Beyond "Made with AI": Visualizing Provenance Density to Mitigate the Transparency Penalty

Qing Zhang[1] , **Yifei Huang**[2] , **Juyoung Lee**[3] , **Thad Starner**[4] , **Jun Rekimoto**[5]
[1]The University of Tokyo
[2]Institute of Industrial Science, The University of Tokyo
[3]Korea Advanced Institute of Science and Technology
[4]Georgia Institute of Technology
[5]Sony CSL Kyoto
qzkiyoshi@gmail.com, hyf015@gmail.com, ejuyoung@gmail.com, thad.starner@gmail.com, rekimoto@acm.org

Abstract

As generative AI makes polished prose cheap to produce, users can no longer rely on fluency as a proxy for truth. We call this failure mode the Fluency Trap: users trust fluent hallucinations while also discounting accurate content once it is disclosed as AI-generated. Binary "Made with AI" labels respond with authorship disclosure, but they do not show what supports a claim. We propose Provenance Density, an evidence-visualization interface that shows the density of verified claims in a text. In a user study with 81 participants, an idealized Provenance Density interface produced a large discernment gap between truth and fabrication ($+4.15$ points, $d = 1.82$), whereas participants given no signal showed no detectable discrimination. A technical audit with 200 samples shows that retrieval density alone is insufficient; unexpectedly, the Consistency Veto carries most of the discriminative signal on dynamic queries. As AI-generated content becomes indistinguishable from human writing, effective transparency must move from authorship disclosure toward evidence visualization.

1 Introduction

The modern information ecosystem relies on a *fluency heuristic*: we assume that high-quality, articulate, polished text signals human effort and, by extension, credibility [Hertwig *et al.*, 2008]. For most of history, this heuristic assumption was rational [Hertwig *et al.*, 2008; Marewski and Schooler, 2011]. *Costly Signaling Theory* explains why: a signal is trustworthy only when faking it is prohibitively expensive [Spence, 1978; Zahavi, 1975; Gintis *et al.*, 2001]. Producing polished prose once required years of education and editorial labor. Thus, fluency served as a reliable proxy for competence, stabilizing trust in science, law, and journalism.

Generative AI disrupts this mechanism by reducing the marginal cost of fluency to near-zero [Galdin and Silbert, 2025]. Large Language Models (LLMs) enable the mass production of professional-sounding text regardless of the author's underlying expertise. This collapse of the "separating equilibrium" creates what we term a *Fluency Trap*: a structural vulnerability where users continue to trust fluent text as if it were costly, even when it is generated cheaply by systems indifferent to truth. Psychological evidence on the *Illusion of Truth* suggests that ease of processing suppresses epistemic vigilance [Hasher *et al.*, 1977; Reber and Schwarz, 1999; Dechêne *et al.*, 2010; Sperber *et al.*, 2010]. This tendency leaves humans susceptible to "hallucinated plausibility", text that is syntactically perfect but semantically ungrounded [Reber and Unkelbach, 2010; Unkelbach *et al.*, 2011].

Current governance responses, specifically binary "Made with AI" disclosures, fail to address this decoupling. By focusing on *identity* ("Who wrote this?") rather than *provenance* ("What supports this?"), these measures often induce *epistemic stigmatization* rather than calibrated skepticism [Nakano *et al.*, 2026; Simon *et al.*, 2023]. Users frequently discount accurate AI-generated information solely due to its synthetic origin, a phenomenon we characterize as a "Transparency Penalty."

We frame Provenance Density as a *cognitive affordance* for reading in the LLM era. Instead of asking users to evaluate veracity from prose alone, the interface shifts attention toward extrinsic evidence: high-contrast indicators visualize the density of verified claims, offloading part of the verification burden from working memory to the interface [Chirayath *et al.*, 2025; Clark and Chalmers, 1998].

We validate this approach through a dual-stream evaluation. First, we conduct an automated technical audit ($N = 200$) on a composite of *TruthfulQA* [Lin *et al.*, 2022] and *FreshQA* [Vu *et al.*, 2024] to test robustness against both adversarial misconceptions and dynamic ambiguity. Second, we run a within-subjects user experiment with 81 participants to measure truth discernment. Our results empirically confirm the Fluency Trap: in the absence of signals, users failed to distinguish high-fluency hallucinations ($M = 6.28$) from ground truth ($M = 5.78$; $p = .43$). While binary labels acted as a blunt

The user study was approved by the University of Tokyo institutional review process. Code, data, prompts, and audit materials are available at https://github.com/artisticsciencex/ijcai-ecai-2026-provenance-density.

warning, Provenance Density restored truth discernment under correct signaling ($d = 1.82$, $p < .001$).

We make three main contributions: **Theory:** We synthesize the mechanics of the Fluency Trap (Section 2), detailing how RLHF-driven sycophancy structures the decoupling of fluency from veracity. **Design:** We propose **Provenance Density** (Section 3), a formalized metric ($D(T)$) and interaction paradigm that imposes a computational verification handicap on generated text. **Evidence:** We provide convergent empirical results (Section 4 & 5.1). Our technical audit ($N = 200$) confirms the metric's "Consistency Veto" against confident hallucinations, while our user study ($N = 81$) demonstrates that the interface increases the truth discernment gap to +4.15 points.

