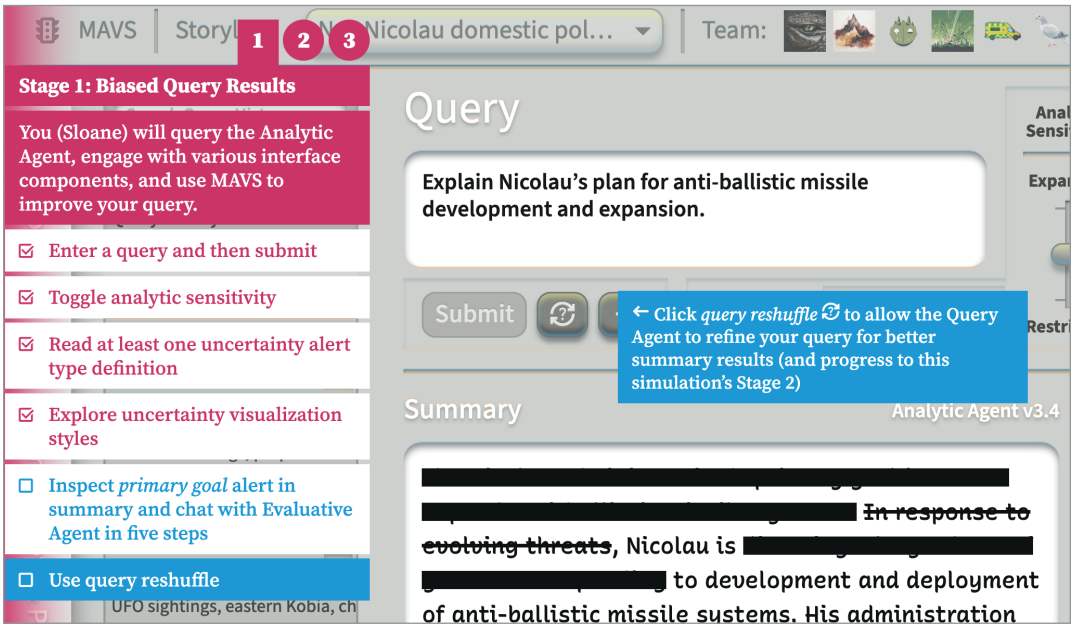

Figure 6. Simulation interface instructional panel with task hint. As listed tasks are completed, they are checked off. Hovering over incomplete tasks activates hints that point to interface elements.

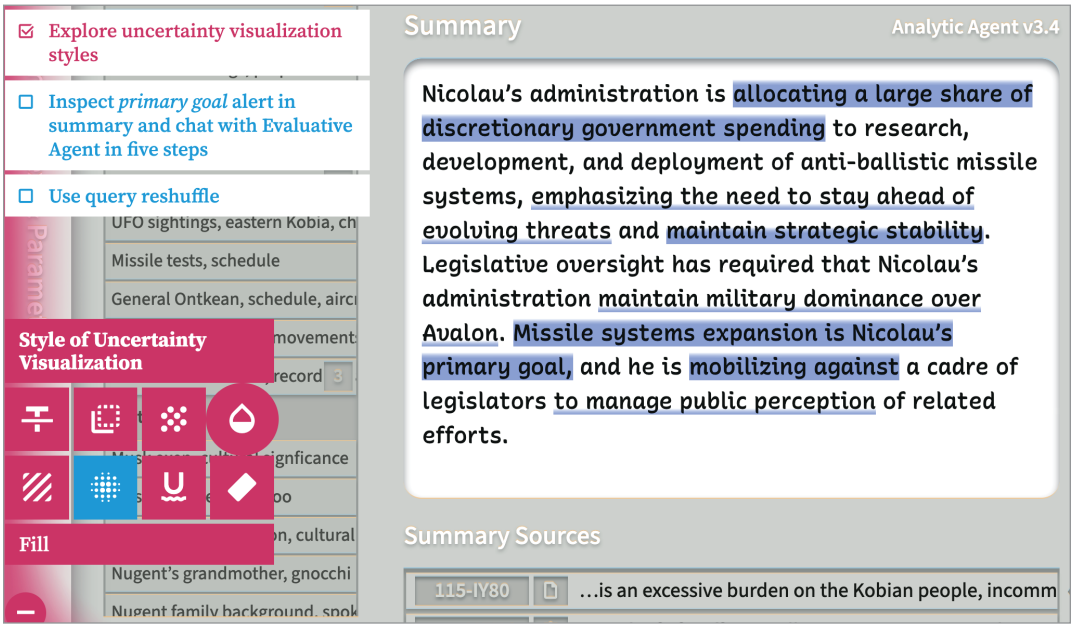

Figure 7. Simulation interface visualization panel with corresponding visual convention displayed in the summary.

conventions and consider their own preferences, leaving open the possibility of an end-user-informed determination of the ideal visual convention to adopt. In this way and others the simulation interface is a rich experimental stimulus that could be used to empirically test MAVS before costly development efforts are undertaken.

7.2. Narrative Interfaces

The simulation interface leverages familiarity by suggesting a control panel metaphor. However, this metaphor limits the potential capabilities of emergent technology as envisioned in MAVS. We thus also developed scenario videos for three distinct MAVS prototypes that are based on unconventional UX patterns, which may more naturally exemplify MAVS functionality. To focus the resultant narrative interfaces, we looked to the trust calibration (TC) literature and related interface design strategies TC1–TC5 (see Section 2). Based on these sources, we wrote three prompts to guide our interfaces.

1. *Transparency through interrogation and verification:* How might the interface utilize query recommendations, nudging, verification, and source inspection to calibrate trust between users and virtual agents? (Corresponds with TC1: Transparency and TC3: Alignment.)
2. *Multi-agent dialogue:* How might the interface use conversational AI to calibrate trust between users and virtual agents? (Corresponds with TC3: Alignment and TC5: Virtual Agents.)
3. *Context-driven:* How might the interface respond to the needs of specific users, customers, or storylines to calibrate trust between users and virtual agents? (Corresponds with TC3: Alignment and TC4: Interactivity.)

