Eroding the Truth-Default: A Causal Analysis of Human Susceptibility to Foundation Model Hallucinations and Disinformation in the Wild*

Alexander Loth
alexander.loth@stud.fra-uas.de
Frankfurt University of Applied Sciences
Frankfurt am Main, Germany

Martin Kappes
kappes@fra-uas.de
Frankfurt University of Applied Sciences
Frankfurt am Main, Germany

Marc-Oliver Pahl
marc-oliver.pahl@imt-atlantique.fr
IMT Atlantique, UMR IRISA, Chaire Cyber CNI
Rennes, France

arXiv:2601.22871v3 [cs.CY] 29 Oct 2026

Abstract

As foundation models (FMs) approach human-level fluency, distinguishing synthetic from organic content has become a key challenge for Trustworthy Web Intelligence.

This paper presents **JudgeGPT** and **RogueGPT**, a dual-axis framework that decouples "authenticity" from "attribution" to investigate the mechanisms of human susceptibility. Analyzing 918 evaluations across five FMs (including GPT-4 and Llama-2), we employ Structural Causal Models (SCMs) as a principal framework for formulating testable causal hypotheses about detection accuracy.

Contrary to partisan narratives, we find that political orientation shows a negligible association with detection performance ($r = -0.10$). Instead, "fake news familiarity" emerges as a candidate mediator ($r = 0.35$), suggesting that exposure may function as adversarial training for human discriminators. We identify a "fluency trap" where GPT-4 outputs (HumanMachineScore: 0.20) bypass Source Monitoring mechanisms, rendering them indistinguishable from human text.

These findings suggest that "pre-bunking" interventions should target cognitive source monitoring rather than demographic segmentation to ensure trustworthy information ecosystems.

CCS Concepts

• **Computing methodologies** → **Natural language processing**;
• **Human-centered computing** → *Empirical studies in HCI*.

Keywords

trustworthy foundation models, truth-default theory, human susceptibility, disinformation, hallucinations, causal analysis, pre-bunking, web intelligence

1 Introduction

The rapid proliferation of foundation models, particularly large language models (LLMs), presents both opportunities and challenges for web intelligence. Han et al. [7] define trustworthy machine learning as encompassing robustness, privacy, and causal integrity—properties that extend beyond model performance to the entire data-model-human pipeline. As these models improve, distinguishing between human-written and machine-generated content has become a central challenge for maintaining trustworthy information ecosystems online [20]. Recent work has examined the causal reasoning capabilities of LLMs [4] and their tendency to hallucinate [16], yet how users perceive and evaluate foundation model outputs "in the wild" remains underexplored.

Building trustworthy AI systems requires understanding not only the technical reliability of models but also the causal factors that influence human susceptibility to AI-generated misinformation and disinformation [18]. Yao et al. [34] frame web data as inherently "imperfect," subject to noise and annotation errors; we extend this insight by treating human perception errors as a form of imperfect annotation that reveals latent variables governing trust. Following Liu et al. [17], we approach the analysis of human judgment patterns as a method of causal discovery—identifying the hidden factors that mediate susceptibility.

To address this gap, we developed **JudgeGPT** (https://github.com/aloth/JudgeGPT), an interactive, open-source platform for systematically studying human perception of AI-generated content, and **RogueGPT** (https://github.com/aloth/RogueGPT), a complementary content generation engine that produces diverse news fragments using multiple foundation models. Together, these tools enable investigation of causal relationships between user characteristics, content properties, and judgment accuracy.

Our research addresses the following questions with implications for trustworthy foundation models:

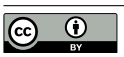
- **Causal Factors in Susceptibility:** What demographic and experiential factors causally influence users' ability to distinguish foundation model outputs from human-written content?
- **Model-Specific Vulnerabilities:** Do different foundation models vary in their ability to produce human-like content, and what are the implications for trustworthy AI deployment?
- **Learning and Adaptation:** Can exposure to AI-generated content improve detection abilities over time, suggesting potential for "pre-bunking" interventions?

1.1 Contributions

This paper offers three contributions:

- **JudgeGPT and RogueGPT Platforms:** Open-source tools for continuous data collection on human perception of foundation model outputs, featuring dual-axis assessment (authenticity and source attribution) with multilingual support.

*This research is funded by the European Union project CyberSecDome under the grant agreement 101120779. Published in WWW Companion '26, pp. 1125–1132. DOI: 10.1145/3774905.3795832.



This work is licensed under a Creative Commons Attribution 4.0 International License.

- **Causal Analysis of Human Perception:** Empirical findings from 154 participants and 918 evaluations revealing how fake news familiarity, political orientation, and other factors causally influence detection accuracy.
- **Model Comparison Insights:** Evidence that GPT-4 produces more human-like text than other models tested, with implications for trustworthy foundation model development and deployment.

1.2 Paper Structure

Section 2 situates our work within related research. Section 3 describes the JudgeGPT and RogueGPT platforms. Section 4 reports our primary findings. Section 5 discusses implications for trustworthy foundation models and future research directions.

2 Related Work

Research on trustworthy foundation models intersects with several established areas that inform our approach.

2.1 Trustworthy AI and Foundation Models

The challenge of building trustworthy AI systems has received attention as foundation models become widespread [7]. Weidinger et al. [32] provide a taxonomy of LLM risks including misinformation generation and human overreliance on fluent outputs. Key concerns include model reliability under noisy or adversarial conditions [8], the causal reasoning capabilities of LLMs [4], and understanding why models hallucinate [16]. While Chi et al. [4] demonstrate that LLMs lack "Level 2" causal reasoning, our data reveals that humans falsely attribute this capability to them due to output fluency—a phenomenon Bender et al. [1] term "stochastic parrots," where coherent-sounding text masks the absence of genuine understanding. These technical dimensions of trustworthiness are complemented by the need to understand human factors—specifically, how users perceive and interact with AI-generated content in web environments [34].

2.2 Cognitive Frameworks for AI Perception

Our analysis draws on established cognitive science frameworks. The Source Monitoring Framework [13] explains how humans use heuristic cues (detail, fluency) rather than verified source information when attributing text origins—we term this the "Fluency Trap." Truth-Default Theory, recently applied to AI [24], suggests that AI systems exhibit truth bias, which may create "validity illusions" for users who assume AI outputs are accurate; fluent foundation model outputs may thus be *eroding the truth-default* that humans apply to textual communication. Finally, dual-process theory [14] explains why detection accuracy degrades over time: sustained analytical effort (System 2) depletes cognitive resources, leading participants to rely on faster but less accurate heuristics.

