LLM Forensic Evaluation: Diagnosing Actionability, Uncertainty, and Human Comprehension in High-Stakes Outputs

Jaye Nias[1], Saurav K. Aryal[1], Christopher Watson[1], Jeremy Blackstone[1], Simone Smarr[1], Lucretia Williams[1], Gloria Washington[1]

[1]Human-Centered AI Institute, Howard University, Washington, DC, USA
jaye.nias@howard.edu, saurav.aryal@howard.edu, christopher.watson@howard.edu, jeremy.m.blackstone@howard.edu, simone.smarr@howard.edu, lucretia.williams1@howard.edu, gloria.washington@howard.edu

Abstract

Large language models are increasingly incorporated into decision support workflows to summarize situations, propose actions, and communicate rationale. These capabilities are valuable in time-sensitive environments, but they also introduce risks related to hallucination, overconfidence, and contextual misalignment. This paper presents Project Comprehension, a forensic evaluation framework that treats language model outputs as artifacts for post hoc analysis rather than isolated successes or failures. Project Comprehension integrates structured empirical probing across operationally grounded scenarios with human-centered annotation instruments designed to capture interpretability and perceived uncertainty. We report early results from empirical testing and scale validation using a labeling set developed to support reliable forensic judgments of model behavior. We describe a failure mode taxonomy for reasoning and communication breakdowns, and we illustrate how forensic insights can inform assurance practices, trust calibration, and human autonomy teaming. The paper concludes with recommendations for building forensic readiness into language-enabled systems used in high-stakes decision support.

Introduction

Large language models (LLMs) are increasingly incorporated into decision-support workflows where information is incomplete, time is constrained, and the consequences of error are high. In these settings, models are not intended to replace human judgment, but to summarize information, propose actions, or support reasoning under uncertainty. However, the introduction of fluent generative outputs into such workflows raises a fundamental evaluation challenge. Many risks associated with LLM use do not stem from obvious factual errors, but from how responses are framed, interpreted, and trusted by human users. Outputs that appear coherent and confident can obscure uncertainty, omit critical assumptions, or invite overreliance, even when they are technically plausible. These concerns are well documented in prior work on human-centered AI and decision support, which shows that failures often arise from how information is framed and interpreted rather than from technical malfunction alone (Morrison et al. 1996; National Academies of Sciences, Engineering, and Medicine 2022). Existing evaluation approaches, which emphasize correctness, benchmark performance, or internal model behavior, offer limited visibility into these interpretive dynamics.

In this paper, we introduce Project Comprehension, a forensic evaluation framework for examining how language model outputs function as decision-relevant artifacts. Rather than treating outputs as isolated responses to be scored for accuracy, the framework supports post hoc inspection of how reasoning cues, uncertainty signaling, grounding, and actionability appear in practice and how they are likely to be interpreted by human decision-makers. Project Comprehension synthesizes established insights from human-centered AI, decision science, and uncertainty communication into a structured, multi-dimensional evaluation instrument designed to surface response-level risks that may not register as errors under conventional metrics. Through empirical testing and scale validation, we demonstrate how this approach reveals stable patterns in human judgment and exposes failure modes in which surface-level response quality masks missing or weak grounding. Empirical results show that surface-level response quality often masks weak or absent grounding, even in responses rated as plausible and actionable. Together, these results position forensic evaluation as a practical method for assessing trust, interpretability, and risk in language-enabled decision-support systems.

Background and Related Work

Research on decision making under stress has long emphasized that failures in high-stakes environments are rarely attributable to technology alone. The DARPA Tactical Decision Making Under Stress (TADMUS) program demonstrated that cognitive overload, time pressure, and ambiguity fundamentally shape how information is perceived, interpreted, and acted upon by human decision makers (Morrison et al. 1996). Subsequent work reinforced that effective decision support depends not only on information availability, but on how well systems align with human cognitive constraints, situational awareness, and sensemaking processes (Hutchins, Morrison, and Kelly 1996). These insights established a foundational principle that remains relevant today: decision quality is mediated by interpretation, not merely by correctness.

As AI-driven decision support systems entered military

and operational contexts, this principle carried forward. Prior work in military decision support emphasized procedural clarity, usability, and alignment with operational workflows as critical success factors (Louvieris, Gregoriades, and Garn 2010; Uziel 2020). At the same time, human factors research highlighted the risks of overreliance and inappropriate trust when automated systems present information with unwarranted confidence or insufficient transparency (Lee and See 2004). These concerns anticipated many of the challenges now observed in large language model (LLM)-based systems.

Recent advances in LLMs have reintroduced these issues in amplified form. Large language models are capable of producing fluent, coherent, and contextually appropriate responses even when underlying reasoning is incomplete or incorrect (Brown et al. 2020). Surveys of hallucination and factual inconsistency document how plausible language can obscure errors, fabricate details, or omit critical constraints without signaling uncertainty (Ji et al. 2023; Huang et al. 2025). Work on evaluation practices has further shown that traditional accuracy-focused metrics fail to capture these interpretive risks, particularly in applied or decision-oriented settings (Gehrmann, Clark, and Sellam 2023). As a result, concerns have shifted from whether models are correct to whether their outputs are understandable, trustworthy, and usable by humans.

