LLM-Mediated XAI Explanations: An AI Advisor for Fast and Calibrated Judgments on Potential Misinformation

Valentin Grimm
Computer Science & Automation
TH OWL University of Applied Sciences and Arts
Höxter, Germany
valentin.grimm@th-owl.de

Eelco Herder
Utrecht University
Utrecht, Netherlands
e.herder@uu.nl

Jessica Rubart
Computer Science & Automation
TH OWL University of Applied Sciences and Arts
Höxter, Germany
jessica.rubart@th-owl.de

Carsten Röcker
Computer Science & Automation
TH OWL University of Applied Sciences and Arts
Lemgo, Germany
carsten.roecker@th-owl.de

Abstract

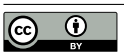
This paper introduces an LLM-mediated AI Advisor that contextualizes and synthesizes heterogeneous explainable AI (XAI) outputs to support fast and calibrated misinformation judgments in time-sensitive social media settings. We define LLM-mediated XAI as a process in which a large language model aggregates, prioritizes, and translates heterogeneous XAI outputs into a context-sensitive explanation tailored to the user's decision situation. Semantic features, XAI modules and LLM-based summarization and synthesis enable the generation of explanations that are adapted in three ways: compressed for time-efficient decisions, translated into non-technical language, and progressively expandable for deeper inspection. Through a mixed-methods user study, including a quantitative study and a qualitative study, we analyze how users interpret, challenge and strategically rely on LLM-mediated explanations during real-world misinformation assessment tasks. The findings indicate that the approach reduces time-to-decision and supports critical inspection without inducing over-reliance. Progressive disclosure and different techniques to present information favored different user needs while conversational functionality was rarely used due to unclear benefits and fear of confusion.

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



WebSci Companion '26, Braunschweig, Germany

ACM ISBN 979-8-4007-2492-3/26/05
https://doi.org/10.1145/3795513.3810452

ACM Reference Format:
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 *18th ACM Web Science Conference (WebSci Companion '26), May 26–29, 2026, Braunschweig, Germany.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3795513.3810452

1 Introduction

Social media platforms have become a primary channel for information consumption, but they are also a major vector for the rapid spread of misinformation.

While automated detection systems have advanced considerably (e.g., X, formerly known as Twitter)[1], fully automated moderation remains problematic: machine learning and decision-support models for social media moderation are inherently imperfect and context-dependent, and they exhibit biases and failure modes that are often hard to foresee or fully characterize before deployment [5]. As a result, the users themselves stay highly responsible for which content to trust and which not, despite algorithmic decision aids.

At the same time, the scale and speed of online information flows place substantial demands on human judgment. Users have to make credibility assessments in short amounts of time, with limited attention and varying levels of domain expertise before they move on to other content. In such settings, simply exposing raw model outputs or complex explanations risks overwhelming users rather than empowering them.

This tension highlights a central challenge for human-centered AI: how to support informed, efficient and calibrated decision-making in fast-paced decision contexts without substantially increasing cognitive burden or encouraging over-reliance. We approach this challenge by framing decision support as a contextualized, situated process mediated by an AI Advisor.

In this work, LLM mediation is conceptualized as a transformation layer between XAI components and the user. Instead of exposing individual explanations directly, the LLM aggregates multiple XAI signals, prioritizes them based on relevance to the decision task and generates a unified natural language assessment. This

[1] https://blog.x.com/en_us/topics/company/2022/introducing-our-crisis-misinformation-policy

way, users are enabled to engage with additional detail only when needed, in the spirit of Shneiderman's information seeking mantra [22].

Adaptation in this context does not rely on explicit user modeling, but emerges from the alignment of explanations with the decision situation, characterized by time pressure, limited attention and non-expert users.

We make three contributions toward understanding how LLMs can mediate explainable decision support in fast-paced, user-driven settings.

First, we introduce the AI Advisor, a system design that integrates a classifier with multiple XAI components and uses an LLM as a mediation layer that aggregates outputs from multiple XAI modules, ranks them by decision relevance, and synthesizes them into a single, user-facing assessment with optional drill-down explanations. Second, we operationalize this mediation in a self-moderated misinformation detection task, deliberately targeting a low-stakes but high-frequency scenario in which explanations must be concise and users retain full judgment authority. Third, through a mixed-methods evaluation, we empirically examine how the AI Advisor affects time-to-decision, reliance and user engagement with explanatory content. Together, these contributions provide guidance on when and how LLM mediation can support efficient human judgment in a variety of individual contexts.