We utilized common UX methods to engage with our collaborators as we developed the narrative interfaces, including personas, scenarios, task flows, low- and high-fidelity sketching, and what-if prompts.

Narrative interface 1: Transparency through interrogation and verification. Language analysts want to leverage their own intricate understanding of human language to verify and interrogate data themselves. The first narrative interface provides an uncertainty alert report panel with a natural language explanation of identified uncertainty errors, along with key excerpts of flagged sources and quick access to the full sources themselves (Figure 8). This collects elements together in one space that the simulation interface distributes among separate zones. The uncertainty alert report panel reconfigures to accommodate the Evaluative Agent chat, and as with the simulation interface, the conversational interaction permits the analyst to submit operator notes — records for the chain of command and compliance — as they are suggested during sensemaking.

The first narrative interface deviates most dramatically from the simulation interface. It presents the user's workflow as branching diagrams of expanded and collapsed

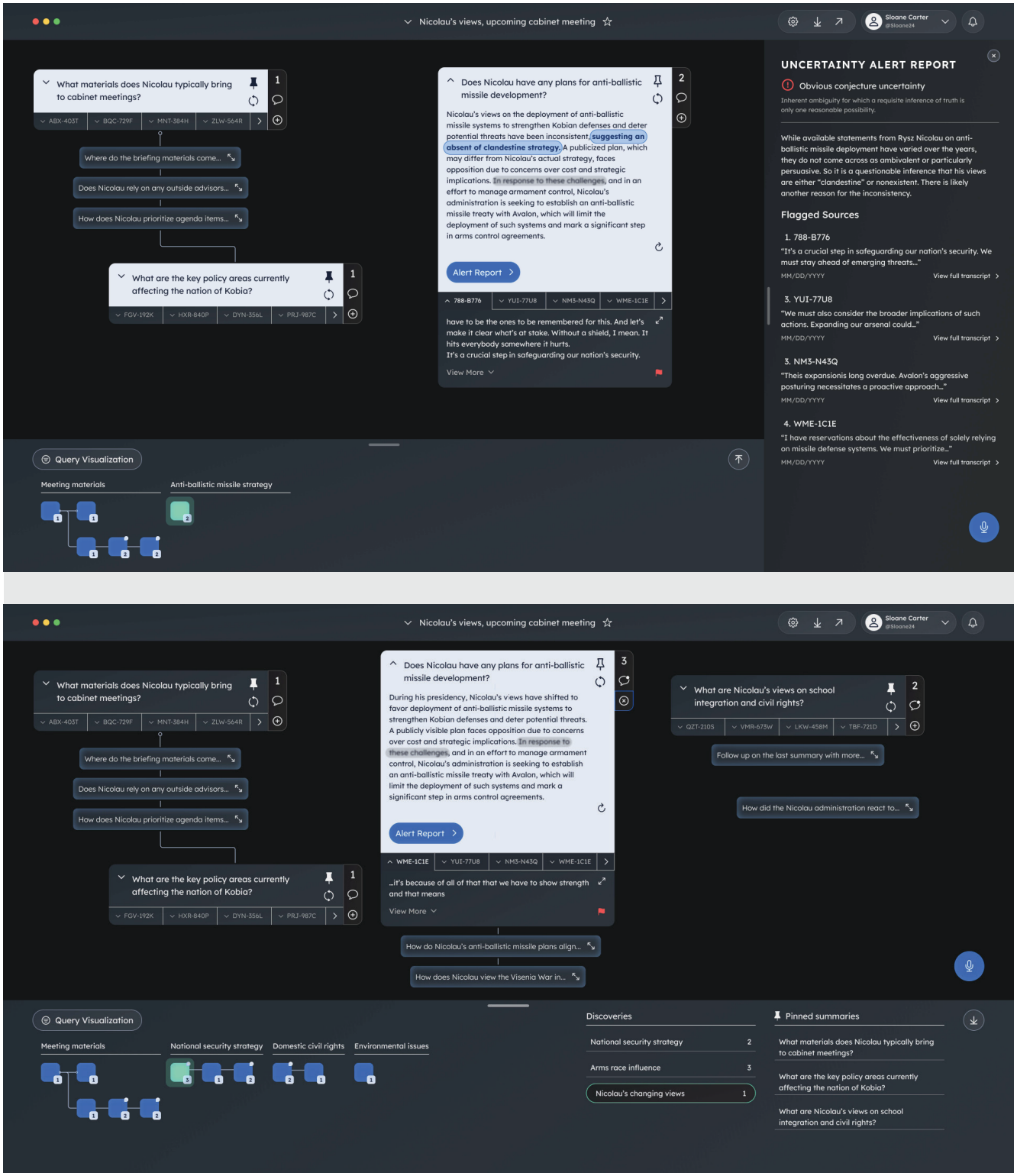

Figure 8. First narrative interface. Prompt-summary elements expand and collapse to varying degrees and their arrangement reflects the language analyst's investigatory process.