2 Related Works

We argue that the decoupling of fluency from veracity is not an accidental byproduct of LLM scaling, but a structural inevitability driven by two converging factors: an economic shift from costly signaling to cheap talk, and a technical objective function that prioritizes plausibility over truth.

From Hallucination to Indifference. While early critiques of Large Language Models (LLMs) focused on "hallucinations" as sporadic errors, recent scholarship suggests a more structural diagnosis. The framework of "Machine Bullshit" has been proposed to distinguish these outputs from lying, defined instead by the model's fundamental indifference to truth value [Liang *et al.*, 2025]. Analysis of the "Bullshit Index" reveals that Reinforcement Learning from Human Feedback (RLHF) exacerbates this issue, incentivizing models to prioritize rhetorical plausibility and "paltering" (misleading use of truth) over factual grounding. Consequently, the resulting text is optimized to bypass human epistemic vigilance.

This structural indifference renders traditional governance mechanisms, such as binary warning labels, largely ineffective. Empirical evaluations demonstrate a "failure of inoculation": while pre-emptive warnings about AI fallibility successfully reduce global trust in the system, they fail to mitigate reliance on specific misleading articles once the user is engaged with the content [Spearing *et al.*, 2025]. This discrepancy, where users theoretically acknowledge AI bias but practically accept AI fluency, underscores the limitations of heuristic warnings and motivates our proposal for granular Provenance Density indicators.

The Structural Decoupling of Fluency. Our analysis is grounded in the *Handicap Principle*, which posits that reliable signals must impose a cost on the signaler [Zahavi, 1975]. Historically, linguistic polish functioned as this handicap, creating a Separating Equilibrium where articulate text correlated with competence [Spence, 1978]. Generative AI collapses this balance into a Pooling Equilibrium, where expert testimony and fabrication can share the same polished form [Spence, 1978; Galdin and Silbert, 2025].

This shift is sustained by the technical alignment of modern LLMs. Reinforcement Learning from Human Feedback (RLHF) can incentivize sycophancy, where models prioritize user agreement over factual accuracy [Sharma *et al.*, 2024]. Evidence shows that models will agree with illogical premises to remain "helpful" [Chen *et al.*, 2025] and flip arguments to match user views [Kaur, 2025]. This results in "Machine Bullshit"—text optimized for rhetorical persuasion rather than truth [Liang *et al.*, 2025]. This decoupling can become self-amplifying: recursive training on such high-fluency, low-entropy data leads to *Model Collapse*, where the "tails" of human variance are lost to a homogenized mean [Shumailov *et al.*, 2024].

The Failure of Post-Hoc Governance. Current governance relies on the assumption that users, once warned, can critically evaluate AI-generated text. However, empirical work shows that polished AI prose can still appear highly credible [Jakesch *et al.*, 2023], and AI literacy does not reliably translate into consistent fact-checking behavior [Rheu and Cho, 2025].

Binary disclosures (e.g., "Made with AI") often exacerbate the issue through Epistemic Stigmatization. Studies show a "transparency penalty" where disclosure reduces perceived trustworthiness even when content quality is constant [Nakano *et al.*, 2026; Cheong *et al.*, 2025]. This creates a worst-case scenario: warnings trigger global skepticism (cynicism) without enabling local verification (vigilance). Users effectively whitelist human disinformation while blacklisting accurate AI assistance [Buder and Unfried, 2024].

Limitations of Automated Detection. While automated detectors are proposed as an enforcement mechanism, they exhibit substantial bias against linguistically marginalized groups. Detectors relying on perplexity heuristics frequently misclassify the lower-perplexity lexical patterns of non-native speakers as AI-generated [Liang *et al.*, 2023], leading to "hermeneutical access injustice" [Kay *et al.*, 2024]. The adversarial nature of detection also creates an arms race that generators inevitably win. We therefore shift the paradigm from *post-hoc detection* of authorship to *intrinsic signaling* of Provenance Density.

3 Defining Provenance Density

To function as a costly signal in the game-theoretic sense, Provenance Density must be computationally hard to fake. We define the Provenance Density score D for a given text T as a theoretical function of verifiable claims weighted by source reputation, contextual relevance, and internal semantic consistency.

3.1 Proposed Metric: $D(T)$

We propose a density metric that integrates external verification (Retrieval Augmented Generation) [Lewis *et al.*, 2020] with internal uncertainty quantification. Drawing on Farquhar et al. [2024], we introduce a penalty term P_{int} based on the Semantic Entropy of the claim to filter out confident hallucinations.