Human-centered AI and explainable AI research offers partial guidance, emphasizing intelligibility, justification, and transparency as prerequisites for responsible system use (Doshi-Velez and Kim 2017; Miller 2019; Cambria et al. 2024). However, much of this work focuses on explaining model behavior rather than evaluating outputs as communicative artifacts. Studies of explanation quality and accountability underscore the importance of traceability and source signaling (Ribeiro, Singh, and Guestrin 2016; Selbst et al. 2019), while usability standards frame comprehension as a function of clarity, relevance, and cognitive support (International Organization for Standardization 2018). Yet these perspectives are rarely integrated into a unified evaluative structure for language model outputs.

Insights from pragmatics and interaction further illuminate why fluency alone is insufficient. Linguistic theory distinguishes surface coherence from meaningful communicative alignment, highlighting how plausibility can mislead when conversational norms are violated (Grice 1975). Conversational grounding research shows that understanding emerges through iterative clarification, repair, and shared context rather than one-shot exchanges (Clark and Brennan 1991). Work on mixed-initiative systems and conversational agents similarly demonstrates that effective interaction depends on a system's ability to invite follow-up, support refinement, and adapt to user needs (Horvitz 1999; Luger and Sellen 2016).

Finally, research on uncertainty communication and calibration underscores the importance of explicitly signaling limitations and confidence. Studies in risk communication show that users make better decisions when uncertainty is conveyed in interpretable forms rather than hidden behind precise language (Gigerenzer et al. 2007). Visualization and presentation choices strongly influence perceived reliability and ordering judgments (Hullman, Resnick, and Adar 2015), while calibration studies reveal persistent gaps between model confidence and actual correctness (Jiang et al. 2021). Together, this work highlights the danger of fluent outputs that suppress uncertainty cues and invite overtrust.

Taken together, these bodies of work converge on a common gap: while prior research has identified many factors that influence human understanding and trust, there is limited work that operationalizes these insights into a coherent framework for evaluating language model outputs as decision-relevant artifacts. Project Comprehension builds on this foundation by synthesizing established evaluative lenses into a structured forensic framework that supports systematic inspection of plausibility, comprehension support, transparency, actionability, interactional affordances, and uncertainty signaling in LLM outputs.

Project Comprehension: Forensic Evaluation Framework

In this paper, we use the term ***LLM forensics*** deliberately, while recognizing that it is not yet a widely established label in the literature. Here, forensics refers to ***post hoc inspection of language model outputs as evidence of system behavior***, rather than attempts to explain internal model mechanisms or optimize performance. The focus is on reconstructing how failures, ambiguities, and confidence cues manifest in outputs, and how those signals are likely to be interpreted by human decision makers. This framing reflects long-standing concerns in human–AI teaming, human factors, and assurance research around use, misuse, and interpretive risk in applied systems, particularly in high-stakes contexts where errors may not be immediately visible (National Academies of Sciences, Engineering, and Medicine 2022; Probasco et al. 2025). Unlike work that treats forensic analysis primarily as attribution or misuse detection, Project Comprehension adopts an evidence-centered, interpretive approach in which model outputs are examined as artifacts relevant to trust, understanding, and decision support (Cambria et al. 2024; Steyvers and Kumar 2024).

Core Forensic Framing

Project Comprehension is designed as a forensic evaluation framework for examining how large language models behave when embedded in decision support workflows. Rather than treating outputs as isolated responses to be scored for correctness, the framework treats each output as an artifact that can be examined after the fact to understand how reasoning, confidence signaling, and contextual alignment emerge in practice. Insights from a prior pilot study are abstracted here into a reusable evaluation structure that emphasizes interpretability, reliability, and trust calibration rather than performance optimization. The framework supports systematic inspection of externally observable behavior and its implications for human decision making. Figure 1 illustrates how Project Comprehension operationalizes these concerns into interpretable evaluation dimensions.

Dimension	Primary Scholarly Traditions	Representative Constructs	Example References
PU	NLP evaluation; trust in automation; linguistic pragmatics	Plausibility vs. correctness; surface credibility; coherence	Grice (1975); Lee & See (2004); Gehrmann et al. (2023)
PP	Risk communication; uncertainty visualization; calibration	Confidence signaling; epistemic uncertainty; overtrust	Gigerenzer et al. (2007); Hullman et al. (2015); Jiang et al. (2021)
CS	Human-centered AI; explainable AI; usability theory	Intelligibility; explanation quality; cognitive scaffolding	Norman (2013); Doshi-Velez & Kim (2017); ISO 9241
GT	Explainable AI; accountability; epistemic transparency	Justification; traceability; source signaling	Ribeiro et al. (2016); Miller (2019); Selbst et al. (2019)
AC	Decision support systems; applied HCI	Procedural clarity; task utility; decision readiness	Endsley (1995); Amershi et al. (2019); Shneiderman (2022)
FW	Conversational AI; mixed-initiative interaction; sensemaking	Dialog grounding; repair; iterative refinement	Clark & Brennan (1991); Horvitz (1999); Luger & Sellen (2016)

Table 1: Epistemic grounding of Project Comprehension evaluation dimensions.