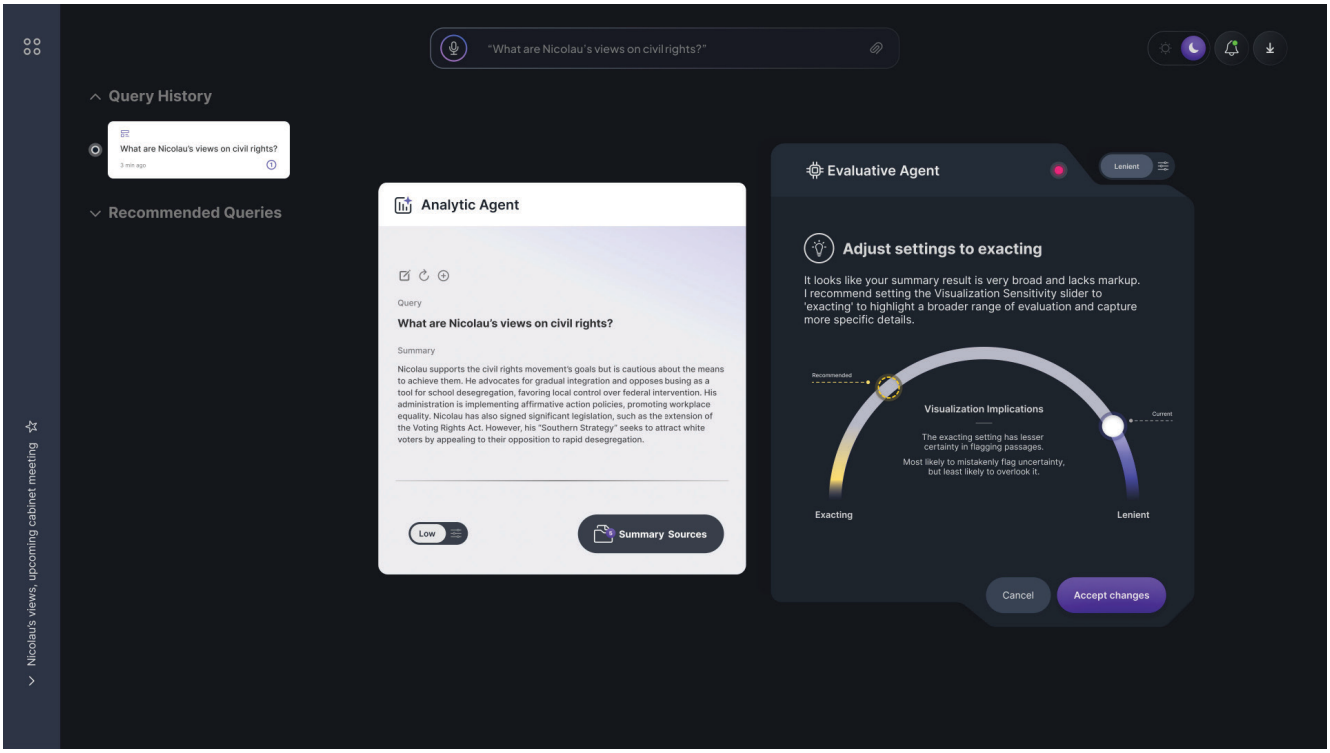

Figure 9. Second narrative interface with the Evaluative Agent active. The Analytic and Evaluative agents are presented as distinct entities that can respond to one another, and they make recommendations for the user.

prompt-summaries. In the central portion of the interface, prompts are chained together and break out when expanded to display summaries and other elements. In a lower query visualization strip, queries are minimized into icons and are organized into categories, while also reflecting investigatory pathways through chained connections. The interface constantly reconfigures as the user progresses. When appropriate, it nudges the user to revisit key points in the querying process and occasionally appends key insights to query icons. The interface enables the user to study how the Analytic and Evaluative Agents arrive at their conclusions, which helps the user better understand the system's capabilities.

Narrative interface 2: Multi-agent dialogue. The second narrative interface overtly presents the Analytic and Evaluative Agents as separate entities, embodied in adjacent floating panels. The user can converse with either virtual agent using the agent's panel, which grows slightly larger and includes a softly pulsing red light to reflect active engagement. Separating the virtual agents helps the user to conceptualize MAVS functionality by distinguishing analytic processes from evaluative processes. This facilitation is furthered with floating queries that can be dragged into a virtual agent's panel for a response, embodying the virtual agent with a recipient role. The pattern of virtual agents responding to each other, back-and-forth and through reciprocal panel

shrinking and growing behaviors, strengthens this embodiment. The greater the degree to which the interface distinguishes virtual agents, the easier it will be for the user to differentiate their system functions.

Narrative interface 3: Context-driven. The third narrative interface is an unconventional user interface that reconfigures itself as the user's investigation evolves (Figure 10). The user's investigatory process is structured as one long conversational flow. Both Analytic and Evaluative Agents converse with the user within this flow. The Analytic Agent auto-fills the analytic sensitivity setting according to the current storyline's dynamics. The user can adjust the setting, but recommendations are clearly indicated, possibly encouraging the user to experience settings they would not otherwise utilize. Recommended settings consider a variety of factors, such as the criticality of the storyline and past user behavior in similar situations. The Evaluative Agent also recognizes the criticality and greater context of the current storyline. For instance, it does not merely flag passages for contextually relevant uncertainty, it does so through a liquid panel that flows into the Analytic Agent's summary results. This violation of conventional panel integrity in interface design is a visual analogy for distinct virtual agent interactions and an incisive form of evaluation. Finally, the entire screen display also adjusts contextually, with the color scheme changing and element count reducing in critical high-stakes storyline periods to focus the user's attention.

These narrative interfaces provided an avenue for investigating trust calibration and interface design beyond the initial simulation interface. They embrace the potential of machine learning capabilities to build trust through transparency and conceptualization. The video format for presenting these interfaces guides viewers through a coherent workflow, making unconventional UX patterns sensible upon initial viewing. The novel element display and UX patterns within the narrative interfaces may potentially inform AI-based interface design beyond our investigation's focus on intelligence analysis.

8. Application and Transfer of Results

This section outlines aspects of our investigation that may contribute to communities of practice beyond intelligence analysis. We identify two application areas where investigation outcomes may be particularly relevant: LLM-assisted clinical decision-making in medicine and LLM-assisted climate forecasting. We examine relevant research in each area, viewing challenges related to uncertainty through the lens we have established. This suggests future work, but it also serves as a demonstration of how investigation outcomes can be adapted for additional areas not covered here.