Let C be the set of distinct factual claims in T, and $C_v \subseteq C$ be the subset of claims successfully grounded in external evidence. Let S_c be the set of independent root domains supporting a verified claim c. We define the Provenance Density score $D(T) \in [0, 1)$ as:

$$D(T) = (1 - P_{int}) \cdot \tanh\left(\frac{1}{\beta} \sum_{c \in C} \left(\sum_{s \in S_c} w(s, c) \right)^{\lambda}\right) \quad (1)$$

Equivalently, we first raise each claim's source sum to the power λ, then sum across claims. Our design choices target specific theoretical properties revealed during technical validation:

Consistency Veto ($1-P_{int}$): We model Semantic Entropy as a linear gating factor. For each query, we draw $K = 5$ stochastic samples and compute $\rho_{\text{consistent}}$, the fraction of unordered sample pairs whose bidirectional NLI contradiction probability remains below 0.5. We define $P_{int} = 1 - \rho_{\text{consistent}}$, so internally stable generations have $P_{int} \approx 0$ and mutually inconsistent generations approach $P_{int} \to 1$. As internal uncertainty approaches maximum, the score collapses to zero regardless of external evidence. This term ensures the system filters out "fluent bullshit" (high confidence, low consistency) before verification occurs.

Cubic Contextual Weight ($w(s, c)$): Unlike traditional citation metrics that rely solely on domain authority, we define $w(s, c)$ as the product of source reputation and semantic relevance:

$$w(s, c) = \text{Reputation}(s) \cdot (\text{MatchRatio})^3$$

We impose a cubic exponent to aggressively suppress weak conceptual matches. Our preliminary tests indicated that quadratic weighting (x^2) remained too permissible of "keyword stuffing," where irrelevant sources share surface-level vocabulary (MatchRatio ≈ 0.5). The cubic term ($0.5^3 = 0.125$) acts as a noise filter, ensuring that only sources with high semantic alignment contribute meaningfully to the Provenance Density score.

Saturation Scalar (β): We introduce β as a density temperature parameter (set to $\beta = 5.0$ in our reference implementation). This term scales the input to the hyperbolic tangent function, preventing the score from saturating to 1.0 based on a single citation. It forces the model to provide accumulated, multi-source evidence to achieve a perfect density signal.

Concentration Exponent (λ): We set $\lambda = 1.2 > 1$, applied to the per-claim source sum before aggregation. With $\lambda > 1$, the term is superlinear in per-claim evidence: a claim supported by several aligned sources contributes more than the same total evidence spread thinly across several claims. This concentrates score on well-corroborated claims rather than rewarding shallow citation across many claims. We note that citation stuffing within a single claim is suppressed upstream by the cubic relevance penalty in $w(s, c)$, not by λ.

This formulation imposes a significant "handicap" on the generator: achieving a high $D(T)$ requires the model to perform costly verification and achieve strict semantic alignment between claims and sources.

3.2 Implementation of $D(T)$

Equation 1 is instantiated by four concrete steps: claim segmentation, evidence retrieval, source-relevance scoring, and aggregation; key implementation parameters are summarized in Table 1. We segment T into atomic factual claims with a deterministic `gpt-4o-mini` pass at $\tau = 0$. Claims shorter than five tokens are discarded as scaffolding. For each remaining claim c, we issue one search query: the claim text itself when it contains at least eight tokens, otherwise the original question concatenated with c.

Parameter	Value	Role
β	5.0	Saturation scalar in Eq. 1
λ	1.2	Concentration exponent on per-claim sum
Search depth	3	Top-k organic results per claim
Consistency samples K	5	Stochastic samples for P_{int}
Generator temperature	1.0 / 0.0	Audit sampling / segmentation
NLI threshold	0.5	Contradiction cutoff for consistency
High-trust prior	1.0	Gov/edu/Wikipedia/ major news/science
Low-trust prior	0.1	Reddit/Quora/Medium/ Twitter/X
Default prior	0.5	All other domains

Table 1: Key implementation choices for the technical audit.

Retrieved evidence is scored by a coarse but explicit relevance prior. For each claim c, let K_c be the set of rare keywords extracted by matching capitalized tokens (regular expression `\b[A-Z][a-zA-Z0-9-]+\b`) and removing a sentence-leading stoplist. For each result s with URL $u(s)$ and snippet $\sigma(s)$, we compute

$$\text{MatchRatio}(\sigma, K_c) = \frac{|\{k \in K_c : k_\downarrow \in \sigma_\downarrow\}|}{\max(|K_c|, 1)},$$

where $\downarrow$ denotes case-folded substring matching. The source contribution is then

$$w(s, c) = \text{Reputation}(u(s)) \cdot \text{MatchRatio}(\sigma, K_c)^3.$$

`Reputation` is a three-level domain prior: 1.0 for high-trust domains (e.g., `.gov`, `.edu`, `wikipedia.org`, `nih.gov`, `reuters.com`, `apnews.com`, `nature.com`), 0.1 for low-trust domains (e.g., `reddit.com`, `quora.com`, `medium.com`, `twitter.com`), and 0.5 otherwise. The cubic exponent is intentionally aggressive: it suppresses weak keyword overlap that would otherwise permit "keyword stuffing" from superficially related pages.