Epistemic Grounding of Evaluation Dimensions

The evaluation dimensions used in Project Comprehension are not proposed as novel psychological constructs, but as an integrative synthesis of recurring concerns in prior work on human-centered AI, explainable systems, decision support, conversational interaction, and uncertainty communication. Each dimension reflects a well-established evaluative lens that has been adapted here to support systematic forensic inspection of language model outputs. Table 1 situates each dimension within its primary scholarly lineage and representative constructs.

Methods

Project Comprehension employs a mixed-method evaluation approach designed to support forensic inspection of language model outputs. Rather than benchmarking performance, the methods focus on how outputs are constructed and how they are likely to be interpreted by human decision makers in applied decision support contexts.

Scenario Set and Data Generation

The evaluation scenarios used in Project Comprehension were designed to reflect applied decision-support contexts in which language models are expected to summarize information, propose procedural steps, or support reasoning under uncertainty. Scenarios were motivated by high-stakes settings in which model outputs may influence human judgment, but were not intended to simulate operational doctrine or assess task performance. Instead, they were constructed to elicit behaviors relevant to interpretability, confidence signaling, and contextual alignment.

To ground this evaluation, the prompt set included representative decision-support cases drawn from maritime and military-adjacent contexts, including:

- Determining whether an approaching object constitutes a ship based on standard operating procedures.
- Deciding whether to engage or withdraw when encountering a friendly vessel.
- Evaluating engagement or retreat strategies when facing an adversarial ship.

Scenario prompts incorporated role framing, task objectives, and contextual constraints, while allowing for variation in ambiguity, completeness of information, and required specificity. Model outputs were generated under consistent prompt conditions and recorded as prompt-response pairs with associated metadata to support traceability during annotation and analysis.

Annotation Scheme and Labeling Workflow

Human-centered annotation is central to the Project Comprehension framework. A structured rubric was used to capture interpretive dimensions of model outputs that are not reflected in accuracy-based metrics. The rubric defines dimensions including perceived plausibility, grounding transparency, actionability, and perceived uncertainty. Full rubric definitions and labeling instructions are provided in Appendix A.

Annotators applied the rubric independently based on how each output would likely be interpreted by a human decision maker, rather than by identifying specific factual errors. Each annotated item was associated with a unique identifier, annotator identifier, and timestamp to support consistency and traceability. Annotation guidelines were refined iteratively to address ambiguities observed during labeling.

Scale Refinement and Validation

Exploratory analyses were conducted to assess whether annotation dimensions functioned consistently across annotators. Inter-rater reliability analyses appropriate for ordinal labels were used to examine agreement patterns. These analyses were not intended to establish formal psychometric validity, but to assess whether the rubric could be applied as a stable forensic instrument. Relationships among rubric dimensions were also examined to identify potential overlap or redundancy. Observations from these analyses informed iterative refinement of scale definitions.

Failure Mode Synthesis

Failure mode synthesis focused on identifying recurring patterns in how model outputs behave and how those behaviors are interpreted by human reviewers. Annotated outputs were examined to characterize patterns involving plausibility, grounding, confidence signaling, and actionability.