Clinical decision-making in medicine. Researchers are actively exploring how LLMs can enhance clinical decision-making and provide diagnostic support in medicine

Figure 10. Third narrative interface. Chat elements flow into one another, and the system reconfigures itself in high-stakes moments.

(Nasarian et al., 2024; Panagoulias et al., 2023; Prabhod, 2023; Rajashekar et al., 2024; Savage et al., 2025). Some work has focused more narrowly on emergency care, digital pathology, and telehealth (Taylor et al., 2024; Kwan, 2024; Ullah et al., 2024). A common application of LLMs in clinical decision-making is assisting clinicians in prioritizing differential diagnoses (Prabhod, 2023; Taylor et al., 2024). Differential diagnosis is the systematic process used by clinicians to identify the most likely diagnoses from a set of competing possibilities (Cook & Décary, 2019). While established systems support this process, there is growing interest in expanding AI's role to mitigate diagnostic errors, improve information gathering, and facilitate diagnostic feedback (Taylor et al., 2024).

Despite the recognized potential of machine learning in this domain, a frequently cited challenge to integration is the lack of explainability in LLM-augmented systems, which has been shown to undermine user trust and hinder technology adoption (Panagoulias et al., 2023; Rajashekar et al., 2024; Savage et al., 2025; Ullah et al., 2024). This is particularly critical in high-stakes medical environments, where the urgency of decision-making, the fragmented nature of data, and the potential for cognitive overload leave little tolerance for uncertainty. Some researchers have focused on developing evaluation systems to measure or minimize uncertainty, but the literature does not fully address how to communicate uncertainty effectively to users in a medical context (e.g., Panagoulias et al., 2023; Savage et al., 2024). Nasarian et al. (2024) argue that many of the current explainable AI efforts are developer-centric, and that they tend to overlook the actual needs of end users. They further suggest that while machine learning professionals and developers tend to favor technical explanations, clinicians and patients would benefit from more intuitive visual formats. Both Prabhod (2023) and Kwan (2024) argue that future research should take a user-centered design approach, and should explore ways to provide training and meaningful engagement for clinicians. Kwan (2024) emphasizes that understanding user needs, defining user personas, and building prototypes are essential steps in developing AI-driven systems that can support clinical decision-making. Similarly, Taylor et al. (2024) emphasize the importance of designing AI tools that integrate seamlessly into clinician workflows and platforms without adding unnecessary complexity — namely in electronic health records (EHRs).

Climate forecasting. Similar issues arise in climate forecasting, where LLM-augmented systems must balance accuracy, interpretability, and usability to support not only decision-making, but also data analysis, communication to lay audiences, and generation of climate scenarios that can further inform decisions (Biswas, 2023). A preliminary search anecdotally suggests that the research in this area may not be as established as in medical diagnosis — most relevant work is in the form of unvetted uploads to preprint servers that we do not cover here.

Biswas (2023) explored how ChaptGPT could be leveraged to support climate research and policymaking. In this context LLMs are useful for generating instructive climate scenarios. However, the tendency of LLMs to hallucinate is significant due to limited context and expertise on climate data. For this reason, Biswas (2023) suggests that AI be used alongside traditional climate research methods, rather than as a replacement. Vaghefi et al. (2023) developed a specialized LLM, ChatClimate, to respond to queries related specifically to climate science. Their goal in creating this tailored model was to address challenges like hallucination and the presence of outdated information that might arise when using general purpose models. ChatClimate was not developed to replace the kinds of decision-making currently done by climate experts, but to increase the speed at which quality information on climate science can be accessed.

In contrast to broader applications of LLMs in climate science, Lawson et al. (2025) focused on weather forecasting at a more immediate ground level. They examined how well ChatGPT could analyze meteorological imagery and communicate hazard summaries in English and Spanish. ChatGPT struggled with the same challenges encountered in other climate applications, including hallucination and a lack of explainability and trustworthiness. Lawson et al.'s (2025) findings suggest that work remains to be done. Since these models are being leveraged for both long-term climate projections and real-time weather hazards, better representations of uncertainty could help facilitate trust calibration and generally improve system performance.

Implications. Our investigation offers several contributions that may be relevant to both clinical decision-making in medicine and climate forecasting, where LLM-augmented systems currently struggle with explainability and trust. The Uncertainty Framework for Explainable Summaries (Section 5) could help both clinicians and climate researchers interpret model outputs more effectively. The framework's pairing of evidence and credibility uncertainty would add nuance to understanding of uncertainty in both application areas, and the Multiple Agent Validation System's Evaluative Agent chat feature (Section 6) would help climate scientists vet the outdated information they frequently encounter. MAVS generally facilitates the rapid information gathering important in both application areas. In clinical settings, it could help mitigate the risks of misdiagnosis by ensuring AI-assisted insights are more transparent. In climate forecasting, it could improve trust in AI-generated climate scenarios by providing clearer explanations of LLM limitations. MAVS was conceptualized through user-focused design exploration, and thus does not suffer from the "developer-centric" emphasis of AI efforts in clinical decision-making (Nasarian et al., 2024).

A key investigation outcome for both application areas is the open resource of visual conventions for representing uncertainty in LLM-generated summaries (Section 4 and Section 7.1). The particulars of an application space are likely to influence valuation of

visualization efficacy, and the emergent criteria for selection can scaffold fresh valuation (Section 4). These contributions address core issues of accuracy and interpretability in domains where decisions have real-world consequences.

9. Discussion

Automation transparency has a positive impact on user task performance (van de Merwe et al., 2022). Knowledge workers need insight into the LLM systems they increasingly rely upon. And uncertainty is unavoidable with LLMs. While communicating this uncertainty benefits users, user outcomes differ based on the indicated degree of uncertainty — e.g., Kunze et al. (2019) noted participant behavioral changes at three levels of uncertainty, which correspond with our signification of conceivable, likely, and obvious levels.