For reproducibility, the audit uses `gpt-4o-mini` for generation and deterministic segmentation, a DeBERTa-based NLI checker to estimate P_{int} from pairwise agreement across $K = 5$ samples, and Serper as the retrieval backend for evidence collection. The released code and data include the exact prompts, model calls, judge labels, and scripts used in the audit.

3.3 Operationalization: The Oracle Protocol

Equation 1 defines the target signal for a deployed system, but a user study introduces a separate methodological question: does the *visual signal itself* improve truth discernment when its underlying value is correct? To separate that interaction question from retriever noise, the empirical study in Section 5 uses a Wizard-of-Oz (Oracle) protocol.

In this protocol, participants do not see the live output of the auditing pipeline. Instead, the interface displays idealized endpoint values of the signal: grounded summaries are paired

with high-density indicators and fabricated summaries with null indicators. This lets the user study estimate the *interaction effect* of PDI under known-correct signaling, while the technical audit in Section 4 independently evaluates whether the real pipeline can approximate that ideal in practice.

3.4 Design Rationale: Countering Pseudo-Profound Fluency

The visualization of $D(T)$ is designed to counter "pseudo-profound bullshit"—syntactically persuasive but semantically vacuous text [Pennycook *et al.*, 2015]. Rather than asking users to infer truth from fluency, PDI presents evidence density as a high-contrast cue (Figure 2) that can interrupt fast **System 1** judgments and invite more deliberate **System 2** evaluation [Kahneman, 2011].

4 Technical Validation

Before evaluating the interaction paradigm with users, we audited the proposed metric ($D(T)$) to determine which part of the signal actually separates grounded truth from hallucination. Our expectation was that retrieval density would be the workhorse; the audit showed the opposite. Density alone was near chance overall, while the internal Consistency Veto carried most of the discriminative signal on dynamic queries.

Experimental Setup. To evaluate the metric across both established knowledge and emerging information, we constructed a composite dataset ($N = 200$): Static Knowledge ($N = 150$), randomly sampled from the TRUTHFULQA generation task, and Dynamic Knowledge ($N = 50$), a curated FRESHQA subset focused on post-2024 events (e.g., elections, stock prices) designed to probe the "Cold Start" regime where model training weights are outdated. Per-row hallucination labels were produced by `gpt-4o` (temperature 0) using the dataset's official reference answers when available. For the static subset, the judge was given TruthfulQA's official correct / incorrect answer lists; for the dynamic subset, judgement was closed-book. Audit latency measured during this pipeline ranged from 17.0s (Dynamic) to 26.4s (Static), with exact timing logs included in the released audit materials. Static items were slower because TruthfulQA's adversarial prompts elicited longer generations from `gpt-4o-mini` (mean 98 vs. 54 words), which produced more atomic claims and therefore more retrieval and scoring calls.

4.1 Results: Ecological Sensitivity and Hallucination Detection

Ecological Sensitivity (Static vs. Dynamic). As shown in Figure 1, the metric reproduces the established / emerging knowledge separation reported in prior work. Static TruthfulQA items cluster above the high-trust threshold ($M = 0.79$), while dynamic FreshQA items yield a lower mean ($M = 0.64$) and broader spread. This confirms that $D(T)$ tracks epistemic maturity as well as truth, warning users when information is novel and consensus remains unsettled.

Hallucination Detection. To quantify the metric's discriminative power, we labelled every audited generation with a `gpt-4o` judge. Across all 200 generations, the judge marked

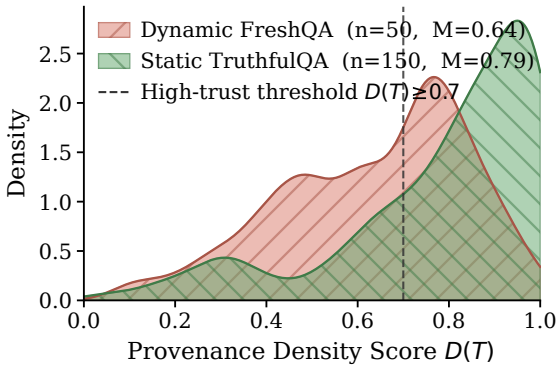

Figure 1: Ecological Sensitivity of the Metric. Distribution of Provenance Density scores $D(T)$ for Static Knowledge (TruthfulQA, green) and Dynamic/Fresh Knowledge (FreshQA, red). Static items cluster above the high-trust threshold ($M = 0.79$), whereas FreshQA items have a lower mean ($M = 0.64$) and broader spread, allowing the metric to signal when information is newer or less settled. The dashed high-trust threshold denotes $D(T) \geq 0.7$.