2 Related Work

Explainable AI (XAI) has traditionally aimed to improve the transparency and interpretability of machine learning models by exposing internal reasoning, counterfactual reasoning [24] or feature importance [16]. Foundational surveys, such as Nunes and Jannach [18], conceptualize explanations as components of decision-support systems, identifying goals such as transparency, justification, trust support and user acceptance.

Studies across several domains demonstrate that users often struggle to operationalise explanations, particularly when they are verbose, technical or insufficiently contextualized [21].

Verbose, technical or poorly contextualized explanations can overwhelm users, decrease decision-quality, affect over-reliance and slow down decision-making [11]. Prior work suggests that the effectiveness of explanations depends not only on what information is shown and its accuracy but also on how and when [1]. The level of detail seems to be a central property to handle the trade-off between high effort and high assurance [15].

Cognitive forcing functions like adding wait-periods [2] or progressive disclosure [25] have been used to address these issues. While these measures help reduce over-reliance, they have been investigated in rather isolated, non-contextualized and non-personalized settings.

LLMs enable flexible summarization, translation, and synthesis of model outputs and explanations, lowering the technical barrier for non-expert users. Systems such as TalkToModel [23] and XAgent [27] explored conversational interfaces for interacting with machine learning explanations.

Several recent studies caution that conversational interfaces may unintentionally foster over-trust and over-reliance when uncertainty, limitations, or conflicting evidence are not made explicit. He et al. [10] show that conversational explanations can amplify persuasive effects if not carefully designed. Approaches such as "LLMs as Devil's Advocates" [26] and tool-augmented systems like ECHO [28] attempt to mitigate these risks by encouraging critical dialogue and counterfactual reasoning.

Automated misinformation detection has been widely studied using supervised and semi-supervised models trained on textual, linguistic, and graph-based features [3, 14, 30]. While such systems can achieve competitive accuracy, prior work consistently reports limitations in real-world deployment, including sensitivity to domain shifts, evolving narratives, and context-dependent interpretation.

As a result, recent research increasingly emphasizes human-in-the-loop approaches that assist rather than replace human judgment [6, 7, 17]. In these settings, the goal is not only to flag suspicious content but also to support users in understanding why a post is flagged and how confident the system is in its assessment. This perspective aligns with broader human-centered AI research highlighting the importance of meaningful human oversight even in comparatively low-stakes domains such as social media moderation [20, 29].

Some recent work explores LLM-based approaches for combating misinformation by generating corrective or personalized responses. Proma et al. [19], for example, propose a fact-grounded, personalized LLM pipeline that tailors persuasive responses to user characteristics. While effective in generating diverse and persuasive outputs, this approach primarily targets belief change rather than supporting independent inspection or calibrated decision-making.

Prior research reveals a gap between the theoretical promise of explainable, conversational, and mediated AI systems and their actual use in practice. Our work addresses this gap by studying LLM mediation not as a conversational endpoint, but as a decision support layer that emphasizes on adaptive, human-centered AI systems that respect user control while supporting efficient and effective decision-making.

3 A System for Self-Moderated Misinformation Detection

Political misinformation is one of the most persistent and socially consequential forms of online deception. Unlike professional fact-checking, which is slow and resource-intensive, everyday social media use requires rapid, repeated credibility judgments made by non-expert users under limited attention and time pressure. In our use case, we explicitly enable self-moderated misinformation detection, where users remain responsible for the final judgment but are supported by an AI-based decision aid.

3.1 Scope and Data Basis

The system is grounded in political fact-checking data derived from the LIAR2 dataset [31] and an additional curated subset of recent (2024 & 2025) PolitiFact articles[2]. Statements include contextual metadata such as speaker, source, and topic. For the purpose of user-facing moderation, the original six truthfulness labels are mapped to three categories (True, Neither, False), with the study focusing on True and False cases. This simplification reflects the goal of

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

supporting fast credibility judgments by non-expert users rather than detailed journalistic fact-checking.

The intended user is a non-expert social media user who encounters potentially misleading political content in their personal feed. The task is not to verify claims exhaustively, but to decide whether a post should be considered trustworthy enough to accept or share, or sufficiently suspicious to ignore, question or further inspect and report.

3.2 Application Instantiation

From the user's perspective, the interaction with the System and the AI Advisor follows a defined set of properties, designed to balance speed, transparency and optional depth. The most relevant components in the interface are depicted and explained in Figure 1.