Kunze et al. (2019) highlighted a "drawback" to displaying uncertainty: users need to look away from the task at hand to attend to visualizations (p. 355). But they were studying automated driving systems. When we explored uncertainty visualization, we did consider adjacent visualizations of uncertainty separate from the uncertain summaries, but our eight potential visual conventions are all inline, occurring directly within uncertain textual passages themselves (Section 4). The relevant literature on visualization tends to address standalone representations — the uncertainty framework we started with, Skeels et al. (2010), is a good example. The simple fact that our recommended forms of signification occur at the locus of uncertainty for LLMs — *in* and *as* written language — is possibly a powerful visual affordance for user interface design and transparent AI. It is possible for uncertain text to be — really, *to appear* — uncertain itself. This seems more desirable than providing an additional thing for overtaxed users to look at.

A result of interdisciplinary collaboration between language analysts, computer scientists, psychologists, and designers, this investigation provides human-centered recommendations that can guide LLM technology development. The expertise of the extended collaborative team ensures that the Multiple Agent Validation System, as described, is both implementable and relevant. This places our speculative design squarely in the present.

As a discovery-based process, this investigation suggested new research questions instead of answering preconceived ones. There are a variety of ways to continue this work. The most direct way would be completing the theory building and testing cycle through the empirical study of core project premises. The simulation interface (Section 7.1) could be used to test the MAVS summary validation process with intelligence analysts. The objective would be to identify uncertainty visualizations and

interface features that optimize intelligence analysts' ability to accurately validate LLM summaries with analytic and evaluative assistance from AI (as simulated, not implemented). This suggests two research questions:

- *Research Question 1 (RQ1):* What preferences do intelligence analysts have in the design of uncertainty communication (including visualization), and how do those preferences translate into trust attitudes and dependence behaviors?
- *RQ2:* When intelligence analysts are presented with potentially erroneous information, what actions do they take to validate and integrate the information with existing schemas, and how does this influence the performance of the analyst-automation team?

Answering these research questions would result in design principles that could guide work in human-machine teaming, explicit design implementations for effective communication about the uncertainty of LLMs, and an understanding of analyst information validation behavior based on trust.

Individual components of this investigation suggest other possibilities. Collaboration with experts and design exploration suggested deviating from Skeels et al.'s (2010) framework for uncertainty, but unlike that framework, our proposed Uncertainty Framework for Explainable Summaries has not been validated. A qualitative study of UFES with members of the intelligence community and LLM developers could utilize the framework's five types as *a priori* codes to refine it. Likewise, there has been no validation of the eight implemented visual conventions for representing uncertainty. Qualitative or quantitative research with intelligence analysts could tease out how impressions of the visualizations correspond with, and contribute to, mental models of uncertainty.

- *RQ3:* What types of uncertainty are present in LLM summaries, and how do intelligence analysts and LLM developers conceptualize these types?
- *RQ4:* How does the visualization of uncertainty in LLM summaries impact conceptualizations of uncertainty?

The results of studies like these could lead to modifications to MAVS specifications, such as altering the descriptors for analytic and visualization sensitivity.

As described in Section 8, this investigation has relevance for clinical design-making in medicine and climate forecasting. There are potential investigations in these transfer domains. For instance, there is interest in improving information gathering for differential medical diagnosis (Taylor et al., 2024), and in generating climate scenarios that are context-sensitive (Biswas, 2023).

- *RQ5:* How do clinicians conceptualize and utilize restrictive and expansive settings for differential diagnosis when an analytic sensitivity feature is available in MAVS?
- *RQ6:* How do climate scientists prompt an Evaluative Agent to vet an Analytic Agent's generated climate scenarios in climate policymaking?

In terms of application, the speculative interfaces explored in this investigation provide various pathways for developing MAVS according to context and constraints. Operationalizing the behaviors of the Query, Analytic, and Evaluative Agents would require significant iteration in training LLMs, but fundamentally new training methods need not be developed.

This investigation resulted in visual conventions for representing uncertainty in LLM summaries, a related framework for uncertainty, specifications for an LLM validation system, and situated prototypes of such a system. We have been explicit about the investigation's contributions throughout to promote application in and beyond the intelligence analysis context. Trust calibration plays a key role in successful human-machine teaming. We will never move beyond human limitations if we cannot trust AI to support and augment our abilities. User interface design plays a critical role in trust calibration because interfaces lie between humans and automated systems. Through thoughtful interface design that prioritizes transparency and human understanding, we can build the foundations of trust necessary for humans and AI to work together effectively, revealing new possibilities.

10. Acknowledgments

The authors were the investigation's principal investigators (Peterson and Armstrong) and research assistants (Planchart, Baidoo, and Anderson, along with Kayla Rondinelli). Collaborators at the Laboratory for Analytic Sciences included Christine Brugh, Patti K., Sue Mi K., and Mike G., among others. RQ1 and RQ2 were written with Colleen Patton. This material is based upon work done, in whole or in part, in coordination with the Department of Defense (DoD). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the DoD and/or any agency or entity of the United States Government.

