The XAI-Seeking Principle: Structuring Explainability for LLM-Mediated Decision Support

Valentin Grimm
valentin.grimm@th-owl.de
Institute Industrial IT
Ostwestfalen-Lippe University of
Applied Sciences and Arts
Lemgo, Germany

Jessica Rubart
jessica.rubart@th-owl.de
Institute Industrial IT
Ostwestfalen-Lippe University of
Applied Sciences and Arts
Lemgo, Germany

Eelco Herder
e.herder@uu.nl
Information and Computing Sciences
Utrecht University
Utrecht, The Netherlands

Carsten Röcker
carsten.roecker@th-owl.de
Institute Industrial IT
Ostwestfalen-Lippe University of
Applied Sciences and Arts
Lemgo, Germany

Abstract

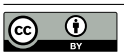
We investigate how LLM-mediated explanations can preserve evidence trails, support orientation, and reduce interpretation effort under cognitive and temporal constraints. While LLMs can make XAI artifacts accessible, fluent summaries may obscure provenance and hinder verification. We introduce the XAI-Seeking Principle (XSP), a hypertextual design principle that structures explainability as linked abstraction layers: overview-links connect low-level evidence to higher-level summaries, while detail-links preserve inspectable evidence. Higher levels support rapid action; lower levels enable justification and verification. XSP guides transformation monitoring, feature alignment, performance–latency trade-offs, and validation-driven refinement. We examine it in a misinformation-support prototype for social media posts. Results reveal layer-specific effects of model scale and reasoning, trade-offs between semantic grounding and assessment stability, and the value of validation signals for adapting transformations. XSP thus provides a design principle for constructing explainability as a navigable representation space, with technical evidence on how abstraction layers, feature grounding, latency, and validation interact in LLM-mediated XAI.

CCS Concepts

• **Computing methodologies** → **Information extraction**; **Natural language generation**; • **Human-centered computing** → **Empirical studies in interaction design**; • **Information systems** → **Decision support systems**.

Keywords

Large Language Model Mediation, Explainable AI, Decision Co-Pilot Systems, Misinformation Detection



HT '26, London, United Kingdom

ACM ISBN 979-8-4007-2564-7/2026/09
https://doi.org/10.1145/3800935.3830869

ACM Reference Format:
Valentin Grimm, Jessica Rubart, Eelco Herder, and Carsten Röcker. 2026. The XAI-Seeking Principle: Structuring Explainability for LLM-Mediated Decision Support. In *37th ACM Conference on Hypertext and Social Media (HT '26), September 14–18, 2026, London, United Kingdom.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3800935.3830869

1 Introduction

Large Language Model (LLM)-mediated decision-support systems (DSSs) transform technical evidence into fluent, high-level explanations. While this can reduce cognitive effort, it also changes how users navigate, verify, and contest information. From a hypertext perspective, this is a problem of orientation, provenance, and inspectable trails (cf. [10]).

In domains such as content moderation, information filtering, and operational decision-making, users often need to act on uncertain model outputs under limited attention [22]. In such settings, the bottleneck is whether users can understand what the system is claiming, why it claims this, and how to inspect the underlying evidence.

In this paper, we introduce the XAI-Seeking Principle (XSP), a hypertextual design principle that organizes LLM-mediated explanations as linked abstraction layers. Overview-links synthesize low-level XAI artifacts into higher-level summaries, while detail-links allow users to inspect the evidence behind these summaries.

Classical Explainable AI (XAI) artifacts can expose technically relevant evidence, but they rarely define how users should move from a high-level recommendation to lower-level evidence, or how intermediate summaries should remain linked to the artifacts they summarize [13].

LLMs can turn technical XAI artifacts into accessible summaries, but they can also collapse multiple evidence sources into a single fluent statement. This may hide the path from evidence to assessment and make it harder for users to verify whether a summary faithfully reflects the underlying model behavior (cf. [10]).

Prior work on conversational navigation shows that LLM interfaces can obscure links, trails, provenance, and orientation cues

that are normally central to hypertextual sensemaking [10]. XAI-seeking addresses this by making the transformations between evidence, summaries, and assessments explicit and navigable.