Each post in the timeline is accompanied by a lightweight machine-learning-based flag that indicates one of the three veracity categories (left). The flag serves as an attention cue rather than a decision mandate, signaling that a post may warrant closer inspection. For instance, when a post seems trustworthy but is flagged as misinformation or vice versa.

When a user selects a flagged post, the system presents a concise assessment of that flag (how trustworthy is it?) generated by an LLM via the advisor view. This assessment integrates multiple XAI signals through LLM-mediated synthesis, where signals are filtered and prioritized based on their relevance for a quick credibility judgment. The resulting explanation is adapted to the decision context by emphasizing high-level cues (e.g., conflicting evidence, uncertainty) while deferring detailed technical information to optional layers.

If a user wishes to better understand why the system reached its assessment, they can progressively expand the interface. Additional layers reveal structured textual descriptions of influential features, visual explanations and conversational interaction. The AI Advisor implements LLM-mediated explanation as a multi-stage transformation pipeline:

(1) Feature abstraction: input posts are transformed into semantic features relevant to misinformation detection.
(2) Model inference: a classifier produces a prediction and associated XAI signals.
(3) Explanation translation: individual XAI outputs are translated into natural language statements.
(4) LLM mediation: the LLM aggregates and prioritizes these statements to produce a concise assessment and rationale.
(5) Progressive disclosure: more detailed explanation layers are made available for optional inspection.

This design explicitly follows Shneiderman's information seeking mantra, "overview first, zoom and filter, details on demand" [22], to avoid overwhelming users and enable users to choose the information they need.

The flagging is based on the following approach: posts are encoded into semantic feature values (with an LLM) that are related to misinformation patterns, such as polarization or weak claims of sources (cf. [7]). Following these abstractions, a supervised machine learning classifier is used to generate a misinformation flag. Then, for the AI Advisor, statistical data from the training process including data distribution and XAI information ("modules") are calculated. Each piece of information is translated into natural language by the LLM, that summarizes and highlights most significant patterns (→ Dashboard Setting).

Finally, the individual information from the XAI modules are then passed on to an assessment step where the LLM aggregates all provided information, generates a trust rating (low or high) towards the machine learning decision and provides a description for their choice (→ Assistant Setting).

All generated explanations are adaptively transformed into multiple representations, including a concise summary for rapid decisions and extended explanations for deeper inspection, enabling users to control the level of detail according to their needs.

The prototype uses an LLM-based feature extraction pipeline combined with a random forest classifier, which serves as a functional proof-of-concept.

4 User Studies

To investigate how the AI Advisor supports decision-making in self-moderated misinformation detection, we conducted a mixed-methods evaluation consisting of (1) a controlled quantitative user study and (2) a qualitative follow-up study. The two studies are designed to be complementary: the quantitative study is meant to inform the qualitative study by surfacing general phenomena such as decision-time differences and initial subjective feedback by evaluating observable effects on efficiency, trust and cognitive load while the qualitative study aims to uncover interpretive strategies and tensions that are difficult to capture through surveys alone. We treat the quantitative results as indicative rather than conclusive.

4.1 Quantitative Study

4.1.1 Design and Procedure. The quantitative study followed a within-subject design with two groups of students with a combined $n = 14$ participants. One, with 8 students from the Master IT and 6 students from the Bachelor on applied computer sciences. All participants had a technical background, but only with no- to moderate experience with XAI applications. Each participant completed a sequence of misinformation assessment tasks under two conditions.

The study compared two conditions with the task to decide whether they considered the content of a post trustworthy or not.

(1) XAI Dashboard: Participants interacted with a baseline interface providing individual XAI explanations (e.g., feature importance visualizations), supported by LLM-generated textual descriptions.
(2) AI Advisor: Participants used the full system described in Section 3.2, where XAI outputs are additionally synthesized and contextualized by an LLM into a concise assessment and rationale.

The key difference between conditions is the presence of an LLM-mediated aggregation layer that summarizes and prioritizes explanatory information.