2.3 Human Perception and Fake News Detection

The intersection of generative AI and misinformation presents a dual challenge: while LLMs can assist in detection, they also lower barriers to fake news generation [20, 21]. Automated fake news detection has employed techniques from multimodal architectures [31] to transformer-based systems like GROVER [35]. The RAID benchmark [6] provides the current standard for evaluating machine-generated text detectors, demonstrating that automated systems struggle with adversarial attacks and domain shifts. Our work complements RAID by providing *human-centric* metrics: while RAID measures algorithmic detection, JudgeGPT measures human detection—both are necessary for a complete picture of trustworthiness in web intelligence.

Studies have investigated how demographic factors influence credibility judgments [3, 30, 33] and factors influencing trust in machine-generated content [10]. Clark et al. [5] found that human evaluation of generated text is unreliable, with evaluators often unable to distinguish GPT-3 outputs from human writing. Ippolito et al. [12] demonstrated that automatic detection is easiest precisely when humans are fooled, suggesting complementary roles for algorithmic and human detection. The MIST test [23] measures misinformation susceptibility using psychometric analysis. However, most work focuses on binary classification (real vs. fake), missing the nuanced spectrum of human perception.

2.4 Pre-bunking and Psychological Inoculation

Our finding on learning effects connects to the literature on psychological inoculation against misinformation. Roozenbeek et al. [28, 29] have demonstrated that gamified "pre-bunking" interventions can improve resilience against misinformation on social media. The Debunking Handbook [15] synthesizes evidence on effective correction strategies, emphasizing that prebunking is more effective than debunking. Google Jigsaw's collaboration on EU election pre-bunking campaigns [2] demonstrates real-world applications of this research for web intelligence. We position JudgeGPT as a *dynamic pre-bunking environment*: beyond passive measurement, the platform's iterative exposure to AI-generated content may itself serve as a form of cognitive inoculation.

Table 1 positions JudgeGPT relative to existing approaches.

Tool	Human Eval	Dual-Axis	Multi-LLM
GROVER [35]	✗	✗	✗
MIST [23]	✓	✗	✗
FakeGPT [9]	✗	✗	✗
JudgeGPT (Ours)	✓	✓	✓

Table 1: JudgeGPT provides dual-axis human evaluation across multiple foundation models.

JudgeGPT extends prior work by: (1) capturing nuanced perceptions using continuous scales for both authenticity and source attribution; (2) enabling demographic analysis to identify causal factors in AI content perception; (3) comparing multiple foundation models to assess relative "human-likeness"; and (4) supporting multilingual evaluation for cross-cultural insights.

3 JudgeGPT and RogueGPT Platforms

Our dual-system framework combines **JudgeGPT** (https://github.com/aloth/JudgeGPT), an interactive platform for evaluating human perception of content authenticity and source, with **RogueGPT** (https://github.com/aloth/RogueGPT), a generative engine producing news fragments using multiple foundation models. Together, these platforms form the empirical foundation for our ongoing research program on human-AI interaction in the context of misinformation [20, 21].

3.1 JudgeGPT: Human Evaluation Platform

JudgeGPT implements a Streamlit-based web interface that presents news fragments and captures user evaluations along two continuous dimensions:

- **Perceived Authenticity (LegitFakeScore):** From "Completely Fake" to "Completely Legitimate" — capturing how believable the content appears
- **Source Attribution (HumanMachineScore):** From "Definitely Human-Written" to "Definitely Machine-Generated" — capturing perceived authorship

This dual-axis approach enables analysis of whether humans can distinguish AI-generated content (source attribution) independently of whether they find it credible (authenticity assessment). The platform supports four languages (English, German, French, Spanish) with automatic browser language detection. After obtaining informed consent, participants provide demographic information (age, gender, education, political orientation, fake news familiarity) before evaluating news fragments. The system logs response times and interaction patterns alongside judgments, enabling temporal analysis of learning and fatigue effects.

3.2 RogueGPT: Content Generation Engine

RogueGPT generates news fragments using multiple foundation models, enabling comparative analysis of how different models vary in producing human-like content. The currently supported models include:

- **OpenAI:** GPT-4, GPT-4o, GPT-3.5-turbo
- **Meta:** Llama-2-13b-chat
- **Google:** Gemma-7b-it
- **Mistral AI:** Mistral-7b-instruct
- **Microsoft:** Phi-3-mini-4k-instruct

This multi-model approach informs trustworthy AI deployment: if certain models consistently evade human detection, they pose greater risks for misinformation propagation on the web.

3.3 Data Collection and Metrics

The platform collects three types of data for causal analysis:

(1) **Participant demographics:** Age, gender, education level, political orientation, fake news familiarity, newspaper subscriptions, and geographic location — enabling analysis of susceptibility factors
(2) **Fragment metadata:** Source (human or AI), topic, language, generating model, and veracity (based on factual accuracy) — enabling model comparison
(3) **Assessment metrics:** Continuous slider positions for both dimensions, response times, and interaction patterns — enabling behavioral analysis

As of December 2025, the dataset includes 154 validated participants and 918 evaluations across 2301 news fragments in four languages.

4 Findings

This section presents findings from 918 valid participant responses, examining factors associated with human perception of foundation model outputs and comparative model performance.

4.1 From Correlation to Causal Hypotheses via SCMs

While correlational analysis provides initial insights, we frame our findings within a Structural Causal Model (SCM) [11, 26] to formulate explicit, testable causal hypotheses. We propose a Directed Acyclic Graph (DAG) with the following structure:

- T **(Treatment):** Model Architecture (e.g., GPT-4 vs. Mistral vs. Llama-2 [25])
- M **(Mediator):** Textual Features—decomposed into three interpretable dimensions: (1) *Surface Fluency*: sentence length variance, type-token ratio (lexical diversity), grammatical error rate; (2) *Discourse Coherence*: connective frequency ("however," "therefore"), referential consistency, topic continuity; and (3) *Causal Consistency*: logical entailment patterns, temporal plausibility, factual grounding. While we do not computationally extract these features in the current study, prior work [5, 12] demonstrates their differential role in human detectability—with surface fluency being the most salient cue and causal consistency the least utilized.
- C **(Confounders):** User Demographics (fake news familiarity, age, education, political orientation)
- Y **(Outcome):** Source Attribution (HumanMachineScore)

This formalization enables claims using Pearl's *do*-calculus: "Controlling for political orientation (C), the direct causal effect of Model Architecture (T) on detection failure (Y) is mediated by Perceived Fluency (M)." Formally, $P(Y|do(T))$ captures the interventional effect of deploying different foundation models on human susceptibility. This language aligns with the causal inference focus of the workshop [7, 17].