The XSP conceptualizes explainability as a layered transformation pipeline, in which low-level model artifacts are incrementally synthesized into higher-level, actionable representations. These representations serve as decision points that allow users to act quickly or selectively access deeper evidence when needed. From this structure, we derive four design characteristics for XAI-seeking systems: monitoring transformation quality, aligning features with the problem space, balancing explanation quality and latency, and using validation signals for targeted refinement.

We instantiate XSP in a misinformation-support prototype for social media posts and use it as a technical proof of concept to examine selected design characteristics. The technical evaluation examines how model scale and reasoning affect different abstraction layers, how feature design influences explanation quality and assessment stability, and how validation signals can diagnose divergence patterns and guide prompt adaptation.

The three contributions of this paper are as follows:

(1) Conceptual: XSP as a hypertextual design principle for structuring LLM-mediated explainability as linked abstraction layers.
(2) Design: Four operational characteristics for building XAI-seeking systems: monitoring/validation, feature-problem relation, performance-latency trade-off, and validation-driven improvement.
(3) Technical: A misinformation proof-of-concept showing layer-specific effects of model scale/reasoning, feature-grounding trade-offs, and validation-driven adaptation.

2 Related Work

Early approaches to XAI focused on post-hoc techniques that make opaque model predictions more interpretable to humans. Methods such as LIME [11], SHAP [8] and Anchors [12] expose local or global evidence, but this evidence is often technical, fragmented and difficult to use under time constraints and does not by itself define how users move between overview, intermediate interpretation, and detailed inspection.

Recent LLM-XAI systems show that language models can translate feature attributions, counterfactuals, or other XAI artifacts into more accessible explanations [17, 18]. However, these systems often focus on generating or querying explanations, while the relation between abstraction layers, evidence trails, validation signals, and interaction structure remains under-specified.

LLM-mediated explanations introduce a specific validation problem: generated rationales may appear plausible while failing to faithfully reflect the model evidence on which they are based [1]. Prior work has therefore proposed automatic and LLM-based checks for hallucination, factuality, and rationale quality [5, 9, 16], but these approaches often evaluate outputs at the level of individual responses, models, or benchmarks rather than as part of an interactive explanation structure. In XSP, validation is instead treated as a structural mechanism for inspecting transformations between abstraction layers to support scalable diagnosis and refinement of the explanation pipeline.

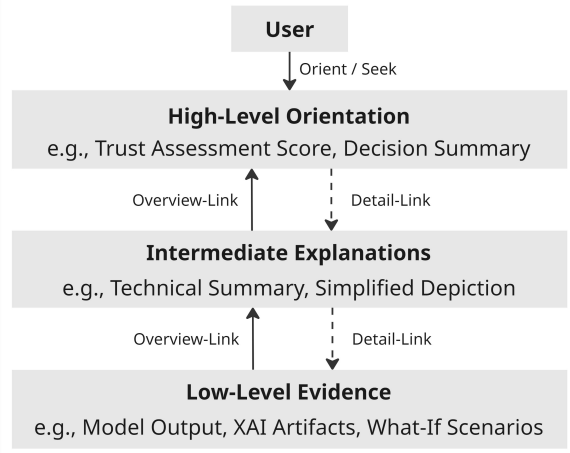


Figure 1: The XAI-Seeking Principle as linked abstraction layers.

Within the hypertext community, XAI appears as an emerging, strongly human-centered line of work rather than as a purely model-centric one. Barria-Pineda et al. [3] explicitly study *explainable recommendations*, showing how textual explanations can support users' interaction with adaptive systems, while Burel and Alani [4] make misinformation monitoring more interpretable through automatically generated, human-readable reports about the co-spread of misinformation and fact-checks.

This perspective connects naturally to recent work on generative AI: Ayoobi et al. [2] do not treat LLM-generated fake news only as a detection task, but identify simple linguistic cues intended to help users themselves become more skeptical of AI-generated news. Rahdari and Brusilovsky [10] argue that conversational navigation challenges core hypertext properties such as visible links and inspectable provenance, underscoring the need for interfaces that preserve orientation, traceability, and user agency. This work motivates treating LLM-mediated explainability as a hypertextual problem where summaries, evidence, provenance, and interaction paths must remain linked so that users can move between rapid orientation and deeper verification.