The study followed a fixed order in which all participants first interacted with the XAI dashboard condition, followed by the Ai Advisor condition. This design choice was made because the AI Advisor interface builds upon the components of the dashboard, making it difficult to isolate the baseline condition after exposure to the full system. However, this introduces a potential learning

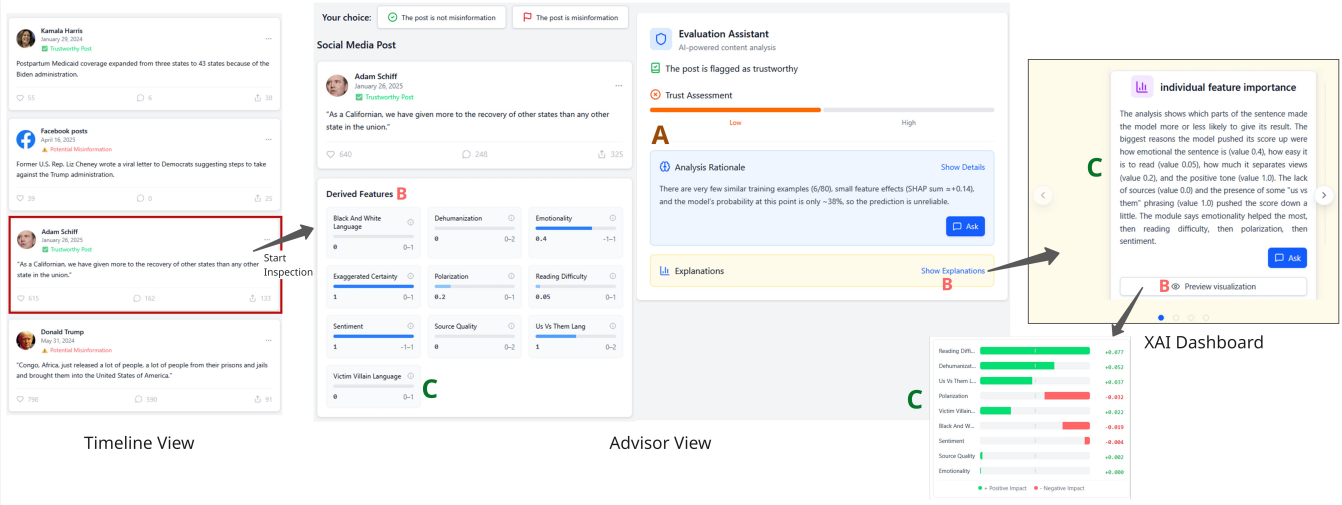


Figure 1: The AI Advisor interface follows Shneiderman's mantra: (A) Overview First, (B) Zoom & Filter, and (C) Details on Demand.

or order effect. To address this, we excluded the first trial of each condition from analysis and used only subsequent trials for comparison. We acknowledge this as a limitation when interpreting the results.

The study procedure consisted of four stages, as depicted in Figure 2. Participants first received a short introduction to the system and then completed a sequence of misinformation assessment tasks under two conditions. Then, a classical social media timeline with 30 position-randomized posts is shown to the users that carry misinformation flags (either "trustworthy" or "potential misinformation"). Users have to select at least 6 of them. Afterwards, three posts are analyzed with the XAI dashboard followed by the analysis of three posts with the dashboard combined with the AI Advisor. The participants were instructed to decide whether they considered each post trustworthy or not, using any interface component they deemed helpful. After each of the two conditions, a survey is conducted. We collected both subjective and behavioral measures:

- Time-to-decision: Measured as the time (in seconds) from post selection to final decision.
- Decision accuracy: Binary correctness of the participant's decision compared to the ground truth label.
- Agreement with AI: Whether the participant's decision aligned with the system's prediction.
- Interaction behavior: Number and type of interactions with UI components (toggling of short and long rationale (light blue box), toggling of the XAI dashboard (light yellow box), movement between XAI explanations and display of visualizations).
- Subjective measures (survey): Trust in automation [13], perceived suitability [12], and cognitive load via NASA TLX [9] were assessed using post-condition questionnaires on Likert scales.

4.1.2 Results. At the level of subjective ratings, we observed no statistically significant differences between the baseline and assistant

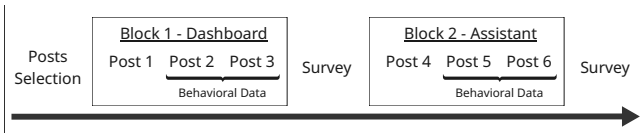


Figure 2: Process depiction of the quantitative study

conditions across the trust, suitability and cognitive load dimensions. A Mann–Whitney-U-test revealed no statistically significant differences between the two conditions across all survey items (all $p > .15$). Effect size estimates (Cohen's d and Cliff's delta) indicated small to moderate differences for some items. However, these effects were inconsistent in direction and did not reach statistical significance.

Participants generally reported moderate trust, perceived the explanations as timely and rated the task as relatively mentally demanding in both conditions.