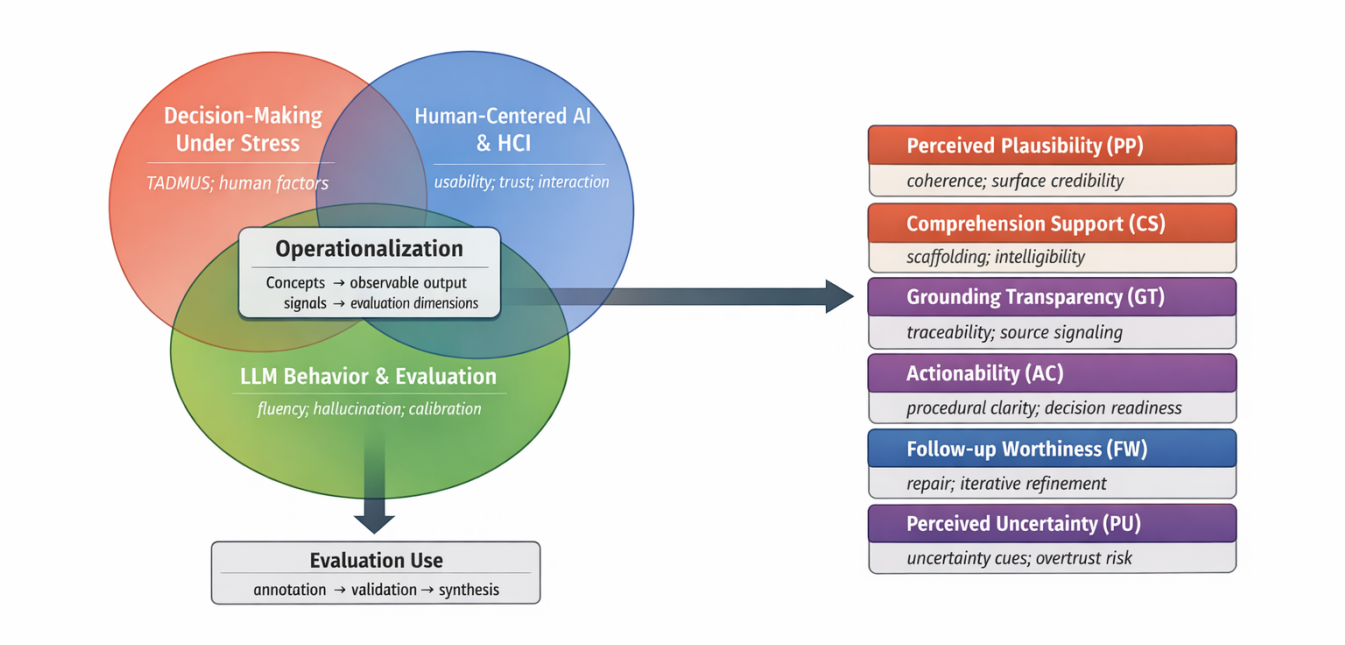

Figure 1: Project Comprehension framework operationalization

Quantitative summaries were used to guide inspection, but the identification of failure modes relied on qualitative judgment grounded in applied decision support use. This synthesis process emphasizes patterns of interpretation rather than isolated errors.

Results

This section reports empirical observations from Project Comprehension based on annotated model outputs and repeated empirical testing. The goal of these results is not to benchmark overall model performance, but to characterize patterns that emerge when language model responses are examined across multiple scenarios, repeated runs, and human judgments. We first describe the annotated dataset, then report reliability and distributional properties of the evaluation dimensions, followed by empirical findings related to response stability and recurring failure modes. Throughout, we focus on behaviors that are likely to matter in applied decision support contexts.

Dataset Overview

The dataset consists of 240 unique prompt–response items, each evaluated across six dimensions: *Perceived Uncertainty* (PU), *Perceived Plausibility* (PP), *Comprehension Support* (CS), *Grounding Transparency* (GT), *Actionability* (AC), and *Follow-up Worthiness* (FW). In total, 960 individual ratings were collected using five-point Likert scales, with an average of approximately 3.25 raters per item and up to five raters for some items. A subset of items was re-rated by the same annotators to support assessment of rating stability over time.

Table 2 summarizes item-level mean scores across all dimensions. Mean values varied in ways that align with the conceptual intent of each dimension. *Perceived Uncertainty* exhibited a lower mean (2.22), indicating that uncertainty was present for a subset of responses rather than uniformly expressed. In contrast, *Perceived Plausibility*, *Comprehension Support*, *Actionability*, and *Follow-up Worthiness* fell in the moderately positive range (means between 3.4 and 4.2). *Grounding Transparency* showed a lower and more neutral mean (2.43), reflecting variability in whether responses explicitly conveyed assumptions, sources, or reasoning.

All dimensions made use of the full response range, with no substantial floor or ceiling effects. These distributional properties suggest that the scale differentiates meaningfully across prompt–response pairs and is sensitive to variation in model behavior.

Instrument Validation: Reliability Across Dimensions

Inter-rater reliability (IRR) was assessed to examine the consistency of judgments across annotators for each evaluation dimension. Each prompt–response pair was independently rated by multiple annotators, with rater assignments intentionally unbalanced. Given the ordinal nature of the ratings and variable rater overlap, Krippendorff's Alpha (ordinal) was used as the primary global reliability metric. Reliability was computed separately for each dimension.

Across the dataset, each prompt was rated by between three and five annotators, though overlap between specific rater pairs varied substantially. As a result, IRR values are best interpreted as reflecting relative convergence across dimensions rather than absolute agreement thresholds.

Krippendorff's Alpha (Ordinal) Krippendorff's Alpha values ranged from 0.04 to 0.23 across dimensions (Table 2).

Dimension	Krippendorff's α	Avg. Weighted κ
PU	0.039	0.602
PP	0.223	0.502
CS	0.104	0.000
GT	0.132	0.116
AC	0.233	0.756
FW	0.188	0.225

Table 2: Inter-rater reliability by evaluation dimension.

Dimension	Floor (1)	Ceiling (5)	Extreme Total
PU	0.42%	1.25%	1.67%
PP	4.17%	2.50%	6.67%
CS	0.00%	11.25%	11.25%
GT	4.58%	0.00%	4.58%
AC	7.08%	1.25%	8.33%
FW	0.00%	0.42%	0.42%

Table 3: Floor and ceiling effects by dimension

Dimension	Skew	Shape Interpretation
PU	1.40	Strong skew
PP	-1.62	Strong skew
CS	-1.15	Strong skew
GT	0.27	Approximately symmetric
AC	-1.30	Strong skew
FW	-0.97	Mild skew

Table 4: Distributional skew of evaluation dimensions

Dimensions related to *Actionability* and *Perceived Plausibility* exhibited higher alpha values, while *Comprehension Support* and *Grounding Transparency* showed lower agreement. These differences indicate that some dimensions elicited more convergent judgments than others under sparse rater overlap.

Pairwise Agreement To complement the global alpha estimates, pairwise agreement was examined using weighted Cohen's κ (linear) for rater pairs with at least five shared items. Average pairwise κ values mirrored the alpha patterns, with higher agreement for *Actionability* and *Perceived Plausibility* and lower agreement for *Comprehension Support* and *Grounding Transparency*. These results indicate that when raters evaluated the same outputs, judgments related to operational clarity and plausibility were more likely to converge than judgments involving explanation quality or grounding.