3 XAI-Seeking Principle

LLM-mediated explainability creates a representational problem where technical evidence, natural-language summaries, and higher-level assessments are produced at different levels of abstraction, but their relations are often hidden. The XSP addresses this problem by treating explainability as a linked representation space rather than as a single explanatory output.

3.1 Explainability as Linked Abstraction Layers

The XSP defines explainability as a layered process in which multiple representations of the same underlying evidence are connected across abstraction levels. Figure 1 illustrates this structure and shows how low-level XAI artifacts are progressively synthesized into higher-level summaries through overview-links, while detail-links preserve access from summaries and assessments back to the

evidence they summarize. It builds on Shneiderman's information-seeking mantra—"overview first, zoom and filter, details on demand" [15] — but adapts this idea to LLM-mediated XAI. Prior work demonstrates the value of coordinating multiple views and contextual information for decision support [14]. XSP extends this perspective by formalizing explainability as linked, LLM-mediated transformations whose coherence can be inspected and validated. Users first encounter an overview, can then inspect intermediate summaries, and can finally access detailed model evidence when needed. In contrast to a single generated explanation, XSP treats these representations as parts of a navigable structure.

At lower levels, the system exposes model outputs, feature attributions, performance indicators, or other XAI artifacts. Higher levels synthesize these artifacts into technical summaries, layperson summaries, trust assessments, or decision-oriented rationales; the number and type of layers depend on the domain and task. The central mechanism is the explicit connection between layers: overview-links compress evidence upward, while detail-links preserve verifiable paths back down. A high-level assessment should therefore remain connected to the intermediate summaries and low-level evidence from which it was derived.

In this structure, higher-level representations function as entry points for action or inspection. Users may use them for initial orientation, but the system preserves the option to seek additional evidence before relying on them. Explainability thus becomes an on-demand, goal-directed process, enabling users to move upward for abstraction and downward for justification. The following section derives design characteristics for constructing such XAI-seeking systems and for technically probing whether transformations between abstraction layers remain useful, traceable, and reliable.

3.2 Design Characteristics for XAI-Seeking Systems

Linked abstraction layers make explanations navigable, but they also introduce potential failure points. Each transformation from low-level evidence to higher-level summaries can distort, overcompress, or delay information. XAI-seeking systems therefore require design characteristics that make these transformations inspectable and adaptable. We derive four such characteristics from the XSP structure: monitoring and validation, feature–problem relation, performance–speed trade-off, and scalable improvement. These characteristics are not fixed implementation requirements and rather define what should be made explicit when constructing and technically evaluating an XAI-seeking system (see table 1).

Monitoring and validation are central because XSP relies on transformations between representations rather than on a single explanation output. Validation should assess whether individual outputs are understandable and whether the relations between outputs remain coherent. In this paper, we focus on two validation targets: truthfulness/coherence across abstraction levels and clarity within generated representations. Because these targets are partly semantic, they cannot be fully captured through deterministic checks alone. LLM-based judging (cf. [21]) can provide scalable diagnostic signals for such cases, but must be interpreted cautiously and not as independent ground truth.

Feature–problem relation concerns the degree to which the system's internal features are meaningful for the user's decision problem. In XSP, features are not only model inputs but they also serve as navigational cues. Strong feature–problem relations improve information scent because users can more easily anticipate why a feature matters and whether deeper inspection is useful. Weak feature–problem relations, by contrast, may reduce orientation and make LLM-mediated summaries harder to verify.

Performance–speed trade-off is relevant because XSP assumes that users can move between abstraction depths. This movement depends on system responsiveness. Larger models, reasoning modes, and multi-step synthesis may improve explanation quality, but they can also increase latency. XAI-seeking systems should therefore align computational effort with the expected value of additional explanation depth, especially in time-constrained decision settings.

Scalable improvement means that validation signals should not only evaluate explanations after generation, but also support targeted refinement of the explanation pipeline. When recurring divergence patterns are detected, designers can adapt prompts, feature representations, model configurations, or linking strategies. In this paper, we illustrate this through prompt adaptation based on divergence analysis, showing how validation can guide refinement without assuming fully automated self-improvement.