Methodological Note: We acknowledge that our observational design cannot fully establish causality in the interventional sense. While SCMs provide a principled framework for reasoning about causal relationships, the correlations we report should be interpreted as *causal hypotheses* requiring validation through randomized experiments (e.g., manipulating fake news exposure or model assignment). Our DAG represents a testable causal model rather than confirmed causal knowledge. Critically, we assume no unmeasured confounders between perceived fluency (M) and susceptibility (Y)—yet factors such as cognitive load, reading speed, or domain expertise could confound this relationship. Future work should employ sensitivity analyses [11] to quantify how robust our estimates are to violations of this assumption, and experimental designs that directly manipulate textual features while holding content constant.

4.2 Associational Patterns: Fake News Familiarity and Political Orientation

We analyzed associations between participant characteristics (confounders C) and judgment accuracy (Figure 1):

- **Fake news familiarity is associated with accuracy:** Moderate positive correlations exist between fake news familiarity and both source attribution ($r = 0.35$) and legitimacy assessment ($r = 0.29$). Under our proposed DAG, this association is consistent with prior misinformation experience improving detection capabilities—a hypothesis supported by findings from our larger reproduction study [22].

- **Political orientation shows weak associations:** Only weak negative correlations exist between political extremity and judgment accuracy ($r = -0.10$ for source, $r = -0.05$ for legitimacy), suggesting political orientation is not a primary factor—consistent with Pennycook and Rand's [27] finding that analytical thinking, not partisan bias, predicts fake news susceptibility.

These associational patterns, interpreted through our causal framework, suggest that interventions targeting fake news awareness may be more effective than those targeting political groups—a hypothesis warranting experimental validation.

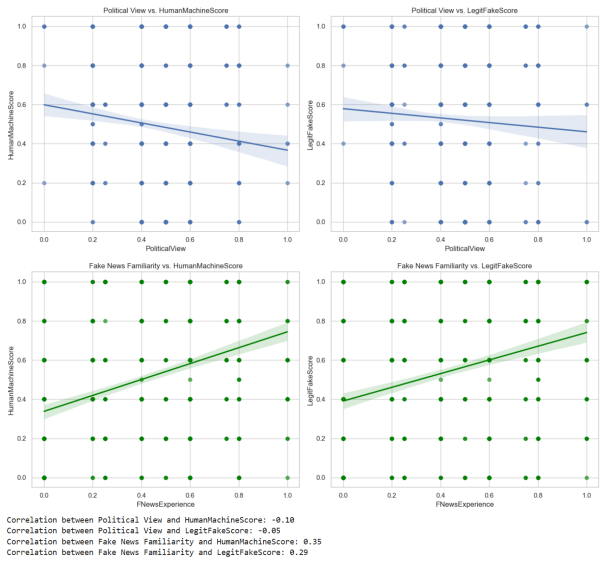


Figure 1: Correlations between participant characteristics and judgment scores. Fake news familiarity shows stronger predictive power than political orientation.

4.3 Foundation Model Comparison: Perceived Humanness

A notable finding for trustworthy foundation models is the variation in how "human-like" different models appear to users. Table 2 shows the average HumanMachineScore by model (lower = more human-like).

Result: GPT-4 outputs received lower scores (0.20) than all other models (next best: 0.44), indicating GPT-4 produces text that participants perceive as more human-like. This finding has implications for foundation model trustworthiness: as models become more capable, human detection becomes less reliable. Strikingly, when comparing human-written and AI-generated fragments in aggregate (Figure 2), independent t-tests found no significant mean differences between groups for either HumanMachineScore or LegitFakeScore—participants could not reliably distinguish AI content from human content at the population level.

Table 2: Average HumanMachineScore by foundation model (lower = more human-like).

Model	Avg. Score
openai_gpt-4	0.200
mistralai_mistral-7b	0.442
openai_gpt-4o	0.468
google_gemma-7b	0.499
meta_llama-2-13b	0.526
openai_gpt-35-turbo	0.548
microsoft_Phi-3-mini	0.580

4.4 Learning Effects and Behavioral Patterns

We examined whether participants' judgment criteria changed over the course of the experiment, possibly due to learning or fatigue. Figure 4 plots the average scores as a function of response order.

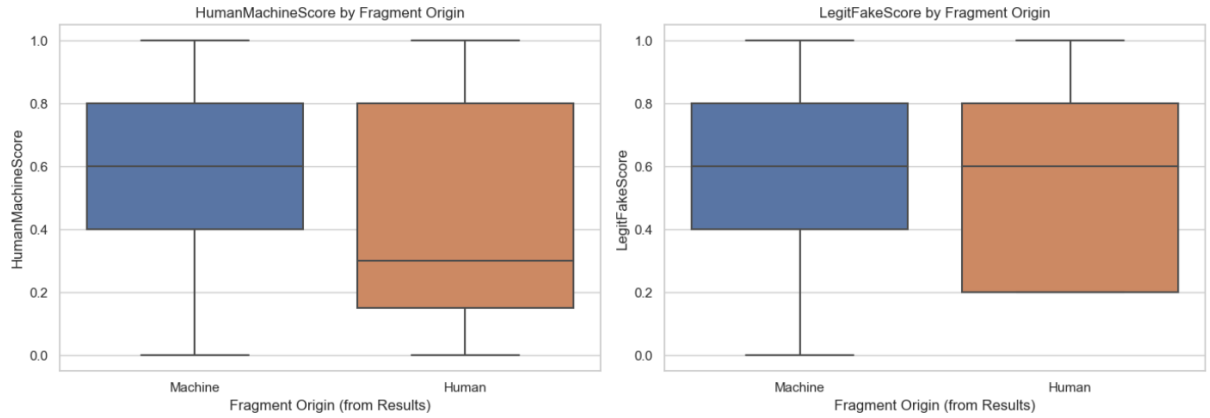
Learning Effects: The HumanMachineScore shows a downward trend early in sessions before stabilizing (Figure 4), suggesting participants improve at detecting AI text with exposure. This aligns with research on psychological inoculation against misinformation [29]. Our larger-scale analysis reveals asymmetric cognitive fatigue: fake news detection accuracy degrades under sustained exposure while AI source detection remains relatively stable [21]. This fatigue effect is consistent with dual-process theory [14], where sustained System 2 effort leads to resource depletion. This learning effect suggests that targeted "pre-bunking" training could leverage natural learning mechanisms to improve detection capabilities.

Behavioral Clustering: Clustering analysis (Figure 3) revealed distinct "judging profiles" among participants:

- **Skeptical judges:** Consistently rated content as machine-generated and fake
- **Trusting judges:** Tended to rate content as human-written and legitimate—consistent with Truth-Default Theory applied to AI [24]
- **Variable judges:** Showed inconsistent patterns, possibly reflecting genuine uncertainty

These profiles were influenced by factors including response time patterns and fake news familiarity, suggesting that individual differences in cognitive style affect susceptibility to AI-generated misinformation.