However, behavioral data revealed a clearer picture. Time-to-decision decreased by approximately 35% in the assistant condition. This reduction occurred despite similar subjective cognitive load ratings, suggesting increased efficiency. Also, the differences are relatively consistent over different quartiles and showed low variance.

In the XAI dashboard condition, participants relied on the visual explanations, with an average of 7.5 visualization interactions per user and no observed chat interactions. When the AI Advisor was introduced in the second condition of the study, interaction patterns shifted noticeably. Visualization interactions dropped to an average of three per user, corresponding to a reduction of approximately 60%. At the same time, participants made limited use of assistant-related features, with "Show Details" accessed on average 0.25 times per user, and chat interactions occurring 0.21 time per user.

This indicates that participants substantially reduced their direct interaction with visual XAI components when an LLM-mediated

assessment was present. In both cases, the conversational option to further engage with the LLM was basically unutilized. Decision accuracy remained comparable across the conditions. In the baseline condition, participants achieved an accuracy of 54% compared to 50% in the AI Advisor condition.

With respect to reliance, agreement with the recommendations was observed in 58% of cases. The rate of correct AI predictions that were initially disagreed with by users and later corrected, was similar across conditions: 50% in the baseline and 48% in the assistant condition.

Overall, these findings provide a first indication that LLM mediation in the context of the AI Advisor primarily *reshapes* how users arrive at decisions rather than what they decide. It appears to function as a cognitive shortcut or synthesis layer, reducing the need for detailed inspection of visual explanations without replacing user judgment or inducing over-reliance.

4.2 Qualitative Study

The qualitative study aims to further analyze the effects of the AI Advisor and to get a deeper understanding of the user behavior.

4.2.1 Motivation and Design. While the quantitative study captured aggregate effects on efficiency, trust and cognitive load, it provided limited insight into how participants interpreted and combined the different explanatory components during decision-making. In particular, survey-based measures could not fully explain why reduced interaction with visual explanations did not translate into lower perceived cognitive load, nor how users reasoned about the assistant's assessment.

To address these questions, we conducted a qualitative follow-up study using semi-structured interviews based on the method by DeJonckheere and Vaughn [4] with 8 participants that differed from the participants in the quantitative study. After clarifying the purpose and scope to each participant, they were introduced to the use case and did a hands-on misinformation assessment with our prototype. They studied each component and were prompted to make a rough estimate if they believe that the post at hand is rather misinformation or not.

Afterwards, we conducted a semi-structured interview that contained a grand tour question ("Briefly describe how you approached the task and what you focused on to reach your decision."), followed by 5 core questions, follow-up questions and unplanned follow-up questions that were related to each component of the user interface. These components are the semantic features (Fig.1, bottom-left), the assessment score (upper-right), the assessment rationale (center-right), the detailed insights (bottom-right) and the ask functionality that enables further discussion with regards to the rationale or detailed insights. We wrapped up the interview with "integration" questions that were about the general strategy to work with such a system, relative value of individual components and which components increase confidence and efficiency.

4.2.2 Observations. Interviews were documented and analyzed using an iterative thematic analysis approach. The analysis focused on identifying recurring strategies, points of confusion, and patterns of reliance on system components.

The main observations are presented in table 1. Across interviews, participants consistently followed a human-first, AI-second decision strategy. They initially formed an independent judgment based on the post content before consulting the AI Advisor, which was primarily used as a second opinion rather than a directive.

Semantic features mostly functioned as fast, intuitive cues rather than evidence. Participants used them to form or confirm expectations and to decide whether deeper inspection was warranted. However, some features were occasionally perceived as ambiguous on the first glance, leading to confusion before the AI Advisor provided clarification. Most importantly, the semantic features seemed to serve as anchor points that built initial expectations, which counteracted high reliance on the AI Advisor.

The trust assessment score served a meta-cognitive role. It was rarely decisive but shaped the depth of subsequent engagement. If expectations were clearly aligned with the AI Advisors output, it led to reduced or no further inspection, while disagreement always triggered closer inspections.

Textual rationales acted as a gateway to deeper engagement. Short rationales were generally sufficient and preferred, while longer explanations were selectively consulted, mainly to handle conflicts of expectation or heightened interest. Several participants described long explanations as cognitively costly and less practical under time constraint.

Visual explanations, particularly feature attributions, supported quick orientation and confidence building, when easily interpretable. However, dense or technically framed visuals were often skipped. Overall, visuals seemed to support faster decisions, whereas text supported confidence and justification.