Together, these characteristics operationalize XSP as a design principle that can be technically probed. The following section instantiates them in a misinformation-support prototype and examines how model configuration, feature grounding, and validation-driven adaptation affect the quality and responsiveness of LLM-mediated explanation layers.

4 Misinformation Proof of Concept and Technical Probes

To examine how the XAI-Seeking Principle can be operationalized, we instantiate it in a misinformation-support proof of concept. The domain is suitable for this purpose because users may encounter frequent, low-stakes but time-sensitive judgments about short social-media claims. The prototype is not intended as a state-of-the-art misinformation detector but provides a controlled setting for studying how model evidence, XAI artifacts, and LLM-mediated summaries can be linked across abstraction levels. The system uses political statements from LIAR-2 [19] and a curated PolitiFact [1] subset, with labels simplified into three categories: *True*, *Neither*, and *False*.

4.1 Prototype and Evidence Pipeline

From the user perspective, the prototype follows the XSP interaction flow in Figure 2. A social media post is first presented with a high-level flag and trust assessment in the advisor view. Users can then inspect intermediate rationales and follow detail-links to relevant XAI modules, such as SHAP or feature distributions, where generated summaries accompany lower-level artifacts. The interface therefore implements the movement from overview to progressive exploration and detailed inspection described in Section 3.1.

[1] https://www.politifact.com

Table 1: Design characteristics for XAI-seeking systems and their corresponding technical probes.

Characteristic	Role in XSP	Technical probe in this paper
Monitoring & validation	Checks whether transformations between abstraction layers preserve coherence, clarity, and traceability.	Truthfulness/coherence checks, clarity ratings, label–assessment consistency, and divergence analysis.
Feature–problem relation	Ensures that features provide meaningful information scent and can be interpreted in relation to the task.	Comparison of semantically grounded features and more abstract linguistic features.
Performance–speed trade-off	Balances deeper synthesis and reasoning against latency in time-constrained interaction.	Comparison of LLM scale, reasoning mode, and response latency.
Scalable improvement	Uses validation signals to identify recurring failure modes and refine the transformation pipeline.	Prompt adaptation based on categorized divergence cases.

Figure 3 abstracts this interaction as a technical XSP pipeline with three functional components. First, a flagging component transforms each post into semantically grounded features for classification. Second, an XAI component generates model-specific evidence, including local feature attributions, partial-dependence information, global feature importance, feature distributions, and performance indicators. Third, an LLM-mediated explanation component transforms these artifacts into higher-level representations.

The classifier operates on ten semantically grounded features extracted from the post text: polarization, source quality, emotionality, sentiment, us-versus-them language, exaggerated uncertainty, dehumanization, black-and-white framing, victim-villain framing, and reading difficulty. As interpretable rhetorical or credibility-related cues, these features serve both as model inputs and as navigational cues in the explanation interface. In the XSP structure, they shape the information scent available to users and the LLM-mediated explanation pipeline.

The resulting feature vectors are used to train a random-forest classifier, chosen for its compatibility with efficient XAI methods such as TreeSHAP [7] and its suitability as an interpretable evidence generator. Predictive optimization is not the prototype's primary goal; rather, the classifier produces stable model outputs and XAI artifacts for studying evidence transformation across abstraction layers.

The LLM-mediated explanation component operationalizes XSP as a two-stage transformation pipeline. At abstraction level 1, low-level XAI artifacts are translated into technical and layperson summaries. At abstraction level 2, these summaries are synthesized into a trust assessment (the high-level assessment score), a technical rationale, and a short rationale. Thus, the prototype connects low-level evidence to higher-level assessments through overview-links while preserving detail-links to the XAI artifacts that ground the assessment.