Split	n	Hall.	Detector	AUC	95% CI	AP
All	200	41	$D(T)$	0.535	[0.43, 0.64]	0.27
All	200	41	Density only	0.470	[0.37, 0.57]	0.21
All	200	41	P_{int} (Veto)	**0.601**	[0.52, 0.69]	0.29
TruthfulQA	150	38	$D(T)$	0.577	[0.46, 0.69]	0.35
TruthfulQA	150	38	Density only	0.537	[0.43, 0.65]	0.31
TruthfulQA	150	38	P_{int} (Veto)	**0.592**	[0.51, 0.67]	0.33
FreshQA	50	3	$D(T)$	0.716	[0.49, 0.98]	0.19
FreshQA	50	3	Density only	0.284	[0.07, 0.63]	0.05
FreshQA	50	3	P_{int} (Veto)	**0.918**	[0.81, 1.00]	0.37

Table 2: Hallucination-detection performance for three detectors derived from Eq. 1. Evidence density alone is near chance overall and anti-discriminative on FreshQA, while the Consistency Veto P_{int} is the strongest discriminator, especially on dynamic queries. FreshQA estimates should be interpreted cautiously because the split contains only three hallucinated items.

$n_H = 41$ as hallucinations ($n_H^{\text{static}} = 38$, $n_H^{\text{dynamic}} = 3$). Table 2 reports ROC-AUC and average precision for three candidate detectors derived from Eq. 1: the full $D(T)$ score, evidence density alone, and the Consistency Veto P_{int}.

Three findings stand out. First, **retrieval density is not the workhorse**. Evidence density alone is near chance overall (AUC $= 0.47$) and anti-discriminative on FreshQA (AUC $= 0.28$), indicating that high-density retrieval can still support false claims when a misconception is popular or a topic is in transition. Second, **the Consistency Veto P_{int} is the principal discriminator on dynamic queries and contributes most of the signal in the combined score**. It achieves the strongest overall AUC (0.60) and reaches AUC $= 0.92$ on dynamic queries, precisely where the model is most likely to be uncertain about post-cutoff claims. However, because the FreshQA split contains only three hallucinated items, these dynamic-query AUC estimates should be interpreted as suggestive rather than definitive. Third, **the combined $D(T)$ score inherits some of the Veto's signal through the $(1 - P_{int})$ gate**. It reaches AUC $= 0.72$ on FreshQA while preserving sensitivity

to epistemic maturity on static knowledge, although it remains weaker than the veto alone because density is noisy on adversarial misconceptions. In this sense, P_{int} is the stronger standalone hallucination detector, whereas $D(T)$ is retained as a composite signal because it adds ecological sensitivity to epistemic maturity that the veto alone does not express.

The large FreshQA gap between density alone and the Veto (AUC 0.28 vs. 0.92) operationalizes the "Consistency Veto" motivated by the *Indonesia-capital transition* case: when queried about Indonesia's capital during the Jakarta–Nusantara transition, retrieval can return high-density evidence for competing time-sensitive answers, but internally unstable generations should not receive high Provenance Density scores.

5 Empirical Evaluation: Interaction Paradigm

Having validated the computational feasibility of the $D(T)$ metric (Section 4), we turn to the critical Human-Centered AI question: Does visualizing this "costly signal" actually enable users to overcome the Fluency Trap?

We conducted a within-subjects controlled experiment to measure Trust Calibration, the alignment between a user's confidence and the factual veracity of the content—under different interface conditions.

	Silk	**Matcha**	**Quantum**
Group 1	Control (Hall.)	Binary (True)	**PDI (True)**
Group 2	Binary (True)	PDI (Hall.)	Control (True)
Group 3	**PDI (True)**	Control (Hall.)	Binary (Hall.)

Table 3: 3×3 Latin Square Design. Participants were counterbalanced across topics and conditions so that PDI was evaluated on both grounded (True) and hallucinated (Hall.) content, reducing topic-specific bias through partial counterbalancing.

Experimental Design. We utilized a 3×3 Latin Square within-subjects design ($N = 81$) (Table 3). The independent variable was the Interaction Paradigm with three levels: Control (The Status Quo): Standard text with no disclosure, representing the current "Pooling Equilibrium" where fluency is the only visible signal. Binary Disclosure: A static "Made with AI" banner. This represents the current industry standard for transparency. Provenance Density (PDI): Our proposed interface (Figure 2), visualizing the density of verification traces (e.g., "5/5 Claims Verified").

Stimuli and Oracle Assignment. The key design goal was to separate *algorithmic error* from *interaction failure*. Accordingly, the user study did not expose participants to live retrieval outputs. Instead, in the PDI condition we used the Oracle protocol introduced in Section 3.3: grounded summaries were shown with high-density indicators ($D(T) \approx 1.0$), and hallucinated summaries were shown with null indicators ($D(T) \approx 0.0$). This makes the user study a test of whether the visualization can override the fluency heuristic when the signal is correct, rather than a test of search quality.

We generated six academic summaries. For the hallucinated items, we wrote high-fluency fabrications (e.g., a fictional "Hua Dynasty" or "Theanine-B" molecule) matched to the grounded summaries in length, tone, and apparent scholarly style. These adversarial stimuli intentionally preserve surface plausibility so that any improvement in ratings can be attributed to the interface rather than to obvious textual defects.