Heterogeneity of Treatment Effect (HTE):. Following Yao et al. [34], we embrace the "imperfect data" inherent to web intelligence. The variation in human response times and the distinct "Skeptical vs. Trusting" clusters (Figure 3) are not merely noise—they represent *heterogeneity of treatment effect* (HTE). Different users exhibit differential susceptibility to different models: Trusting judges are particularly vulnerable to GPT-4's fluency, while Skeptical judges show relative immunity. Identifying *which users* are susceptible to *which models* constitutes a causal discovery task [17]. JudgeGPT provides the granular, user-level data necessary to estimate these heterogeneous effects—a prerequisite for *personalized AI safety interventions* that target specific cognitive profiles rather than broad demographic categories.




```
Human vs. Machine - HumanMachineScore: t-statistic = -1.08, p-value = 0.3168
Human vs. Machine - LegitFakeScore: t-statistic = -0.36, p-value = 0.7259
```

Figure 2: Comparison of HumanMachineScore and LegitFakeScore for fragments by actual source (Human vs. AI). Boxplots show median and interquartile ranges. Independent t-tests found no significant mean differences between human-written and AI-generated groups for either metric, highlighting the challenge of distinguishing foundation model outputs.

5 Conclusion

This research addresses the human element of Web Intelligence. Through our JudgeGPT platform, we collected 918 human evaluations across four languages and identified the specific cognitive and demographic factors that mediate susceptibility to AI-generated misinformation and disinformation.

Our analysis yielded three findings:

(1) **Fake news familiarity associated with detection accuracy:** Prior experience with misinformation is associated with detection accuracy ($r = 0.35$ for source attribution, $r = 0.29$ for legitimacy), while political orientation shows a negligible association ($r = -0.10$). This suggests that media literacy interventions may be more effective than demographic targeting.
(2) **The GPT-4 "humanness gap":** GPT-4 produces text perceived as more human-like than other models (HumanMachineScore: 0.20 vs. 0.44+), quantifying the challenge for trustworthy web intelligence. This gap has implications for AI governance and deployment policies.
(3) **Learning effects and cognitive inoculation:** Participants exhibit learning over evaluation sessions, with distinct judgment profiles emerging. This aligns with psychological inoculation research [29] and suggests that "pre-bunking" training could leverage natural learning mechanisms to improve detection capabilities.

Toward Cognitive Inoculation: We align with the call by Han et al. [7] to view trustworthiness as a property spanning data, model, and human interaction. Our findings support a shift from *debunking* (post-hoc fact-checking) to *pre-bunking* (proactive inoculation).

Inoculation theory [28, 29] posits that exposing individuals to a weakened form of a persuasive argument (the "virus") triggers the production of counter-arguments (the "antibodies"), conferring resistance to future, stronger attacks. We position JudgeGPT as a *digital vaccine*: by exposing users to AI-generated text and providing immediate feedback (the "weakened dose"), we force users to update their Source Monitoring heuristics [13]. The learning curve observed in Figure 4 visualizes this "antibody production"—the strengthening of cognitive defenses through controlled exposure.

Policy Recommendations for Web Intelligence: Based on our findings, we propose three complementary interventions:

(1) **Mandatory Provenance (C2PA):** Since human detection of GPT-4 is failing (HumanMachineScore: 0.20), relying on "user vigilance" is a failed policy. We advocate for cryptographic watermarking standards (e.g., C2PA) [19] to provide an external verification layer that does not depend on fallible human perception.
(2) **Cognitive Security Education:** Digital literacy programs must pivot from "checking sources" (increasingly impossible with AI-generated content) to "checking logic" (causal consistency) [18]. Users must be trained to identify the *absence* of causal reasoning [4] in hallucinations—a skill that targets the fluency trap directly.
(3) **Adversarial User Training:** Platforms should deploy adversarial training for users—periodically presenting synthetic content with verification prompts—to keep Source Monitoring heuristics calibrated to evolving model capabilities. Inspired by Google Jigsaw's EU election pre-bunking campaigns [2], such interventions transform passive consumption into active cognitive exercise.

Toward Model-Side Transparency: While our findings focus on human susceptibility, we identify concrete opportunities for operationalizing insights into model-level constraints and training objectives:

- **Stylistic Fingerprinting:** Models could be trained with an auxiliary loss that maintains detectable stylistic signatures (e.g., characteristic n-gram distributions, punctuation patterns) while preserving task performance. This creates a "watermark by design" that automated detectors can exploit.



Figure 3: Pair-plot visualization of participant judging behavior clusters. Each point represents a participant, colored by cluster. The plots show relationships between HumanMachineScore, LegitFakeScore, TimeToAnswer, PoliticalView, and FNewsExperience. Distinct clusters indicate different judgment behavior patterns that may be causally linked to participant characteristics.

- **Uncertainty Surfacing:** Rather than producing uniformly confident prose, models should be trained with calibration objectives that surface epistemic uncertainty through hedging language ("evidence suggests," "it is unclear whether") when generating claims with low confidence scores from the model's own probability distributions.
- **Provenance Metadata:** API responses should include machine-readable provenance headers (model ID, generation timestamp, temperature settings) enabling downstream verification pipelines [19]. This can be enforced at the infrastructure level without modifying model weights.
- **RLHF for Detectability:** Reinforcement learning from human feedback could incorporate a multi-objective reward: $R = \alpha R_{\text{utility}} + \beta R_{\text{detectability}}$, where the detectability reward is derived from human evaluator accuracy on held-out samples. This explicitly trades off deceptive capability against task performance.
- **Fluency-Gated Safety Classifiers:** Our finding that GPT-4's fluency bypasses Source Monitoring suggests that safety classifiers should weight fluency as a *risk factor*—highly fluent outputs warrant additional scrutiny for factual grounding before release.