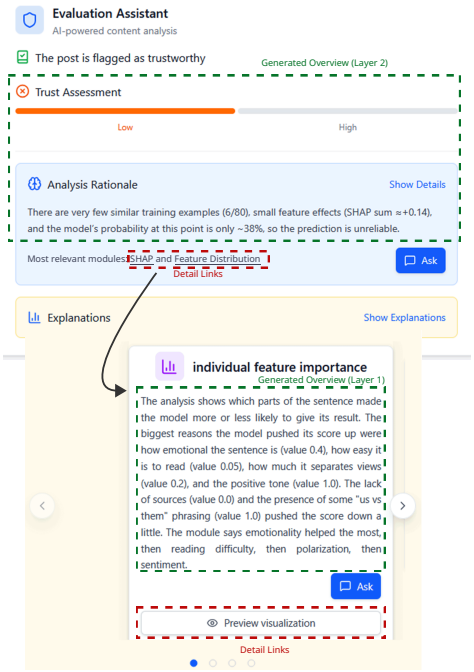


Figure 2: User-facing XSP interaction in the misinformation advisor.

4.2 Technical Evaluation Design

The evaluation technically probes the design characteristics introduced in Section 3.2. It does not evaluate user decision quality directly but rather examines whether the LLM-mediated transformation pipeline preserves clarity, coherence, and traceability across abstraction levels. The main analysis uses 102 balanced samples with five runs per configuration to account for variance in LLM outputs. We focus on the Qwen3 model family [20] to compare model scale and reasoning under a consistent evaluation setup (see table 2). Additional closed-model runs were conducted as exploratory checks but are not central to the short-paper analysis because they used a smaller sample.

We evaluate three types of signals (cf. Figure 3). First, truthfulness and coherence checks assess whether summaries and rationales remain faithful to the lower-level evidence they summarize. Second,

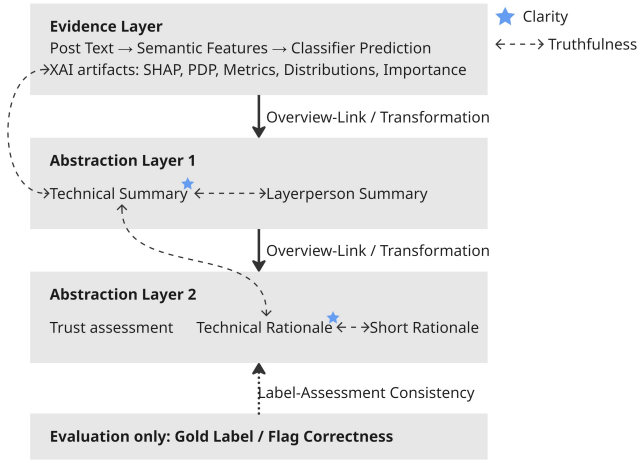


Figure 3: Instantiation of XSP in the misinformation proof of concept. Solid arrows indicate LLM-mediated transformations; dashed arrows indicate validation probes for truthfulness/coherence, clarity, and label–assessment consistency.

Table 2: Model configurations used for the model-scale and reasoning probe.

Condition	Model	Size	Reasoning
Small	Qwen3 4B	4B	No
Medium	Qwen3 8B	8B	No
Large	Qwen3 30B Instruct	30B	No
Large+Reasoning	Qwen3 30B Think	30B	Yes

clarity ratings assess whether generated representations are understandable. Third, label–assessment consistency checks whether the trust assessment aligns with the model output. Because truthfulness and clarity are semantic properties, LLM-based judging is used as a scalable diagnostic signal but not considered as independent ground truth.

The technical probes address three questions. First, how do model scale and reasoning affect different abstraction layers? Second, how does the relation between features and the decision problem influence explanation quality and assessment stability? Third, can validation signals reveal recurring failure modes and guide targeted refinement of the transformation pipeline?

4.3 Technical Probes and Findings

Table 3 summarizes the three technical probes and their implications for the XSP. The following paragraphs describe the findings in more detail.

Model scale and reasoning. The model-scale probe shows that increased LLM capacity does not improve all abstraction layers uniformly. Larger models improved explanation-oriented metrics, especially technical XAI clarity and XAI description truthfulness, but high-level assessment metrics were less stable and often non-monotonic. Reasoning reduced variance in some cases, but introduced substantial latency. This aligns with the performance–speed trade-off characteristic. XAI-seeking systems should not simply use the strongest model everywhere, but should align model capacity with the abstraction layer and with the expected value of deeper synthesis.