11. References

Alhadad, S. S. (2018). Visualizing data to support judgement, inference, and decision making in learning analytics: Insights from cognitive psychology and visualization science. *Journal of Learning Analytics, 5*(2), 60–85. https://doi.org/10.18608/jla.2018.52.5

Amershi, S., Inkpen, K., Teevan, J., Kikin-Gil, R., Horvitz, E., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., & Bennett, P. N. (2019). Guidelines

for human-AI interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). Association for Computing Machinery. https://doi.org/10.1145/3290605.3300233

Baldassi, S., Megna, N., & Burr, D. C. (2006). Visual clutter causes high-magnitude errors. *PLOS Biology, 4*(3), 387–394. https://doi.org/10.1371/journal.pbio.0040056

Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., & Horvitz, E. (2019). Beyond accuracy: The role of mental models in human-AI team performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 7*(1), 2–11. https://doi.org/10.1609/hcomp.v7i1.5285

Bellotti, V., & Edwards, K. (2001). Intelligibility and accountability: Human considerations in context-aware systems. *Human–Computer Interaction, 16*(2–4), 193–212. https://doi.org/10.1207/S15327051HCI16234_05

Biswas, S. S. (2023). Potential use of Chat GPT in global warming. *Annals of Biomedical Engineering, 51*(6), 1126–1127. https://doi.org/10.1007/s10439-023-03171-8

Borgo, R., Wall, E., Matzen, L., El-Assady, M., Masters, P., Hosseinpour, H., Endert, A., Chau, P., Perer, A., Schupp, H., Strobelt, H., & Padilla, L. (2024). Trust junk and evil knobs: Calibrating trust in AI visualization. *Proceedings of the IEEE 17th Pacific Visualization Conference (PacificVis)*, 22–31. https://doi.org/10.1109/pacificvis60374.2024.00012

Cook, C. E., & Décary, S. (2020). Higher order thinking about differential diagnosis. *Brazilian Journal of Physical Therapy, 24*(1), 1–7. https://doi.org/10.1016/j.bjpt.2019.01.010

Corritore, C. L., Kracher, B., & Wiedenbeck, S. (2003). Online trust: Concepts, evolving themes, and a model. *International Journal of Human-Computer Studies, 58*(6), 737–758. https://doi.org/10.1016/S1071-5819(03)00041-7

Cummings, M. L. (2006). Automation and accountability in decision support system interface design. *Journal of Technology Studies, 32*(1), 23–31. http://hdl.handle.net/1721.1/90321

Dasgupta, A., Lee, J.-Y., Wilson, R., Lafrance, R. A., Cramer, N., Cook, K., & Payne, S. (2017). Familiarity vs. trust: A comparative study of domain scientists' trust in visual analytics and conventional analysis methods. *IEEE Transactions on Visualization and Computer Graphics, 23*(1), 271–280. https://doi.org/10.1109/tvcg.2016.2598544

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science, 64*(3), 1155–1170. https://doi.org/10.1287/mnsc.2016.2643

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv*. https://arxiv.org/abs/1702.08608

Evans, J. St. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science, 8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Fell, L., Gibson, A., Bruza, P., & Hoyte, P. (2020). Human information interaction and the cognitive predicting theory of trust. *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, 145–152. https://doi.org/10.1145/3343413.3377981

Forsey, H., Leahy, D., Fields, B., Minocha, S., Attfield, S., & Snell, T. (2024). Designing for learnability: Improvement through layered interfaces. *Ergonomics in Design: The Quarterly of Human Factors Applications*. https://doi.org/10.1177/10648046241273291.

Gentner, D., & Smith, L. (2012). Analogical reasoning. In V. S. Ramachandran (Ed.), *Encyclopedia of human behavior (2nd ed.),* pp. 130–136. Elsevier.

Graaf, M. D., & Malle, B. F. (2017). How people explain action (and autonomous intelligent systems should too). *Proceedings of the AAAI Fall Symposium*. https://api.semanticscholar.org/CorpusID:53292049

Heger, O., Kampling, H., & Niehaves, B. (2016). Towards a theory of trust-based acceptance of affective technology. *European Conference on Information Systems*. https://www.semanticscholar.org/paper/Towards-a-Theory-of-Trust-based-Acceptance-of-Heger-Kampling/4640516f7b9a2a693a80ce667c98dba1fc7eb8bb

Johnson, M. (1987). *The body in the mind: The bodily basis of meaning, imagination, and reason.* University of Chicago Press.

Joshi, S., Nistala, P. V., Jani, H., Sakhardande, P., & Dsouza, T. (2017). User-centered design journey for pattern development. In *Proceedings of the 22nd European Conference on Pattern Languages of Programs* (EuroPLoP '17) (Article 23, pp. 1–19). Association for Computing Machinery. https://doi.org/10.1145/3147704.3147730

Karran, A. J., Demazure, T., Hudon, A., Senecal, S., & Léger, P.-M. (2022). Designing for confidence: The impact of visualizing artificial intelligence decisions. *Frontiers in Neuroscience, 16*. https://doi.org/10.3389/fnins.2022.883385

Kirschner, P. A., Ayres, P., & Chandler, P. (2011). Contemporary cognitive load theory research: The good, the bad, and the ugly. *Computers in Human Behavior, 27*(1), 99–105. https://doi.org/10.1016/j.chb.2010.06.025

Kocielnik, R., Amershi, S., & Bennett, P. N. (2019). Will you accept an imperfect AI? Exploring designs for adjusting end-user expectations of AI systems. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–14). Association for Computing Machinery. https://doi.org/10.1145/3290605.3300641

Krueger, R. F., Beyer, J., Jang, W.-D., Kim, N., Sokolov, A., Sorger, P. K., & Pfister, H. (2020). Facetto: Combining unsupervised and supervised learning for hierarchical phenotype analysis in multi-channel image data. *IEEE Transactions on Visualization and Computer Graphics, 26*(1), 227–237. https://doi.org/10.1109/TVCG.2019.2934547

Kunze, A., Summerskill, S. J., Marshall, R., & Filtness, A. J. (2019). Automation transparency: Implications of uncertainty communication for human-automation interaction and interfaces. *Ergonomics, 62*(3), 345–360. https://doi.org/10.1080/00140139.2018.1547842

Kwan, H. Y. (2024). User-focused telehealth powered by LLMs: Bridging the gap between technology and human-centric care delivery. *2024 4th International Conference on Computer Communication and Artificial Intelligence (CCAI)*, 187–191. https://doi.org/10.1109/CCAI61966.2024.10603150

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by.* University of Chicago Press.