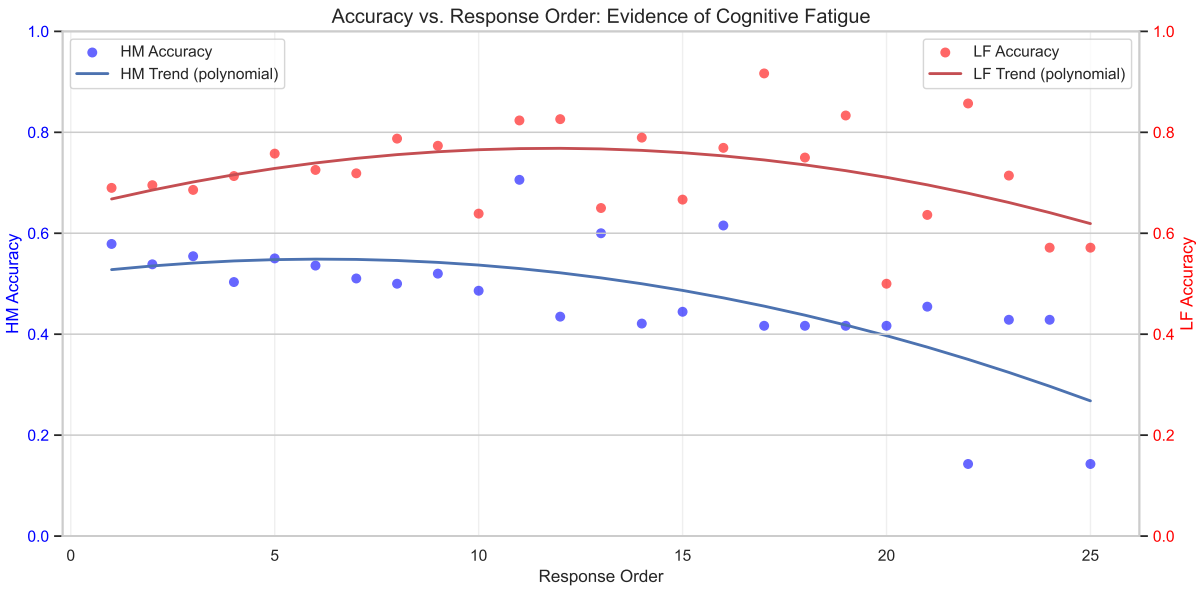


Figure 4: Judgment Scores vs. Response Order. HumanMachineScore and LegitFakeScore are plotted over the sequence of trials for each participant (with smoothing). The HumanMachineScore tends to decrease initially (learning effect), while the LegitFakeScore drops later in the session (possible fatigue or strategy shift toward increased skepticism).

Limitations and Future Work: We acknowledge several constraints that bound our conclusions:

- **Sample Size and Regional Bias:** The current dataset ($N = 154$ participants, 918 evaluations), while multilingual, is skewed toward educated European participants. This limits generalizability to populations with different media ecosystems, education levels, or cultural orientations toward authority and technology. We are actively expanding data collection with more diverse populations [22].
- **Content Domain Variation:** Our corpus spans political and general news topics but does not systematically analyze domain-specific effects. Detection accuracy may vary significantly between political misinformation, health claims, and scientific content—each domain carries distinct prior expectations and verification heuristics.
- **Cultural and Linguistic Variation:** While we support four languages, cultural factors influencing trust (e.g., collectivism vs. individualism, institutional trust levels, media literacy traditions) are not explicitly modeled. Moreover, LLM fluency varies across languages—models are typically more fluent in English than in lower-resource languages, potentially creating language-dependent "fluency traps." Cross-cultural replication with language-stratified analysis is needed to validate whether susceptibility patterns generalize.
- **Longitudinal Effects:** Our study captures within-session learning but does not track retention or decay of detection skills over weeks or months. Inoculation research [28] suggests effects may attenuate without booster exposures; understanding this decay curve is critical for designing sustainable intervention programs.

From Association to Intervention: Future work must move beyond observing correlations to performing interventions. We propose using RogueGPT to generate text with specific "causal flaws" (e.g., impossible physics, logical contradictions, temporal inconsistencies). By measuring whether users can detect these causal errors better than stylistic errors, we can directly test the "Level 2" causal reasoning hypothesis of Chi et al. [4]. This would reveal whether training users on causal consistency—rather than surface fluency—offers a more robust defense against AI-generated misinformation.

Closing the Trustworthy Cycle: We envision a continuous improvement loop grounded in the framework of Han et al. [7]:

(1) **Detection:** Human feedback via JudgeGPT identifies detection failures and susceptibility patterns.
(2) **Causal Discovery:** Analysis identifies the textual features (M) and user characteristics (C) driving these failures [17].
(3) **Model Improvement:** Discovered features inform "safety classifiers" or RLHF fine-tuning to reduce deceptive mimicry.
(4) **Re-evaluation:** Updated models are re-evaluated by humans, restarting the cycle.

This *Trustworthy Cycle*—integrating human perception, causal inference, and model refinement—provides a framework for developing trustworthy web ecosystems. Future work will incorporate additional foundation models (including Claude, Gemini, and Llama-3 [25]) and investigate whether detection accuracy varies by content domain (e.g., political vs. scientific misinformation). Our ongoing research program includes detailed analysis of the AI-driven disinformation kill chain [21] and reproduction of key findings from the disinformation literature [22].

Open Science: Both platforms are available as open-source projects. JudgeGPT is accessible at https://judgegpt.streamlit.app

(code: https://github.com/aloth/JudgeGPT), and RogueGPT at https://github.com/aloth/RogueGPT. We invite researchers to contribute to this ongoing investigation of human-AI interaction in the context of misinformation. As our study continues, we kindly invite experts to participate in our survey: https://github.com/aloth/verification-crisis.

References

[1] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. 610–623. doi:10.1145/3442188.3445922

[2] Mikey Biddlestone, Beth Goldberg, Melisa Basol, Katie Washington, Sara Elnusairi, Anneka Sharpley, Meghan Graham, Sander van der Linden, Ricky Green, Jon Roozenbeek, Rachel Xu, and Andrew Pel. 2026. Video inoculation against election misinformation across 12 EU nations. *Communications Psychology* 4, 1 (2026), 77. doi:10.1038/s44271-025-00379-3

[3] Samantha Bradshaw, Hannah Bailey, and Philip N. Howard. 2021. *Industrialized Disinformation: 2020 Global Inventory of Organized Social Media Manipulation.* Technical Report. Computational Propaganda Research Project, Oxford Internet Institute, Oxford, UK. https://demtech.oii.ox.ac.uk/research/posts/industrialized-disinformation/

[4] Haoang Chi, He Li, Wenjing Yang, Feng Liu, Long Lan, Xiaoguang Ren, Tongliang Liu, and Bo Han. 2024. Unveiling Causal Reasoning in Large Language Models: Reality or Mirage?. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*. 96640–96670. doi:10.52202/079017-3064

[5] Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 7282–7296. doi:10.18653/v1/2021.acl-long.565

[6] Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024. RAID: A Shared Benchmark for Robust Evaluation of Machine-Generated Text Detectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 12463–12492. doi:10.18653/v1/2024.acl-long.674