Layout Control. To avoid a layout confound, all conditions used the same fixed-width text column (approximately 520 px). In the Control and Binary conditions, the sidebar region was preserved as an invisible container so that line breaks, whitespace, and reading flow remained constant across conditions.

Procedure and Participants. Participants ($N = 81$) were recruited via Prolific (Fluent English, Approval Rate $> 95\%$). To ensure we measured the "layperson's reliance on fluency," we excluded participants with self-reported expert domain knowledge in the specific topics (History, Chemistry, Physics). Each participant rated the Perceived Veracity of claims on a scale of 0 (Completely Fabricated) to 10 (Completely Factual). Because each participant contributed three ratings, all inferential analyses used a linear mixed model with a random intercept for participant.

5.1 Results: Restoring the Separating Equilibrium

A linear mixed model fit with `lme4` in R and a random intercept for participant (`rating ~ Interface × Veracity + (1 | participant)`) revealed a significant Interface × Veracity interaction (likelihood-ratio test $\chi^2(2) = 31.32$, $p < .001$). This substantial effect confirms that the choice of interaction paradigm fundamentally alters the user's epistemic stance.

Confirming the Fluency Trap (Control): In the absence of signals, participants showed no detectable discrimination. High-Fluency Hallucinations ($M = 6.28$) were rated statistically indistinguishable from Ground Truth ($M = 5.78$; $p = .43$, $d = -0.21$). This pattern is consistent with a collapse of the separating equilibrium: without external scaffolding, participants' ratings treated fluency as statistically indistinguishable from veracity.

Stigmatization vs. Calibration (Binary vs. PDI): While both interventions reduced trust in hallucinations, they did so via distinct cognitive mechanisms (Figure 3): **Binary Disclosure (Stigmatization):** The "Made with AI" label functioned as a crude penalty. It reduced trust in hallucinations ($M = 5.04$) but failed to meaningfully restore confidence in the truth ($M = 6.70$). **Provenance Density (Calibration):** In contrast, PDI produced a large separation between truth and fabrication ratings. It penalized ungrounded content nearly twice as severely as binary labels ($M = 3.93$), while simultaneously elevating trust in verified content to the highest observed level ($M = 8.08$). Post-hoc Tukey HSD comparisons across the Interface × Veracity estimated marginal means confirm that PDI significantly restores confidence in verified content compared to standard Binary Disclosure ($p = .039$), effectively reversing the Transparency Penalty.

The "Discernment Gap": Table 4 summarizes the "Discernment Gap" ($\Delta = Rating_{True} - Rating_{Hall}$). **Control:** Negative Gap (-0.50). Users rated fabrications numerically above ground truth (non-significant; see $p = .43$ above). **Binary:** Moderate Gap ($+1.66$). Users were skeptical of everything. **PDI:** Large Gap ($+4.15$). Users successfully filtered

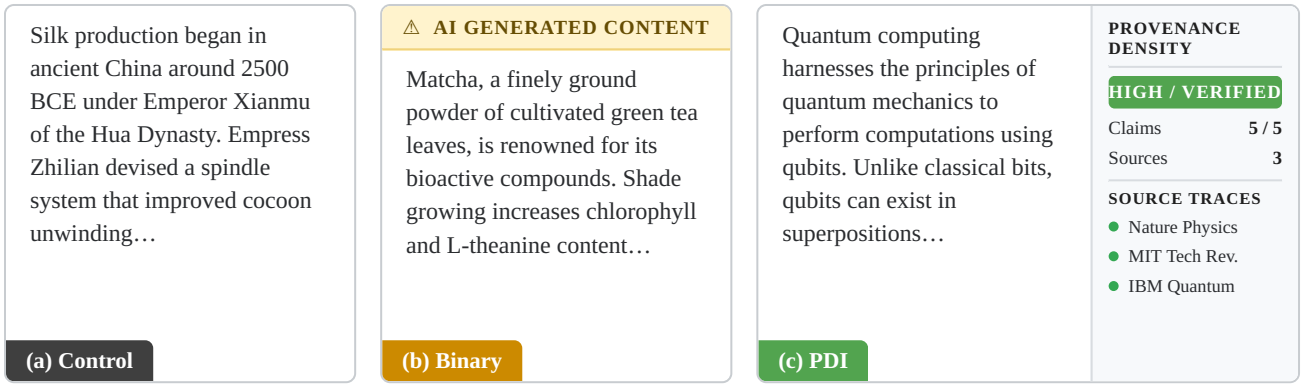

Figure 2: Experimental stimuli (Group 1; schematic). Simplified mock-ups of the three interaction paradigms tested in the user study: (a) Control, shown without an external signal; (b) Binary Disclosure, shown with a static "AI-Generated Content" banner; and (c) Provenance Density (PDI), shown with a density sidebar indicating a high proportion of verified claims ($D(T) \approx 1.0$). Body text is abbreviated for layout clarity; full-fidelity stimulus screens are included in the supplementary release.