Reliability Pattern Summary Across reliability analyses, a consistent pattern emerged. Dimensions tied to more directly observable response properties, such as whether clear actions were proposed or recommendations appeared plausible, exhibited higher agreement. Dimensions that required interpretation of implicit reasoning, assumptions, or informational sufficiency showed lower convergence. Rather than treating lower agreement as annotation failure, we interpret these patterns as reflecting genuine interpretive ambiguity in how certain model behaviors are perceived by human reviewers. In applied decision support contexts, these are precisely the conditions under which language model outputs may be most difficult to assess and most consequential to interpret.

Instrument Validation: Scale Sensitivity and Distributional Properties

To evaluate whether the evaluation instrument was sufficiently sensitive for downstream forensic analysis, we examined distributional properties of item-level mean scores averaged across annotators. This analysis assesses whether each dimension makes effective use of the response range and differentiates among prompt–response pairs independent of rater-level noise.

Across dimensions, item-level means exhibited moderate to high variability (Table 4), with standard deviations ranging from 0.64 to 0.97. This dispersion indicates that the scale is neither compressed nor inert and that it meaningfully differentiates among outputs. Mean score patterns were also consistent with the conceptual intent of each dimension: *Perceived Uncertainty* exhibited a relatively low mean (M=2.22), indicating that uncertainty was identified selectively rather than uniformly, while *Perceived Plausibility*, *Comprehension Support*, *Actionability*, and *Follow-up Worthiness* fell in the moderately to highly positive range (M=3.43–4.21). *Grounding Transparency* occupied a lower and more neutral range (M=2.45), reflecting variability in whether responses explicitly conveyed assumptions, sources, or reasoning.

Floor and ceiling effects were minimal across all dimensions (Table 3). Total extreme mean scores ranged from 0.42% (FW) to 11.25% (CS), with the only notable tendency being a mild ceiling effect for *Comprehension Support* (11.25% at the maximum score). Distributional shape further supported scale sensitivity and theoretical coherence (Figure 2; Table 4). *Perceived Uncertainty* showed strong right skew (skew=1.40), consistent with uncertainty being present for a subset of items. In contrast, *Perceived Plausibility*, *Comprehension Support*, and *Actionability* were left-skewed (skew=-1.62, -1.15, and -1.30 respectively), reflecting generally favorable evaluations while preserving meaningful variation below the ceiling. *Grounding Transparency* was approximately symmetric (skew=0.27), and *Follow-up Worthiness* showed mild skew (skew=-0.97), indicating stable but flexible judgments.

Taken together, these distributional analyses indicate that the instrument uses the response range effectively, avoids problematic floor or ceiling effects, and exhibits variability and shape consistent with the intended meaning of each dimension. These properties provide evidence that the evaluation dimensions are sensitive and interpretable for use as a diagnostic tool in the forensic analysis of language model outputs.

Surface Quality and Grounding Transparency

To examine whether responses that appear strong on surface-level qualities are also well grounded, we conducted an item-level analysis comparing a composite *Surface Quality* score to *Grounding Transparency (GT)*. Surface Quality was operationalized as the mean of *Perceived Plausibility (PP)*, *Com-*