This finding is relevant for the design of linked abstraction layers. In this specific setup and model choices, low-level transformations, such as summarizing XAI artifacts, appeared to benefit from stronger models because they required faithful synthesis of technical evidence. Higher-level assessments, by contrast, required additional judgment about whether the model output should be trusted. These assessments may become less stable when larger or reasoning-enabled models introduce more nuanced interpretations that are not strictly grounded in the classifier evidence. Thus, different layers of an XAI-seeking pipeline may require different model configurations.

Feature–problem relation. The feature-relation probe compares the semantically grounded feature set with a more abstract linguistic feature set based on stylistic and grammatical properties, such as character count, word count, part-of-speech counts, and word density. The results were mixed. Semantic features improved explanation-oriented metrics such as truthfulness and clarity, while linguistic features improved label–assessment consistency and assessment stability.

This suggests that strong feature–problem relations improve the interpretability of LLM-mediated explanations but also introduce a risk. When features such as polarization, emotionality, or source quality are semantically meaningful, the LLM can explain them more naturally and relate them to the misinformation task. At the same time, the LLM may start to reason from its own interpretation of these feature values rather than from the XAI artifacts alone. Feature design therefore affects both information scent and the risk of feature interpretation bias. In XSP terms, features are not neutral inputs and shape the navigational cues that connect model evidence, summaries, and assessments.

Validation-driven adaptation. The validation-driven adaptation probe illustrates how monitoring can support targeted refinement. We analyzed divergence cases in which the assistant's trust assessment conflicted with the classifier evidence and categorized recurring causes. The most prominent failure mode was feature interpretation bias where the LLM sometimes treated strong semantic feature values, such as high polarization or negative emotionality, as direct evidence against the post rather than grounding the assessment in the XAI artifacts.

To address this failure mode, the abstraction-level 2 prompt was refined to reduce direct reliance on feature values and to prioritize evidence from the XAI artifacts. After adaptation, short assessment truthfulness improved by 17%, assessment clarity improved by 9%, interpretation-bias divergences declined by 40%, and total divergences declined by 12%. This result does not show fully automated self-improvement, but it demonstrates how validation signals can make failure patterns inspectable and guide targeted refinement of the explanation pipeline.

Taken together, the probes show that XSP is also a technical design problem. The quality of an XAI-seeking system depends on how model capacity, feature representation, latency, and validation are coordinated across abstraction layers. In the misinformation

Table 3: Summary of technical probes and their implications for XSP.

Probe	XSP characteristic	Main observation	Implication
Model scale and reasoning	Performance–speed trade-off	Larger models mainly improved explanation-oriented metrics at lower abstraction levels, while trust assessment was less stable. Reasoning increased latency substantially.	Different abstraction layers may require different model capacities; costly reasoning should not be applied uniformly.
Feature relation	Feature–problem relation	Semantically grounded features improved explanation-oriented metrics, while linguistic features improved label consistency and assessment stability.	Feature choice may necessitate trade-offs between explanatory accessibility and assessment alignment.
Divergence adaptation	Monitoring, validation, and scalable improvement	Divergence analysis revealed recurring interpretation bias. Prompt adaptation improved abstraction-level 2 metrics and reduced divergent cases.	Validation signals can guide targeted refinement of LLM-mediated transformations.

proof of concept, these probes provide evidence that linked explanation layers can be technically inspected and refined.

5 Discussion and Limitations

The technical probes in this paper support XSP primarily as a design principle for structuring and inspecting LLM-mediated explanation pipelines. They complement prior user-facing work on the same misinformation AI Advisor, which found indicative evidence that LLM mediation can reduce time-to-decision and reshape interaction behavior without clearly increasing blind reliance on the system [6]. In contrast, this paper examines the technical characteristics underlying this instantiation, i.e., how model configuration, feature representation, latency, and validation affect linked abstraction layers. In this sense, the main contribution is a conceptual and technical account of how an XAI-seeking pipeline can be constructed, inspected, and refined.

The probes point to three design implications. First, model capacity should be selected per abstraction layer rather than treated as a uniform system-level property. Stronger or reasoning-enabled models can improve low-level synthesis, but are not necessarily better for high-level assessments and introduce latency. Second, semantically grounded features improve information scent and explanation quality, but also require validation because they can invite feature interpretation bias. Third, validation should inspect relations between abstraction layers rather than only isolated outputs; divergence analysis can expose recurring failure modes and guide prompt or feature refinements. Thus, XSP is a technical design problem in which model configuration, feature representation, linking, and validation must be coordinated.