Conversational interaction was rarely used. Participants reported uncertainty about its benefits, concerns about cognitive load, or lack of ideas for questions. Hypothetically, some saw value in targeted functionalities (e.g. web search), but not as a primary interaction mode.

Overall, participants demonstrated calibrated trust by not relying blindly on the system and using the AI Advisor rather as a supportive second opinion. However, reliance varied substantially, with a minority reporting strong dependence. Perceptions of the weakest system component differed widely, most often pointing to detailed explanations but without clear consensus. This highlights the strength of the approach, to adapt to different user types and needs.

5 Discussion

While we only conducted a small-scale quantitative study, one of the most consistent findings is the substantial reduction in time-to-decision when the LLM assistant is available, despite minimal conversational interaction and only moderate agreement with the system's recommendations. This suggests that LLM mediation supports decision efficiency.

Importantly, this efficiency gain did not coincide with increased over-reliance. Users selectively consulted the LLM-generated summary and proceeded with their own judgment. This behavior aligns with our design goal of building a supportive rather than authoritative system and stands in contrast to concerns that LLM-generated

Analytical Dimension	Similarities	Differences
Initial decision strategy	Participants started with reading the post and semantic features before engaging with the AI Advisor	Some formed early decisions quickly (P1, P2); others deliberately delayed judgment (P3, P4)
Semantic features	Served as cues rather than evidence; some features were initially confusing; used to build expectations	-
Trust assessment score interpretation	Seen as estimate rather than strict recommendation; trigger for further inspection based on own intuition	Varying degree of influence from very strong (P7) to minor (P2, P3); some had higher trust when aligned (P7, P5), while others remained cautious despite alignment (P8)
Rationale usage	Short rationale served as gateway; long rationale in case of expectation conflict or high interest; usually confirmatory to beliefs	Long explanations either clarifying and necessary (P4, P5) or too long / cognitively costly (P7, P8)
Individual Explanations	Text before visuals; feature attribution increased perceived transparency/confidence	For some most confident building (P3), for other hard to interpret (P5, P8)
Time constraint behavior	Semantic features to build expectations; Trust assessment (+ rationale) enough if expectations are met	Some reported to look into detailed visuals only (P8)
Use of conversational interaction	No use due to fear of cognitive overload and potential confusion; unclear benefits	Some saw potential value if it offers specific functionalities like web-search (P1)
Overall trust calibration	No blind trust; AI as second opinion after human expectation	Some reported clear reliance on AI Advisor (P7)
Perceived weakest system component	Mostly detailed explanations but with hesitation	Very diverse opinions: detailed explanations (P2, P5, P1, P6), semantic features (P8, P4), rationale (P7, P3)

Table 1: Comparison of similarities and differences across qualitative interview outcomes.

natural language explanations inherently inflate trust or encourage automation bias (cf. [10]).

Contrary to expectations derived from prior work on natural language explanations, subjective cognitive load did not significantly decrease in the assistant condition. However, behavioral data and interview insights suggest that cognitive effort was redistributed rather than eliminated.

With LLM mediation, users spent less effort in navigating raw visualizations and more cognitive effort engaging with synthesized assessments. At the same time, detailed explanatory layers, especially technical descriptors and feature attributions, introduced new points of confusion for non-expert users. This indicates that LLM-mediated output must account not only for the amount of information but also for *conceptual coherence* across layers.

Conversational interaction with the assistant was rare. Users reported uncertainty about what to ask and preferred to consume explanations passively unless explicitly prompted. This suggests that in time-sensitive, low- to mid-stakes decision-support scenarios, LLMs are more effective as attention-directing components than as open-ended conversational agents (cf. [8]).

Several limitations of this work should be emphasized: the quantitative study was conducted with a relatively small sample size, which limits statistical power and generalizability. While the observed effects, particularly reduced time-to-decision, were consistent, more subtle effects related to trust calibration or cognitive load may not have been detectable. Second, the study focuses on a single, time-sensitive use case of self-moderated misinformation detection.

The role of the LLM as an attention-directing mediator may therefore not directly transfer to domains with different stakes, workflows or accountability structures. Third, the qualitative findings are exploratory and based on a small sample of participants. They should be interpreted as indicative rather than confirmatory, motivating further investigation across domains and especially longer-term use.