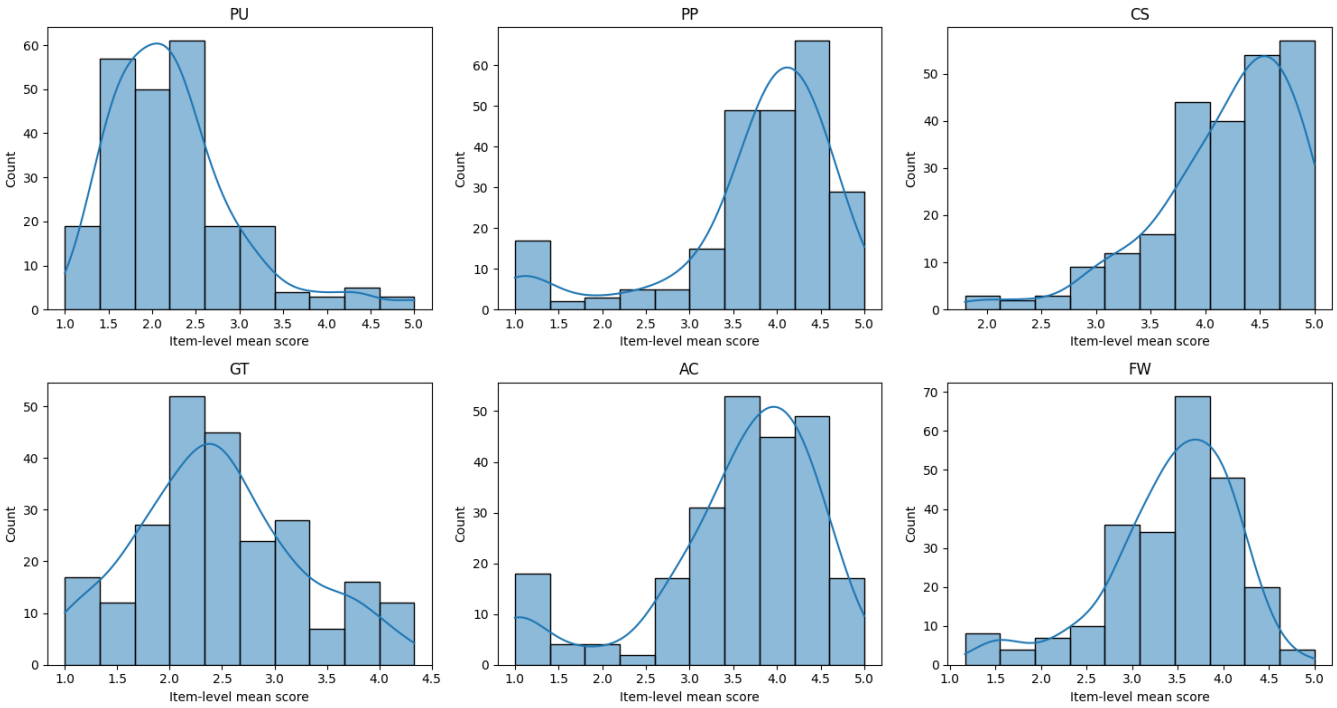

Figure 2: Distribution of item-level mean scores for each evaluation dimension.

prehension Support (CS), and *Actionability (AC)*, averaged across annotators for each prompt–response item.

A Spearman rank-order correlation indicated a moderate positive association between Surface Quality and Grounding Transparency ($\rho = 0.54$, $p < .001$). While higher surface-quality responses tended to exhibit higher grounding transparency on average, substantial variability remained, particularly among items with high surface-quality scores.

Figure 3 illustrates this relationship using a scatter plot of item-level means. At high levels of Surface Quality, Grounding Transparency spans a wide range, indicating that fluent, plausible, and actionable responses are not consistently well grounded. Quadrant analysis further revealed that 14.6% of items fell into a *high surface-quality but low grounding* region, representing responses that appeared strong while providing limited transparency regarding assumptions or informational basis. To assess this pattern conservatively, items were also grouped into tertiles based on Surface Quality. Mean Grounding Transparency increased across bins; however, the increase was gradual rather than proportional. Even among the highest surface-quality items, average grounding transparency remained moderate, reinforcing the observation that surface quality does not reliably predict explicit grounding.

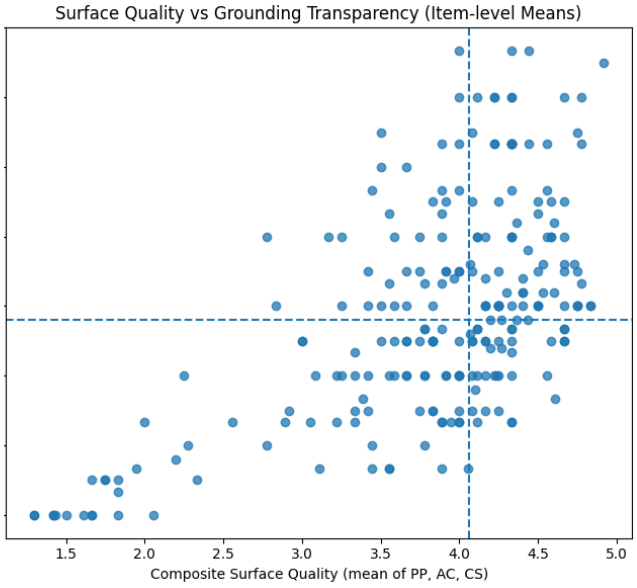

Figure 3: Relationship between composite surface quality and grounding transparency

Discussion

This work examines how language model outputs are interpreted by human reviewers when used in decision-relevant contexts, using a forensic evaluation framework that emphasizes observable response behavior rather than aggre-

gate model performance. Across analyses, the results indicate that human judgments of language model outputs are structured and patterned, not random. Reviewers consistently formed overall impressions of response quality that influenced multiple evaluative dimensions simultaneously, suggesting that surface-level characteristics such as fluency, structure, and tone play a significant role in shaping interpretation.

Variation in inter-rater agreement further clarifies which aspects of language model outputs are more easily and consistently assessed. Dimensions related to Actionability and Perceived Plausibility exhibited higher agreement, indicating that reviewers generally converged on whether responses proposed clear steps or appeared reasonable. In contrast, lower agreement was observed for Comprehension Support and Grounding Transparency, where judgments required interpretation of implicit reasoning, assumptions, or informational sufficiency. Rather than indicating measurement failure, this pattern reflects genuine interpretive ambiguity. From a forensic perspective, dimensions with lower agreement may represent areas of higher interpretive risk, where users are more likely to diverge in how they assess model outputs.

Distributional analyses show that the evaluation instrument makes effective use of the response scale and differentiates meaningfully among prompt–response pairs. Scores were not concentrated at the extremes, and the observed distributional shapes aligned with the conceptual intent of each dimension. Perceived Uncertainty was selectively present rather than uniformly expressed, while Perceived Plausibility, Comprehension Support, and Actionability were generally rated favorably while still exhibiting meaningful variation. Grounding Transparency clustered closer to the midpoint, reflecting inconsistent signaling of sources, assumptions, or justification across responses. Together, these patterns indicate that the scale is sensitive, interpretable, and appropriate for diagnostic analysis of response-level behavior.