The conceptual contribution of XSP is therefore the combination of cognitive compression and cognitive verification. Overview-links allow technical evidence to be transformed into higher-level summaries that support orientation. Detail-links preserve access to the lower-level evidence required for inspection and contestation. This structure is especially relevant for LLM-mediated interaction, where fluent summaries can otherwise hide provenance and collapse multiple evidence sources into a single statement. XSP addresses this by treating explainability as a navigable representation space rather than as a single generated explanation.

Several limitations remain. First, the results provide technical evidence about the explanation pipeline, not comprehensive validation of user-facing decision quality. Second, LLM-based judging enables scalable semantic diagnostics but introduces circularity because LLMs both generate and assess explanations. Its scores should therefore be treated as diagnostic signals rather than independent ground truth. Third, the interpretable random-forest classifier and semantically grounded features facilitate inspection but limit generalization to more complex models and domains with less user-aligned features.

XSP is structurally domain-general: many decision-support settings require movement between evidence, summaries, and higher-level assessments. Yet its layers, features, validation targets, and latency thresholds require domain-specific adaptation. Future work should use larger, more diverse user studies to examine effects on orientation, verification, calibrated reliance, and decision quality.

6 Conclusion

This paper introduced the XAI-Seeking Principle as a hypertextual design principle for LLM-mediated explainability. XSP structures explanation as linked abstraction layers, where overview-links transform low-level XAI evidence into higher-level summaries and detail-links preserve access to inspectable evidence. Through a misinformation-support proof of concept, we examined how model scale, feature grounding, latency, and validation affect such a pipeline. The findings show that abstraction layers differ in their model requirements, that semantically meaningful features improve explanation quality while introducing risks of interpretation bias, and that validation signals can guide targeted refinement. By linking actionable summaries to inspectable evidence, XSP supports both rapid orientation and deeper verification in human–AI decision-making. XSP therefore contributes a design abstraction for constructing explainability as a navigable representation space, while user-facing effects and generalization to more complex domains remain important directions for future work.

References

[1] Chirag Agarwal, Sree Harsha Tanneru, and Himabindu Lakkaraju. 2024. Faithfulness vs. Plausibility: On the (Un)Reliability of Explanations from Large Language Models. arXiv:2402.04614 https://arxiv.org/abs/2402.04614

[2] Navid Ayoobi, Sadat Shahriar, and Arjun Mukherjee. 2024. Seeing Through AI's Lens: Enhancing Human Skepticism Towards LLM-Generated Fake News. In *Proceedings of the 35th ACM Conference on Hypertext and Social Media* (Poznan, Poland) *(HT '24)*. Association for Computing Machinery, New York, NY, USA, 1–11. doi:10.1145/3648188.3675136

[3] Jordan Barria-Pineda, Kamil Akhuseyinoglu, and Peter Brusilovsky. 2023. Adaptive Navigational Support and Explainable Recommendations in a Personalized Programming Practice System. In *Proceedings of the 34th ACM Conference on Hypertext and Social Media* (Rome, Italy) *(HT '23)*. Association for Computing Machinery, New York, NY, USA, Article 29, 9 pages. doi:10.1145/3603163.3609054

[4] Grégoire Burel and Harith Alani. 2023. The Fact-Checking Observatory: Reporting the Co-Spread of Misinformation and Fact-checks on Social Media. In *Proceedings of the 34th ACM Conference on Hypertext and Social Media* (Rome, Italy) *(HT '23)*. Association for Computing Machinery, New York, NY, USA, Article 4, 3 pages. doi:10.1145/3603163.3609042

[5] Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-Verification Reduces Hallucination in Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2024 (ACL 2024)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 3563–3578. doi:10.18653/v1/2024.findings-acl.212 https://aclanthology.org/2024.findings-acl.212/

[6] Valentin Grimm, Eelco Herder, Jessica Rubart, and Carsten Röcker. 2026. LLM-Mediated XAI Explanations: An AI Advisor for Fast and Calibrated Judgments on Potential Misinformation. In *Companion Publication of the 2026 18th ACM Web Science Conference (WebSci Companion '26)*. Association for Computing Machinery, New York, NY, USA, 110–116. doi:10.1145/3795513.3810452