6 Conclusion

In this paper, we investigated how LLM-mediated explanations support user decision-making in a self-moderated misinformation detection setting. Rather than conceptualizing LLMs as conversational agents or standalone explanation generators, we examined their role as mediators that synthesize, contextualize and prioritize existing XAI information for non-expert users who operate under time and cognitive constraints.

Our findings demonstrate the potential of LLM-mediation substantially reducing time-to-decision and reshaping interaction behavior, without increasing blind reliance on the system. More detailed explanations did not uniformly increase trust and, in some cases, productively induced skepticism. Challenging the current focus on conversational systems, users rarely engaged in the extended dialogue with the assistant.

By grounding our study in a realistic self-moderation use case, this work contributes empirical evidence to ongoing discussions on how explanations should be presented in human-centered AI systems for decision-making in highly individualized contexts. Our results suggest that effective decision support does not require exhaustive explanation or continuous interaction, but rather carefully designed mediation that respects users' time, intuition and agency.

Future work will explore how LLM-mediated explanation strategies generalize across domains, user expertise levels and task pressure levels. Further evaluation could provide additional insights. With this work, we aim to encourage the study of LLMs not only as generators of content but as infrastructural components that shape how humans make sense of AI-assisted decisions.

For reproducibility, the code for frontend and backend is available via https://github.com/GrimmV/misinformation-ui-eval and https://github.com/GrimmV/misinformation-study-backend

References

[1] Lamia Alam and Shane Mueller. 2021. Examining the effect of explanation on satisfaction and trust in AI diagnostic systems. *BMC Medical Informatics and Decision Making* 21, 1 (2021), 178. doi:10.1186/s12911-021-01542-6

[2] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 1–21. doi:10.1145/3449287

[3] Anshika Choudhary and Anuja Arora. 2021. Linguistic feature based learning model for fake news detection and classification. *Expert Systems with Applications* 169 (May 2021), 114171. doi:10.1016/j.eswa.2020.114171

[4] Melissa DeJonckheere and Lisa M Vaughn. 2019. Semistructured interviewing in primary care research: a balance of relationship and rigour. *Family Medicine and Community Health* 7, 2 (March 2019), e000057. doi:10.1136/fmch-2018-000057

[5] Andrés Domínguez Hernández, Richard Owen, Dan Saattrup Nielsen, and Ryan McConville. 2023. Ethical, political and epistemic implications of machine learning (mis)information classification: insights from an interdisciplinary collaboration between social and data scientists. *Journal of Responsible Innovation* 10, 1 (Jan. 2023), 1–25. doi:10.1080/23299460.2023.2222514

[6] Jingcheng Du, Sharice Preston, Hanxiao Sun, Ross Shegog, Rachel Cunningham, Julie Boom, Lara Savas, Muhammad Amith, and Cui Tao. 2021. Using machine learning–based approaches for the detection and classification of human papillomavirus vaccine misinformation: Infodemiology study of reddit discussions. *Journal of Medical Internet Research* 23, 8 (2021), e26478.

[7] Mirko Franco, Valentin Grimm, and Eelco Herder. 2025. Preventing Accidental Sharing of Misinformation Using Large Language Models. In *Proceedings of the 2025 International Conference on Information Technology for Social Good (GoodIT '25)*. Association for Computing Machinery, New York, NY, USA, 244–252. doi:10.1145/3748699.3749798

[8] Valentin Grimm, Jessica Rubart, and Patrick Söhlke. 2024. Conversational Data Stories. In *Proceedings of the 7th Workshop on Human Factors in Hypertext* (Poznan, Poland) *(HUMAN '24)*. Association for Computing Machinery, New York, NY, USA, Article 3, 6 pages. doi:10.1145/3679058.3688631

[9] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, Amsterdam, Netherlands, 139–183. doi:10.1016/S0166-4115(08)62386-9

[10] Gaole He, Nilay Aishwarya, and Ujwal Gadiraju. 2025. Is Conversational XAI All You Need? Human-AI Decision Making With a Conversational XAI Assistant. In *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25)*. Association for Computing Machinery, New York, NY, USA, 907–924. doi:10.1145/3708359.3712133

[11] Lukas-Valentin Herm. 2023. Impact Of Explainable AI On Cognitive Load: Insights From An Empirical Study. arXiv:2304.08861 [cs.AI] https://arxiv.org/abs/2304.08861

[12] Andreas Holzinger, André Carrington, and Heimo Müller. 2020. Measuring the Quality of Explanations: The System Causability Scale (SCS): Comparing Human and Machine Explanations. *KI - Künstliche Intelligenz* 34, 2 (june 2020), 193–198. doi:10.1007/s13218-020-00636-z

[13] Moritz Körber. 2019. Theoretical Considerations and Development of a Questionnaire to Measure Trust in Automation. In *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)*, Sebastiano Bagnara, Riccardo Tartaglia, Sara Albolino, Thomas Alexander, and Yushi Fujita (Eds.). Springer International Publishing, Cham, 13–30.