A key empirical finding of this study is that surface-level response quality can mask missing or weak grounding. While surface-quality dimensions were moderately correlated with Grounding Transparency, the relationship was not strong enough to treat surface impressions as a reliable proxy for explanatory adequacy. Item-level, quadrant, and binned analyses revealed substantial variability in grounding among responses that appeared clear, plausible, and actionable. A non-trivial subset of responses occupied a high surface-quality, low grounding region, indicating that polished presentation can obscure the absence of explicit justification or transparency. This masking effect provides quantitative support for concerns raised in prior work regarding overtrust in fluent generative outputs.

Failure mode patterns in this analysis emerged through recurring combinations of evaluation dimensions rather than uniformly poor performance. In particular, responses that scored highly on surface-quality dimensions while exhibiting low grounding transparency represent a distinct and interpretable configuration of risk. These patterns were identified through inspection of dimensional co-occurrence rather than through post hoc labeling or manual coding, illustrating how forensic evaluation can surface response-level vulnerabilities without introducing additional annotation layers.

Implications

The findings have direct implications for the design and evaluation of language-enabled decision-support systems. Responses may rely on retrieved or externally sourced information without clearly exposing that reliance to users, allowing surface-level coherence to obscure the informational basis of the output. Retrieval-augmented architectures offer mechanisms such as source attribution, evidence spans, or retrieval confidence that align directly with the Grounding Transparency dimension identified in this framework. Importantly, the evaluation instrument enables such mechanisms to be assessed empirically rather than assumed to improve transparency or trust.

The results also inform how language model outputs should be evaluated and interpreted by human reviewers. Holistic judgments based on overall impression can overlook critical weaknesses when responses appear confident or well structured. Separating evaluation into distinct dimensions encourages closer inspection of specific response properties and helps expose weaknesses that might otherwise remain hidden. In decision-relevant and human-autonomy teaming contexts, explanations that appear authoritative without adequate grounding may undermine effective oversight. Forensic-style evaluation provides a structured approach for identifying when explanations support human judgment and when they introduce interpretive risk.

Limitations

This study employed an uneven and partially overlapping rater design, which limits absolute estimates of inter-rater reliability. The rater pool was modest in size, and not all prompt–response items were evaluated by the same reviewers. Each evaluation dimension was measured using a single-item rating, prioritizing usability and interpretability over fine-grained psychometric modeling. These choices reflect intentional tradeoffs aligned with the goal of diagnostic, response-level analysis rather than comprehensive measurement of latent traits. The findings characterize how human reviewers interpret model outputs under this framework and should not be interpreted as comparative claims about specific model architectures.

Conclusion

This paper introduces Project Comprehension, a forensic evaluation framework for examining how language model outputs function as decision-relevant artifacts. By focusing on how responses are interpreted rather than on aggregate performance, the framework reveals stable patterns in human judgment and surfaces response-level risks that may not register as errors under conventional evaluation metrics. The results demonstrate that surface-level qualities exert strong influence on perceived reliability and that high-quality presentation can mask missing or weak grounding.

By separating surface quality from grounding transparency and related dimensions, the framework enables more precise identification of interpretive risk. This approach provides a practical safeguard against over-reliance on fluent generative outputs and offers a method for empirically evaluating design choices intended to improve transparency and trust. Together, these contributions position forensic evaluation as a valuable complement to existing assessment approaches for language-enabled decision-support systems.

A: Project Comprehension Evaluation Instrument and Interpretation Guidance

This appendix documents the evaluation instrument used in Project Comprehension and provides guidance on how resulting scores should be interpreted and applied. The instrument is designed to support forensic inspection of language model outputs in applied decision-support contexts, rather than performance benchmarking or model ranking.

Evaluation Dimensions

Each prompt–response pair was evaluated independently by human reviewers using six single-item dimensions. All dimensions were rated on five-point Likert scales (1 = very low, 5 = very high). Reviewers were instructed to assess each dimension based on how the response would likely be interpreted by a human decision-maker, rather than on factual correctness alone. All dimensions were intentionally designed as single-item judgments to preserve interpretive flexibility and reduce reviewer burden. The instrument prioritizes transparency and diagnostic value over strict psychometric optimization.

Interpretation and Use Guidance

Scores produced by the Project Comprehension instrument are intended to be interpreted dimensionally rather than aggregated into a single metric. Each dimension captures a distinct aspect of how a language model response may be understood and acted upon by a human decision-maker.

In practice, the instrument is used by examining profiles of scores across dimensions for individual prompt–response pairs or sets of related items. Particular attention should be paid to divergence across dimensions, such as responses that score highly on perceived plausibility or actionability but low on grounding transparency or uncertainty signaling. Such patterns may indicate interpretive risk even when overall impressions are favorable. This instrument is especially well-suited for comparative analysis across scenarios, prompt formulations, or repeated model runs. Changes in dimensional profiles across these conditions may reveal instability, sensitivity to framing, or shifts in how uncertainty and assumptions are communicated.

The Project Comprehension instrument is not designed to establish normative thresholds, compute composite quality scores, or rank models. Instead, it supports forensic evaluation by surfacing structured signals about interpretability, uncertainty, and decision relevance that require analyst judgment and contextual expertise. Divergence across dimensions is treated as an informative signal rather than as measurement error.