[7] Bo Han, Jiangchao Yao, Tongliang Liu, Bo Li, Sanmi Koyejo, and Feng Liu. 2025. Trustworthy Machine Learning: From Data to Models. *Foundations and Trends® in Privacy and Security* 7, 2-3 (2025), 74–246. doi:10.1561/3300000043

[8] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. 2018. Co-teaching: Robust Training of Deep Neural Networks with Extremely Noisy Labels. In *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*. arXiv:1804.06872 [cs.LG]

[9] Yue Huang and Lichao Sun. 2024. FakeGPT: Fake News Generation, Explanation and Detection of Large Language Models. arXiv:2310.05046 [cs.CL]

[10] Martin Huschens, Martin Briesch, Dominik Sobania, and Franz Rothlauf. 2023. Do You Trust ChatGPT? – Perceived Credibility of Human and AI-Generated Content. arXiv:2309.02524 [cs.HC] https://arxiv.org/abs/2309.02524

[11] Guido W. Imbens and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press. doi:10.1017/CBO9781139025751

[12] Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic Detection of Generated Text is Easiest when Humans are Fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*. 1808–1822. doi:10.18653/v1/2020.acl-main.164

[13] Marcia K. Johnson, Shahin Hashtroudi, and D. Stephen Lindsay. 1993. Source monitoring. *Psychological Bulletin* 114, 1 (1993), 3–28. doi:10.1037/0033-2909.114.1.3

[14] Daniel Kahneman. 2011. *Thinking, fast and slow.* Farrar, Straus and Giroux.

[15] Stephan Lewandowsky, John Cook, Ullrich K. H. Ecker, Dolores Albarracín, Michelle A. Amazeen, Panayiota Kendeou, Doug Lombardi, Eryn J. Newman, Gordon Pennycook, Ethan Porter, et al. 2020. The Debunking Handbook 2020. Databrary. doi:10.17910/b7.1182

[16] He Li, Haoang Chi, Mingyu Liu, and Wenjing Yang. 2024. Look Within, Why LLMs Hallucinate: A Causal Perspective. arXiv:2407.10153 [cs.CL]

[17] Chenxi Liu, Yongqiang Chen, Tongliang Liu, Mingming Gong, James Cheng, Bo Han, and Kun Zhang. 2024. Discovery of the Hidden World with Large Language Models. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*. 102307–102365. doi:10.52202/079017-3249

[18] Alexander Loth. 2021. *Decisively Digital: From Creating a Culture to Designing Strategy.* John Wiley & Sons, Inc., Hoboken, NJ, USA. https://www.wiley.com/en-us/Decisively+Digital%3A+From+Creating+a+Culture+to+Designing+Strategy-p-9781119737285

[19] Alexander Loth, Dominique Conceicao Rosario, Peter Ebinger, Martin Kappes, and Marc-Oliver Pahl. 2026. Origin Lens: Reclaiming Trust on the AI-Mediated Web Through On-Device Image Provenance Verification. In *Companion Publication of the 2026 18th ACM Web Science Conference (WebSci Companion '26)* (Braunschweig, Germany). ACM, New York, NY, USA, 1–5. doi:10.1145/3795513.3806658

[20] Alexander Loth, Martin Kappes, and Marc-Oliver Pahl. 2024. Blessing or Curse? A Survey on the Impact of Generative AI on Fake News. doi:10.48550/arXiv.2404.03021 arXiv:2404.03021 [cs.CL]

[21] Alexander Loth, Martin Kappes, and Marc-Oliver Pahl. 2026. Industrialized Deception: The Collateral Effects of LLM-Generated Misinformation on Digital Ecosystems. In *Companion Proceedings of the ACM Web Conference 2026 (WWW Companion '26)* (Dubai, United Arab Emirates). ACM, New York, NY, USA, 295–302. doi:10.1145/3774905.3795471 arXiv:2601.21963 [cs.CY]

[22] Alexander Loth, Martin Kappes, and Marc-Oliver Pahl. 2026. The Verification Crisis: Expert Perceptions of GenAI Disinformation and the Case for Reproducible Provenance. In *Companion Proceedings of the ACM Web Conference 2026 (WWW Companion '26)* (Dubai, United Arab Emirates). ACM, New York, NY, USA, 980–988. doi:10.1145/3774905.3795484 arXiv:2602.02100 [cs.CY]

[23] Rakoen Maertens, Friedrich M. Götz, Hudson F. Golino, Jon Roozenbeek, Claudia R. Schneider, Yara Kyrychenko, John R. Kerr, Stefan Stieger, William P. McClanahan, Karly Drabot, James He, and Sander van der Linden. 2024. The Misinformation Susceptibility Test (MIST): A psychometrically validated measure of news veracity discernment. *Behavior Research Methods* 56, 3 (2024), 1863–1899. doi:10.3758/s13428-023-02124-2

[24] David M. Markowitz and Jeffrey T. Hancock. 2024. Generative AI Are More Truth-Biased Than Humans: A Replication and Extension of Core Truth-Default Theory Principles. *Journal of Language and Social Psychology* 43, 2 (2024), 261–267. doi:10.1177/0261927X231220404

[25] Meta AI. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783* (2024).

[26] Judea Pearl. 2009. *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge University Press, Cambridge, UK. doi:10.1017/CBO9780511803161

[27] Gordon Pennycook and David G. Rand. 2019. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* 188 (2019), 39–50. doi:10.1016/j.cognition.2018.06.011

[28] Jon Roozenbeek, Eileen Culloty, and Jane Suiter. 2023. Countering Misinformation: Evidence, Knowledge Gaps, and Implications of Current Interventions. *European Psychologist* 28, 3 (2023), 189–205. doi:10.1027/1016-9040/a000492

[29] Jon Roozenbeek, Sander van der Linden, Beth Goldberg, Steve Rathje, and Stephan Lewandowsky. 2022. Psychological inoculation improves resilience against misinformation on social media. *Science Advances* 8, 34 (2022), eabo6254. doi:10.1126/sciadv.abo6254

[30] Sander van der Linden. 2022. Misinformation: susceptibility, spread, and interventions to immunize the public. *Nature Medicine* 28, 3 (2022), 460–467. doi:10.1038/s41591-022-01713-6

[31] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18)*. ACM, 849–857. doi:10.1145/3219819.3219903

[32] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. 2022. Taxonomy of Risks posed by Language Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. ACM, 214–229. doi:10.1145/3531146.3533088

[33] Jeongwon Yang and Yu Tian. 2021. "Others are more vulnerable to fake news than I Am": Third-person effect of COVID-19 fake news on social media users. *Computers in Human Behavior* 125 (2021), 106950. doi:10.1016/j.chb.2021.106950

[34] Jiangchao Yao, Feng Liu, and Bo Han. 2025. Trustworthy AI under Imperfect Web Data. In *Companion Proceedings of the ACM Web Conference 2025 (WWW Companion '25)*. ACM, 65–68. doi:10.1145/3701716.3715859

[35] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending Against Neural Fake News. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*. arXiv:1905.12616 [cs.CL]

Supplementary Materials

This supplement documents the JudgeGPT platform in detail. We provide data schema descriptions and annotated user interface screenshots that illustrate the data collection methodology. These materials facilitate replication and ensure transparency regarding the experimental design.