[7] Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee. 2019. Consistent Individualized Feature Attribution for Tree Ensembles. arXiv:1802.03888 https://arxiv.org/abs/1802.03888

[8] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc., San Diego, CA, USA. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[9] Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 9004–9017. doi:10.18653/v1/2023.emnlp-main.557 https://aclanthology.org/2023.emnlp-main.557/

[10] Behnam Rahdari and Peter Brusilovsky. 2025. From Links to Dialogue; Hypertext Challenges and Opportunities in Conversational Navigation. In *Adjunct Proceedings of the 36th ACM Conference on Hypertext and Social Media*. ACM, Chicago IL USA, 25–29. doi:10.1145/3720533.3750064

[11] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 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* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. doi:10.1145/2939672.2939778

[12] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: high-precision model-agnostic explanations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence* (New Orleans, Louisiana, USA) *(AAAI'18/IAAI'18/EAAI'18)*. AAAI Press, Washington D.C., USA, Article 187, 9 pages.

[13] Yao Rong, Tobias Leemann, Thai-Trang Nguyen, Lisa Fiedler, Peizhu Qian, Vaibhav Unhelkar, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. 2023. Towards human-centered explainable ai: A survey of user studies for model explanations. *IEEE transactions on pattern analysis and machine intelligence* 46, 4 (2023), 2104–2122.

[14] Jessica Rubart, Benjamin Lietzau, Patrick Söehlke, Bastian Alex, Stephan Becker, and Tim Wienböeker. 2017. Semantic Navigation and Discussion in a Digital Boardroom. In *2017 IEEE 11th International Conference on Semantic Computing (ICSC)*. IEEE Press, San Diego, CA, USA, 290–296. doi:10.1109/ICSC.2017.39

[15] Ben Shneiderman. 2003. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *The Craft of Information Visualization*, BENJAMIN B. BEDERSON and BEN SHNEIDERMAN (Eds.). Morgan Kaufmann, San Francisco, 364–371. doi:10.1016/B978-155860915-0/50046-9

[16] Ponhvoan Srey, Xiaobao Wu, and Anh Tuan Luu. 2025. Unsupervised Hallucination Detection by Inspecting Reasoning Processes. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 22117–22129. doi:10.18653/v1/2025.emnlp-main.1124 https://aclanthology.org/2025.emnlp-main.1124/

[17] Vinitra Swamy, Davide Romano, Bhargav Srinivasa Desikan, Oana-Maria Camburu, and Tanja Käser. 2025. iLLuMinaTE: an LLM-XAI framework leveraging social science explanation theories towards actionable student performance feedback. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'25/IAAI'25/EAAI'25)*. AAAI Press, Washington D.C., USA, Article 3167, 9 pages. doi:10.1609/aaai.v39i27.35065

[18] Sebe Vanbrabant, Gilles Eerlings, Gustavo Alberto Rovelo Ruiz, and Davy Vanacken. 2025. ECHO: Enhancing Conversational Explainable AI through Tool-Augmented Language Models. *Proceedings of the ACM on Human-Computer Interaction* 9, 4 (june 2025), 1–33. doi:10.1145/3734191

[19] Cheng Xu and M-Tahar Kechadi. 2024. An Enhanced Fake News Detection System With Fuzzy Deep Learning. *IEEE Access* 12 (2024), 88006–88021. doi:10.1109/ACCESS.2024.3418340

[20] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 Technical Report. arXiv:2505.09388 https://arxiv.org/abs/2505.09388

[21] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., Red Hook, NY, USA, 46595–46623. https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf

[22] Jianlong Zhou, Syed Z. Arshad, Simon Luo, and Fang Chen. 2017. Effects of Uncertainty and Cognitive Load on User Trust in Predictive Decision Making. In *16th IFIP TC 13 International Conference on Human-Computer Interaction — INTERACT 2017 - Volume 10516*. Springer-Verlag, Berlin, Heidelberg, 23–39. doi:10.1007/978-3-319-68059-0_2