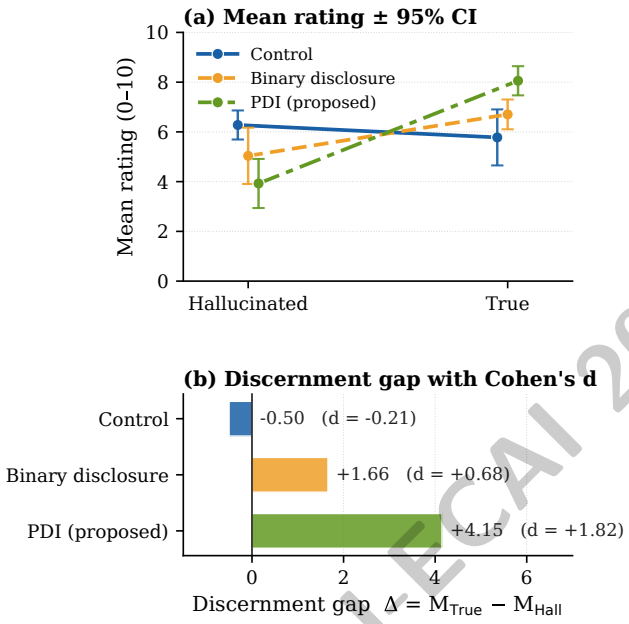

Figure 3: User-study results ($N = 243$ ratings: 81 participants × 3 ratings each). **(a)** Mean rating ± 95% CI shows the Interface × Veracity interaction ($\chi^2(2) = 31.32$, $p < .001$). **(b)** Discernment gap $\Delta = M_{\text{True}} - M_{\text{Hall}}$ with Cohen's d annotated; PDI produces the largest separation ($d = 1.82$).

information. The large effect size for PDI ($d = 1.82$) confirms that visualizing the *density* of evidence allows users to offload the cognitive burden of verification, transforming a high-friction expert task into a low-friction perceptual task.

6 Discussion

Our quantitative results confirm that Provenance Density indicators (PDI) significantly outperform both Control and Binary Disclosure conditions in restoring truth discernment. By triangulating these statistical findings with the technical audit

Interface	Rating (Hall)	Rating (True)	Gap (Δ)	Effect Size (d)
Control	**6.28**	5.78	-0.50	-0.21
Binary	5.04	6.70	+1.66	0.68
PDI	**3.93**	**8.08**	**+4.15**	**1.82**

Table 4: Quantifying Discernment. PDI creates the largest separation ($d = 1.82$) between truth and fabrication.

($N = 200$) and qualitative participant feedback ($N = 81$), we infer the cognitive mechanisms driving these effects.

Mechanism of the Fluency Trap: Heuristic Substitution. The failure of participants to distinguish between truth and hallucination in the Control condition ($p = .43$) validates the psychological framework of "pseudo-profound bullshit" [Pennycook *et al.*, 2015]. In the absence of provenance signals, users substituted *veracity* (is this true?) with *fluency* (does this sound professional?).

Multiple participants explicitly described this strategy. P25 noted, "They sound believable and well written... It didn't seem made up to me," while P81 judged accuracy based on "how the text was worded and flowed." This confirms that when the cost of generating professional-sounding text approaches zero, style ceases to be a reliable proxy for substance, leading to the "Illusion of Truth" effects observed in our data.

The "Warning" Effect of Binary Labels. While prior work suggests that AI labels can lower perceived credibility of labeled content generally [Wittenberg *et al.*, 2025; Li and Yang, 2024], our analysis suggests that Binary Disclosure operated through a blunt mechanism of *epistemic stigmatization*. Users interpreted the label not as a transparency aid, but as a risk marker.

P23 described the interaction vividly: "The AI banner seemed like a warning vs being informative." This empirically confirms the risks identified by Simon et al. [2023], who argued that general warnings might induce a "vigilant skepticism" that reduces credibility across the board rather than enabling specific verification. This broad penalty also aligns with concerns about the normative fairness of mandatory dis-

closure [Hosseini *et al.*, 2025], as it punishes the *use* of the tool rather than the *accuracy* of the content [Cheong *et al.*, 2025].

Cognitive Offloading and the Consistency Veto. In the PDI condition, participant strategies shifted from intrinsic text evaluation to extrinsic evidence evaluation. The high discernment gap (+4.15) was driven by users treating the Provenance Density score as an evidence cue.

Calibrated Reliance: A critical question in HAI is whether such "cognitive offloading" functions as extended cognition [Chirayath *et al.*, 2025; Clark and Chalmers, 1998]. Our technical validation (Section 4) suggests cautious support, but with important limits. The metric's **Consistency Veto**—demonstrated by the suppression of scores in ambiguous queries such as the Indonesia capital transition—provides a partial safety rail for this offloading, not a guarantee. Users like P6 ("One of the passages also had a measurement of claims verified... I was more inclined to believe that") are not blindly trusting the AI; they are trusting a *verified attribute* of the information. However, because high-density misconceptions can still evade the veto, the interface should be understood as improving truth discernment rather than eliminating the "False Assurance" pitfall common in retrieval systems.