A Data Description

This study uses three main data collections: Participants, Fragments, and Results. Each collection is stored in JSON format and provides granular attributes for in-depth analysis of participant attributes, fragment characteristics, and evaluation outcomes.

A.1 Participants Collection

The Participants collection provides demographic, linguistic, and contextual information about each participant. Each participant is identified by a unique `ParticipantID`. Associated attributes include:

- *Demographics:* `Age`, `Gender`, `EducationLevel`
- *Attitudes:* `PoliticalView`, `FNewsExperience`, `NewspaperSubscription`
- *Language:* `Language`, `IsNativeSpeaker`
- *Context:* `ScreenResolution`, `IpLocation`, `UserAgent`

Political views are numerically encoded and mapped to categorical labels, as shown in Table S1. Familiarity with fake news (FNewsExperience) is encoded similarly, ranging from completely unfamiliar to completely familiar, with placeholder options where no choice is made (Table S2). Additional contextual metadata, such as geolocation and browser information, provides environmental context for each participant's judgments.

Table S1: Political View Encoding

Value	Label
0.0	Far Left
0.2	Left
0.4	Center-Left
0.5	Placeholder (no choice)
0.6	Center-Right
0.8	Right
1.0	Far Right

Table S2: Fake News Familiarity Encoding (FNewsExperience)

Value	Label
0.0	Completely unfamiliar
0.2	Mostly unfamiliar
0.4	Somewhat unfamiliar
0.5	Placeholder (no choice)
0.6	Somewhat familiar
0.8	Mostly familiar
1.0	Completely familiar

A.2 Fragments Collection

The Fragments collection contains textual items (fragments) that participants are asked to evaluate. Each fragment has attributes including `Language`, `Style`, `Format`, `SeedPhrase`, `FragmentID`, `Content`, `Origin` (human or machine), and `IsFake`. Some fragments are styled to mimic certain news outlets or social media platforms, while others incorporate seed phrases that set topical or thematic content. The `CreationDate` of each fragment is recorded, allowing temporal analysis. This collection enables comparisons of different fragment properties (e.g., style, origin) and their effects on perceived authenticity.

A.3 Results Collection

The Results collection links participants to fragments and stores their assessment outcomes. Each result record includes `ResultID`, `ParticipantID`, `FragmentID`, `HumanMachineScore`, `LegitFakeScore`, `TopicKnowledgeScore`, `Timestamp`, `TimeToAnswer`, `SessionCount`, `Origin`, and `IsFake`. The HumanMachineScore, LegitFakeScore, and TopicKnowledgeScore are encoded on a 0.0 to 1.0 scale and map to interpretive labels, as shown in Tables S3, S4, and S5. By combining Results with Participants and Fragments data, it is possible to analyze how participant attributes and fragment properties influence judgments of authenticity and how participants' knowledge or familiarity impacts their decision-making.

Table S3: Human vs. Machine Score Encoding (HumanMachineScore)

Value	Label
0.0	Definitely Human Generated
0.2	Probably Human Generated
0.4	Likely Human Generated
0.5	Placeholder (no choice)
0.6	Likely Machine Generated
0.8	Probably Machine Generated
1.0	Definitely Machine Generated

Table S4: Legit vs. Fake News Score Encoding (LegitFakeScore)

Value	Label
0.0	Definitely Legit News
0.2	Probably Legit News
0.4	Likely Legit News
0.5	Placeholder (no choice)
0.6	Likely Fake News
0.8	Probably Fake News
1.0	Definitely Fake News

B JudgeGPT Interface Screenshots

The following figures document the JudgeGPT web application interface, which was developed using Streamlit and deployed as an open-access platform. The platform implements a within-subjects design where each participant first completes a demographic intake form (Figure S1) before evaluating a randomized sequence of news fragments (Figure S2). This design enables us to capture both individual difference variables and learning effects across evaluation trials, consistent with Source Monitoring Framework methodology.

Table S5: Topic Knowledge Score Encoding (TopicKnowledgeScore)

Value	Label
0.0	Not at all
0.2	Slightly
0.4	Somewhat
0.5	Placeholder (no choice)
0.6	Fairly well
0.8	Very well
1.0	Extremely well