Dimension	Prompt to Reviewer	Scale Anchors
Perceived Uncertainty (PU)	To what extent does the response clearly communicate uncertainty, ambiguity, or limitations in the information provided?	1 = No uncertainty signaled 5 = Uncertainty clearly and appropriately communicated
Perceived Plausibility (PP)	How plausible and internally coherent does the response appear given the prompt and context?	1 = Implausible or incoherent 5 = Highly plausible and coherent
Comprehension Support (CS)	To what extent does the response support understanding by explaining reasoning, structure, or relevant context?	1 = Little or no support for understanding 5 = Strong support for understanding
Grounding Transparency (GT)	How clearly does the response indicate sources, assumptions, or the basis for its claims?	1 = No grounding cues 5 = Explicit and transparent grounding
Actionability (AC)	How well does the response support concrete or informed action by a decision-maker?	1 = Not actionable 5 = Highly actionable
Follow-up Worthiness (FW)	To what extent does the response invite or justify further inquiry, clarification, or engagement?	1 = No clear reason to follow up 5 = Strongly warrants follow-up

Table 5: Project Comprehension evaluation dimensions and interpretive rating anchors. All dimensions are rated on five-point Likert scales (1 = very low, 5 = very high).

Acknowledgments

This material is based upon work supported by the Office of Naval Research under Grant No. N00014-22-1-2714. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Office of Naval Research, the Department of Defense or the Department of War.

References

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; et al. 2020. Language Models Are Few-Shot Learners. *Advances in Neural Information Processing Systems*. GPT-3 demonstration of fluency and few-shot capabilities.

Cambria, E.; Malandri, L.; Mercorio, F.; Nobani, N.; and Seveso, A. 2024. XAI Meets LLMs: A Survey of the Relation Between Explainable AI and Large Language Models. *arXiv preprint arXiv:2407.15248*.

Clark, H. H.; and Brennan, S. E. 1991. Grounding in Communication. In Resnick, L. B.; Levine, J. M.; and Teasley, S. D., eds., *Perspectives on Socially Shared Cognition*, 127–149. Washington, DC: American Psychological Association.

Doshi-Velez, F.; and Kim, B. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Gehrmann, S.; Clark, E.; and Sellam, T. 2023. Repairing the Cracked Foundation: A Survey of Obstacles in Evaluation Practices for Generated Text. *Journal of Artificial Intelligence Research*.

Gigerenzer, G.; Gaissmaier, W.; Kurz-Milcke, E.; Schwartz, L. M.; and Woloshin, S. 2007. Helping Doctors and Patients Make Sense of Health Statistics. *Psychological Science in the Public Interest*, 8(2): 53–96.

Grice, H. P. 1975. Logic and conversation. In *Speech acts*, 41–58. Brill.

Horvitz, E. 1999. Principles of Mixed-Initiative User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '99)*, 159–166. ACM.

Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; and Liu, T. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.*, 43(2).

Hullman, J.; Resnick, P.; and Adar, E. 2015. Hypothetical Outcome Plots Outperform Error Bars and Violin Plots for Inferences about Reliability of Variable Ordering. *PLOS ONE*, 10(11): e0142444.

Hutchins, S. G.; Morrison, J. G.; and Kelly, R. T. 1996. Principles for Aiding Complex Military Decision Making. Technical report, U.S. Army Research Institute for the Behavioral and Social Sciences, Alexandria, VA. Technical report; hosted by CORE.

International Organization for Standardization. 2018. ISO 9241-11:2018 Ergonomics of Human-System Interaction – Part 11: Usability: Definitions and Concepts.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.*, 55(12).

Jiang, Z.; Araki, J.; Ding, H.; and Neubig, G. 2021. How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering. *Transactions of the Association for Computational Linguistics*.

Lee, J. D.; and See, K. A. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, 46(1): 50–80.

Louvieris, P.; Gregoriades, A.; and Garn, W. 2010. Assessing critical success factors for military decision support. *Expert Systems with Applications*, 37(12): 8229–8241.

Luger, E.; and Sellen, A. 2016. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*, 5286–5297. ACM.

Miller, T. 2019. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence*, 267: 1–38.

Morrison, J. G.; Kelly, R. T.; Moore, R. A.; and Hutchins, S. G. 1996. Tactical decision making under stress (TADMUS) decision support system.

National Academies of Sciences, Engineering, and Medicine. 2022. *State of the Art and Research Needs*. Washington, DC: The National Academies Press.

Probasco, E. S.; Toner, H.; Burtell, M.; and Rudner, T. G. J. 2025. AI for Military Decision-Making: Harnessing the Advantages and Avoiding the Risks. Technical report, Center for Security and Emerging Technology (CSET), Georgetown University, Washington, DC. Policy report.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. ACM.

Selbst, A. D.; Boyd, D.; Friedler, S. A.; Venkatasubramanian, S.; and Vertesi, J. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, 59–68. ACM.

Steyvers, M.; and Kumar, A. 2024. Three challenges for AI-assisted decision-making. *Perspectives on Psychological Science*, 19(5): 722–734.

Uziel, S. J. 2020. *AI-AUGMENTED DECISION SUPPORT SYSTEMS: APPLICATION IN MARITIME DECISION MAKING UNDER CONDITIONS OF METOC UNCERTAINTY*. Ph.D. thesis, Monterey, CA; Naval Postgraduate School.