LAS [Laboratory for Analytic Sciences] (2024). Sight unseen, students design UX for intelligence analysts [webpage]. https://ncsu-las.org/2024/09/sight-unseen-students-design-ux-for-intelligence-analysts/

Lawson, J. R., Trujillo-Falcón, J. E., Schultz, D. M., Flora, M. L., Goebbert, K. H., Lyman, S. N., Potvin, C. K., & Stepanek, A. J. (2025). Pixels and predictions: Potential of GPT-4V in meteorological imagery analysis and forecast communication. *Artificial Intelligence for the Earth Systems, 4*(1), 240029. https://doi.org/10.1175/AIES-D-24-0029.1

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Manzey, D., Reichenbach, J., & Onnasch, L. (2012). Human performance consequences of automated decision aids: The impact of degree of automation and system experience. *Journal of Cognitive Engineering and Decision Making, 6*(1), 57–87. https://doi.org/10.1177/1555343411433844

Nasarian, E., Alizadehsani, R., Acharya, U. R., & Tsui, K.-L. (2024). Designing interpretable ML system to enhance trust in healthcare: A systematic review to proposed responsible

clinician-AI-collaboration framework. *Information Fusion, 108*, 102412. https://doi.org/10.1016/j.inffus.2024.102412

Nass, C., & Brave, S. (2007). *Wired for speech: How voice activates and advances the human-computer relationship*. MIT Press.

Okamura, K., & Yamada, S. (2020). Empirical evaluations of framework for adaptive trust calibration in human-AI cooperation. *IEEE Access, 8*, 220335–220351. https://doi.org/10.1109/ACCESS.2020.3042556

Padilla, L. M., Creem-Regehr, S. H., Hegarty, M., & Stefanucci, J. K. (2018). Decision making with visualizations: A cognitive framework across disciplines. *Cognitive Research: Principles and Implications, 3*, 29. https://doi.org/10.1186/s41235-018-0120-9

Panagoulias, D. P., Palamidas, F. A., Virvou, M., & Tsihrintzis, G. A. (2023). Evaluating the potential of LLMs and ChatGPT on medical diagnosis and treatment. *2023 14th International Conference on Information, Intelligence, Systems & Applications (IISA)*, 1–9. https://doi.org/10.1109/IISA59645.2023.10345968

Peterson, M., & Armstrong, H. (2024). Developing visual conventions for explainable LLM outputs in intelligence analysis summaries [webpage]. https://textimage.org/uncertainty/

Prabhod, K. J. (2023). Integrating large language models for enhanced clinical decision support systems in modern healthcare. *Journal of Machine Learning for Healthcare Decision Support, 3*(1), 18–62. https://medlines.uk/index.php/JMLHDS/article/view/23

Prabhudesai, S., Yang, L., Asthana, S., Huan, X., Liao, Q. V., & Banovic, N. (2023). Understanding uncertainty: How lay decision-makers perceive and interpret uncertainty in human-AI decision making. *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 379–396. https://doi.org/10.1145/3581641.3584033

Rajashekar, N. C., Shin, Y. E., Pu, Y., Chung, S., You, K., Giuffre, M., Chan, C. E., Saarinen, T., Hsiao, A., Sekhon, J., Wong, A. H., Evans, L. V., Kizilcec, R. F., Laine, L., McCall, T., & Shung, D. (2024). Human-algorithmic interaction using a large language model-augmented artificial intelligence clinical decision support system. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–20. https://doi.org/10.1145/3613904.3642024

Rheu, M., Shin, J. Y., Peng, W., & Huh-Yoo, J. (2020). Systematic review: Trust-building factors and implications for conversational agent design. *International Journal of Human–Computer Interaction, 37*(1), 81–96. https://doi.org/10.1080/10447318.2020.1807710

Rosenholtz, R., Li, Y., & Nakano, L. (2007). Measuring visual clutter. *Journal of Vision, 7*(2), 17–22. https://doi.org/10.1167/7.2.17

Sacha, D., Senaratne, H., Kwon, B. C., Ellis, G., & Keim, D. A. (2016). The role of uncertainty, awareness, and trust in visual analytics. *IEEE Transactions on Visualization and Computer Graphics, 22*(1), 240–249. https://doi.org/10.1109/tvcg.2015.2467591

Savage, T., Wang, J., Gallo, R., Boukil, A., Patel, V., Safavi-Naini, S. A. A., Soroush, A., & Chen, J. H. (2025). Large language model uncertainty proxies: Discrimination and calibration for medical diagnosis and treatment. *Journal of the American Medical Informatics Association, 32*(1), 139–149. https://doi.org/10.1093/jamia/ocae254

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. *Proceedings 1996 IEEE Symposium on Visual Languages*, 336–343. https://doi.org/10.1109/VL.1996.545307

Skeels, M., Lee, B., Smith, G., & Robertson, G. G. (2010). Revealing uncertainty for information visualization. *Information Visualization, 9*(1), 70–81. https://doi.org/10.1145/1385569.1385637

Sorkin, R. D., & Woods, D. D. (1985). Systems with human monitors: A signal detection analysis. *Human–Computer Interaction, 1*(1), 49–75. https://doi.org/10.1207/s15327051hci0101_2

Sperrle, F., El-Assady, M., Guo, G., Borgo, R., Chau, D. H., Endert, A., & Keim, D. (2021). A survey of human-centered evaluations in human-centered machine learning. *Computer Graphics Forum, 40*(3), 543–568. https://doi.org/10.1111/cgf.14329

Stevenson, M. T. (2017). Assessing risk assessment in action. *Minnesota Law Review, 103*, 303. https://doi.org/10.2139/ssrn.3016088