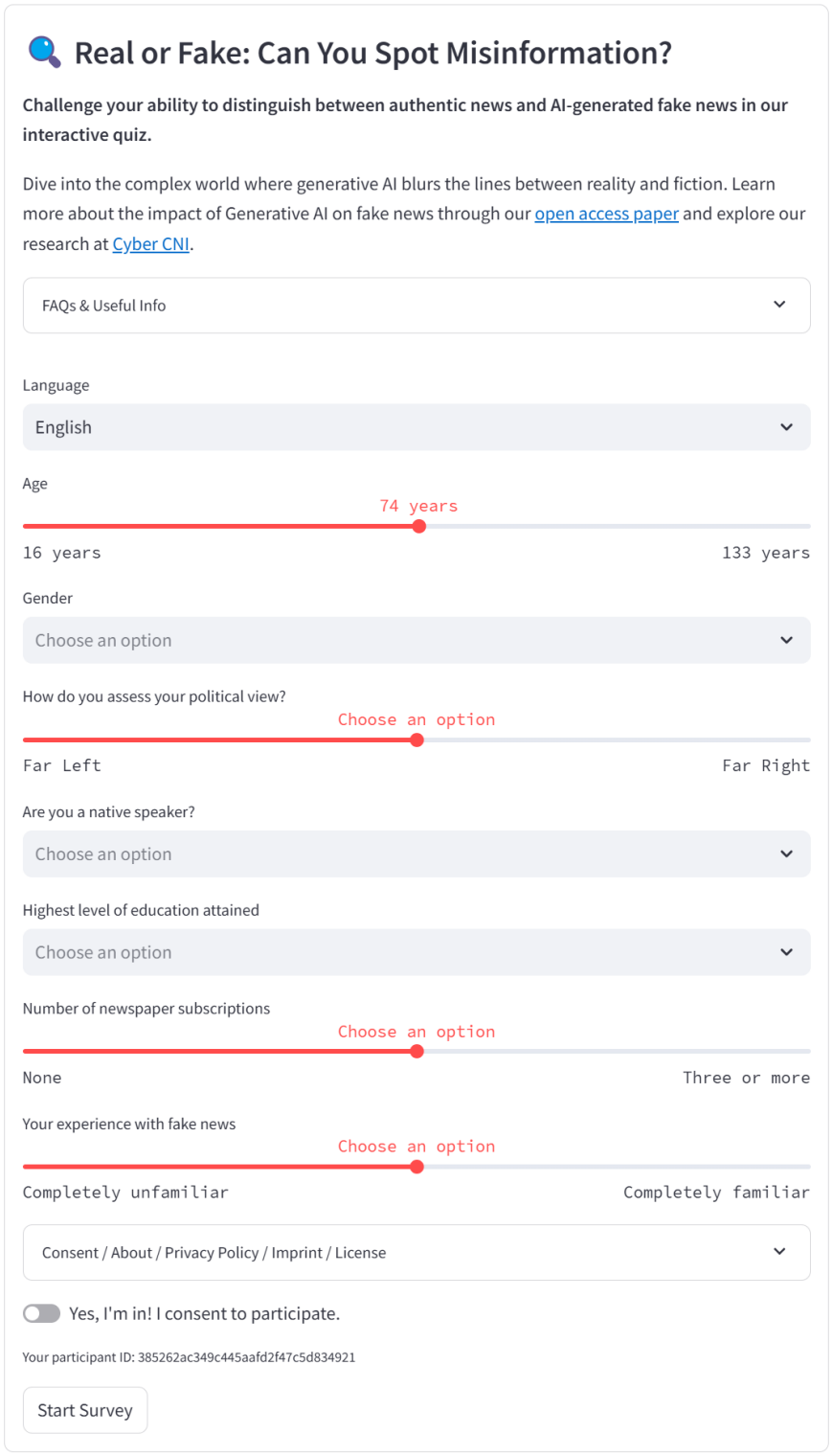


Figure S1: Screenshot of the JudgeGPT Demographic Information Screen. This initial intake screen captures participant demographics including age, gender, education level, political orientation, and self-reported familiarity with fake news—variables central to our analysis (Section 4).

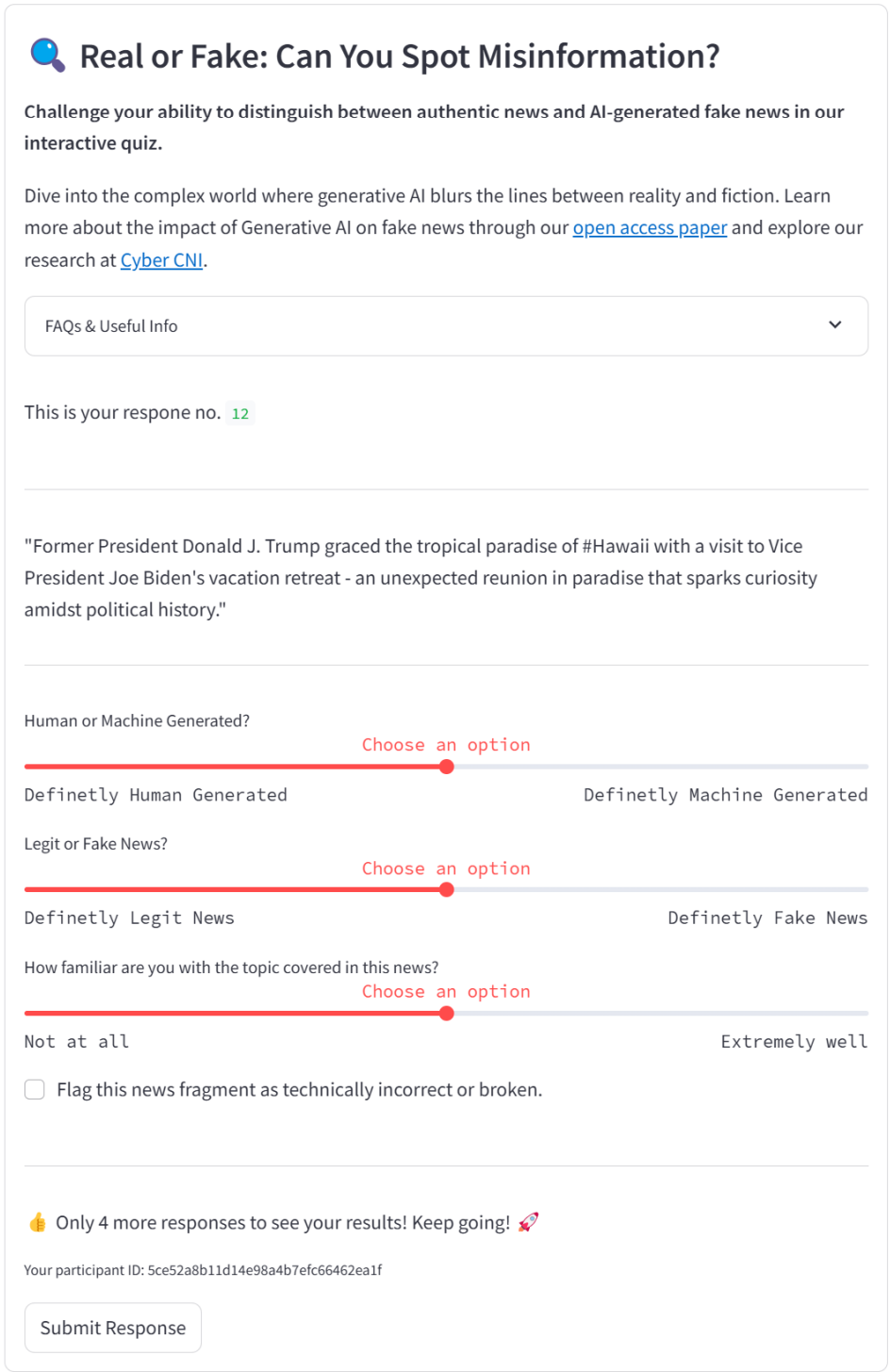


Figure S2: Screenshot of the JudgeGPT Survey Screen. Participants evaluate randomized news fragments using three continuous sliders: the *HumanMachineScore* (ranging from "Definitely Human Generated" to "Definitely Machine Generated"), the *Legit-FakeScore* (ranging from "Definitely Legit News" to "Definitely Fake News"), and a topic familiarity rating. A flag option allows participants to report technically broken content. This multi-dimensional rating design enables independent measurement of source attribution accuracy and content credibility, as analyzed in Section 4.