[14] Pei-Cheng Li and Cheng-Te Li. 2024. TCGNN: Text-Clustering Graph Neural Networks for Fake News Detection on Social Media. In *Advances in Knowledge Discovery and Data Mining*, De-Nian Yang, Xing Xie, Vincent S. Tseng, Jian Pei, Jen-Wei Huang, and Jerry Chun-Wei Lin (Eds.). Springer Nature Singapore, Singapore, 134–146.

[15] Rhema Linder, Sina Mohseni, Fan Yang, Shiva K. Pentyala, Eric D. Ragan, and Xia Ben Hu. 2021. How level of explanation detail affects human performance in interpretable intelligent systems: A study on explainable fact checking. *Applied AI Letters* 2, 4 (2021), e49. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/ail2.49 doi:10.1002/ail2.49

[16] 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., New York, USA. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[17] Christian Meske and Enrico Bunde. 2023. Design principles for user interfaces in AI-Based decision support systems: The case of explainable hate speech detection. *Information Systems Frontiers* 25, 2 (2023), 743–773.

[18] Ingrid Nunes and Dietmar Jannach. 2017. A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction* 27, 3 (2017), 393–444.

[19] Adiba Proma, Neeley Pate, James Druckman, Gourab Ghoshal, and Ehsan Hoque. 2025. Personalizing LLM Responses to Combat Political Misinformation. In *Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*. ACM, New York City USA, 134–143. doi:10.1145/3699682.3728349

[20] Tahereh Saheb, Mouwafac Sidaoui, and Bill Schmarzo. 2024. Convergence of artificial intelligence with social media: A bibliometric & qualitative analysis. *Telematics and Informatics Reports* 14 (2024), 100146.

[21] Max Schemmer, Patrick Hemmer, Maximilian Nitsche, Niklas Kühl, and Michael Vössing. 2022. A Meta-Analysis of the Utility of Explainable Artificial Intelligence in Human-AI Decision-Making. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (Oxford, United Kingdom) *(AIES '22)*. Association for Computing Machinery, New York, NY, USA, 617–626. doi:10.1145/3514094.3534128

[22] Ben Shneiderman. 2003. The eyes have it: A task by data type taxonomy for information visualizations. In *The craft of information visualization*. Elsevier, Amsterdam, Netherlands, 364–371.

[23] Dylan Slack, Satyapriya Krishna, Himabindu Lakkaraju, and Sameer Singh. 2023. Explaining machine learning models with interactive natural language conversations using TalkToModel. *Nature Machine Intelligence* 5, 8 (july 2023), 873–883. doi:10.1038/s42256-023-00692-8

[24] Kacper Sokol and Peter A Flach. 2019. Counterfactual Explanations of Machine Learning Predictions: Opportunities and Challenges for AI Safety. *SafeAI@ AAAI* 2301 (2019), 1–4.

[25] Aaron Springer and Steve Whittaker. 2019. Progressive disclosure: empirically motivated approaches to designing effective transparency. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, Marina del Ray California, 107–120. doi:10.1145/3301275.3302322

[26] Ashley Suh, Kenneth Alperin, Harry Li, and Steven R Gomez. 2025. Don't Just Translate, Agitate: Using Large Language Models as Devil's Advocates for AI Explanations. doi:10.5281/ZENODO.15170455

[27] Jörg Schlötterer Van Bach Nguyen and Christin Seifert. 2024. Xagent: A conversational XAI agent harnessing the power of large language models. 273–280 pages.

[28] 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

[29] Rosa Vicari and Nadejda Komendatova. 2023. Systematic meta-analysis of research on AI tools to deal with misinformation on social media during natural and anthropogenic hazards and disasters. *Humanities and Social Sciences Communications* 10, 1 (2023), 1–14.

[30] Danke Wu, Zhenhua Tan, Haoran Zhao, Taotao Jiang, and Ning Geng. 2024. Domain- and category-style clustering for general fake news detection via contrastive learning. *Information Processing & Management* 61, 4 (july 2024), 103725. doi:10.1016/j.ipm.2024.103725

[31] 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