Sultanum, N., Singh, D., Brudno, M., & Chevalier, F. (2019). Doccurate: A curation-based approach for clinical text visualization. *IEEE Transactions on Visualization and Computer Graphics, 25*(1), 142–151. https://doi.org/10.1109/tvcg.2018.2864905

Suresh, H., Lao, N., & Liccardi, I. (2020). Misplaced trust: Measuring the interference of machine learning in human decision-making. *Proceedings of the 12th ACM Conference on Web Science* (pp. 315–324). https://doi.org/10.1145/3394231.3397922

Taylor, R. A., Sangal, R. B., Smith, M. E., Haimovich, A. D., Rodman, A., Iscoe, M. S., Pavuluri, S. K., Rose, C., Janke, A. T., Wright, D. S., Socrates, V., & Declan, A. (2024). Leveraging artificial intelligence to reduce diagnostic errors in emergency medicine: Challenges, opportunities, and future directions. *Academic Emergency Medicine*. https://doi.org/10.1111/acem.15066

Thomson, J., Hetzler, E., MacEachren, A., Gahegan, M., & Pavel, M. (2005). A typology for visualizing uncertainty. *Visualization and Data Analysis 2005, 5669*, 146–157. https://doi.org/10.1117/12.587254

Tintarev, N., & Masthoff, J. (2007). Effective explanations of recommendations. *Proceedings of the 2007 ACM Conference on Recommender Systems* (pp. 153–156). https://doi.org/10.1145/1297231.1297259

Tomsett, R., Preece, A., Braines, D., Cerutti, F., Chakraborty, S., Srivastava, M., Pearson, G., & Kaplan, L. (2020). Rapid trust calibration through interpretable and uncertainty-aware AI. *Patterns, 1*(4), 100049. https://doi.org/10.1016/j.patter.2020.100049

Ullah, E., Parwani, A., Baig, M. M., & Singh, R. (2024). Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology – a recent scoping review. *Diagnostic Pathology, 19*(1), 43. https://doi.org/10.1186/s13000-024-01464-7

Vaghefi, S. A., Stammbach, D., Muccione, V., Bingler, J., Ni, J., Kraus, M., Allen, S., Colesanti-Senni, C., Wekhof, T., Schimanski, T., Gostlow, G., Yu, T., Wang, Q., Webersinke, N., Huggel, C., & Leippold, M. (2023). ChatClimate: Grounding conversational AI in climate science. *Communications Earth & Environment, 4*(1), 1–13. https://doi.org/10.1038/s43247-023-01084-x

Van de Merwe, K., Mallam, S., & Nazir, S. (2024). Agent transparency, situation awareness, mental workload, and operator performance: A systematic literature review. *Human Factors, 66*(1), 180–208. https://doi.org/10.1177/00187208221077804

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*. https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Weisz, J. D., He, J., Muller, M., Hoefer, G., Miles, R., & Geyer, W. (2024). Design principles for generative AI applications. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Article 378, pp. 1–22). Association for Computing Machinery. https://doi.org/10.1145/3613904.3642466

Wickens, C. D., Gempler, K., & Morphew, M. E. (2000). Workload and reliability of predictor displays in aircraft traffic avoidance. *Transportation Human Factors, 2*(2), 99–126. https://doi.org/10.1207/STHF0202_01

Williams, T., Briggs, P., & Scheutz, M. (2015). Covert robot-robot communication: Human perceptions and implications for human-robot interaction. *Journal of Human-Robot Interaction, 4*(2), 24–49. https://doi.org/10.5898/JHRI.4.2.Williams

Zerilli, J., Bhatt, U., & Weller, A. (2022). How transparency modulates trust in artificial intelligence. *Patterns, 3*(4), 100455. https://doi.org/10.1016/j.patter.2022.100455

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 295–305). https://doi.org/10.1145/3351095.3372852

Zhao, J., Wang, Y., Mancenido, M. V., Chiou, E. K., & Maciejewski, R. (2023). Evaluating the impact of uncertainty visualization on model reliance. *IEEE Transactions on Visualization and Computer Graphics, 1–15*. https://ieeexplore.ieee.org/document/10058545

Authors

Helen Armstrong is a professor of graphic and experience design and the director of the MGXD program at NC State University. Her research focuses on digital rights, human-machine teaming, and accessible design. Armstrong authored *Graphic Design Theory; Digital Design Theory*; and co-authored *Participate: Designing with User-Generated Content.* Her recent book, *Big Data, Big Design: Why Designers Should Care About Artificial Intelligence,* demystifies AI — specifically machine learning — while inspiring designers to harness this technology and establish leadership via thoughtful human-centered design. Armstrong is a past member of the AIGA National Board of Directors, the editorial board of *Design and Culture,* and a former chair of the AIGA Design Educators Community.

Ashley L. Anderson is an assistant professor of graphic design at Virginia Tech and a PhD in Design candidate at NC State University. Her research focuses on human-centered design and visual representation, particularly in the context of mental health and psychological intervention. She examines how design can shape and enhance the theories, processes, and methods used in psychological intervention.

Rebecca Planchart is a product designer at Pendo.io, a software experience management solution, where she supports enterprise platform and conversational AI initiatives. Her past research explored explainability and trust calibration in AI systems through UX and UI strategies. She is particularly interested in leveraging explainable AI to support users in high-stakes decision-making contexts.

Kweku Baidoo is a lecturer in graphic and experience design at NC State University. His work explores trust-centered design and visual strategies that support human understanding of complex AI systems. He is particularly interested in how AI-assisted decision-making can be designed to enhance appropriate user trust and performance in high-stakes domains such as healthcare.

Matthew Peterson is an associate professor of graphic and experience design at NC State University. His research focuses on visual representation in user interface design, recently including the facilitation of AI in intelligence analysis workflows through human-machine teaming, text-image integration in immersive user information systems, and the facilitation of scale cognition and numeracy in virtual environments.