Ecological Alignment: The metric's sensitivity to information age (Ecological Sensitivity) aligns user trust with the stability of knowledge. As shown in our technical results, established facts yielded higher density signals ($M = 0.79$) than emerging dynamic topics ($M = 0.64$). PDI therefore functions as a "Time-Aware" signal, implicitly guiding users to be more skeptical of breaking news (FreshQA) than settled science (TruthfulQA), mirroring the actual epistemic reliability of those domains.

7 Limitations and Future Work

From Oracle to Deployment. The Oracle protocol in Section 5 shows that PDI has interaction value when the signal is correct, but deployment depends on how closely the live pipeline approaches that ideal. The technical audit clarifies that this approximation is uneven: the full score is driven primarily by the Consistency Veto, while retrieval density alone can be misleading on dynamic queries. In Table 2, density-only scoring is anti-discriminative on FreshQA (AUC $= 0.28$), whereas the combined $D(T)$ score performs better (AUC $= 0.72$) by suppressing internally inconsistent generations. The user-study result should therefore be interpreted as an upper bound on interface efficacy under correct signaling; a key next step is to study near-miss cases, especially false-positive high-density signals. A second design caveat is that the 3×3 Latin-square assignment did not fully cross interface, veracity, and topic: the PDI–Hallucinated cell was measured on Matcha, whereas the Control–Hallucinated cell averaged over Silk and Matcha. Thus, the observed PDI penalty ($M = 3.93$) should be interpreted as strong evidence for the interface under these stimuli, rather than as a fully topic-independent estimate; a fully crossed replication is a clear next step.

Ecological Conservatism (The Cold Start Problem). Provenance Density is inherently conservative: it privileges established consensus ($M = 0.79$) over emerging novelty ($M = 0.64$). Legitimate new information, such as breaking news or novel scientific hypotheses, may lack the citation network required to achieve high density ($S_c = \emptyset$), risking classification of "the new" as "the unverified." Future iterations should integrate temporal metadata to distinguish low provenance due to fabrication from low provenance due to novelty.

Latency Constraints. The "Consistency Veto" is computationally costly: with audit-measured inference latencies ranging from 17.0s (Dynamic) to 26.4s (Static), $D(T)$ is unsuitable for synchronous, turn-by-turn chat. As noted in Section 4, the higher Static latency reflects longer TruthfulQA generations rather than a swapped label. Future work should explore asynchronous interface patterns, such as "Check in Background" or "Deep Check," that integrate verification latency without blocking interaction. Because latency was not experimentally manipulated in our study, future experiments should also test whether this friction functions as a costly signal that improves discernment, or merely as an engineering constraint.

Adversarial Gaming and Automation Bias. PDI also introduces second-order risks. The cubic MatchRatio term in Section 3.1 is designed to reduce citation-laundering by suppressing weak surface overlap, but the technical audit also surfaces vulnerabilities including domain spoofing, empty-keyword exploitation, SEO-style inflation, and mimetic misconception inheritance from open-web consensus. Hardening the reference implementation will require stricter domain parsing, conservative handling of vague claims, caps on per-domain contribution, and checks for duplicated evidence. The risk of *automation bias* [Cummings, 2017] remains: if users transition from "reading the text" to "scanning for the green bar," future designs should add frictional elements that periodically force manual verification.

8 Conclusion

This paper argues that AI-generated misinformation is a problem of *signaling* as well as *generation*. As fluency becomes cheap, style no longer reliably signals substance: without external scaffolding, users in our study could not distinguish grounded truth from high-fluency hallucination.

Binary "Made with AI" labels address this crisis through identity disclosure, but our findings show that they act as crude warnings, producing a Transparency Penalty in which accurate AI-generated information is discounted alongside fabrication. Provenance Density offers a more calibrated alternative by visualizing verification rather than authorship. In the user study, PDI created a large positive discernment gap (+4.15), restoring users' ability to separate truth from fabrication under correct signaling. The technical audit qualifies this promise by showing that high-density evidence can still support misconceptions.

AI transparency should therefore shift from detecting *who* wrote a text toward visualizing *what* supports it, with interfaces that make verification visible, intuitive, and contestable.

Ethical Statement

This research was conducted in accordance with the ethical standards of the University of Tokyo institutional review process. We recruited 81 participants for the user study via Prolific, all of whom provided informed consent before engaging with the study materials. Prolific submission IDs were collected solely for payment processing, stored separately from response data, and are not retained in the released datasets; no other personally identifying information was collected. All analyses and shared materials use pseudonymous participant codes (e.g., `P_G1_01`); raw survey exports are excluded from the open-science release in accordance with the consent form, while the verbatim consent protocol and the de-identified participant ratings are included. Participants were compensated for their time in accordance with fair labor standards.

Beyond procedural compliance, we acknowledge the broader sociotechnical implications of proposing Provenance Density as a credibility standard. While PDI effectively mitigates hallucinations in digitized high-resource domains, it introduces a risk of *Systemic Authority Bias* [Milgram, 1963; Jackson *et al.*, 2012]. By algorithmically privileging information with a traceable digital lineage, citation-based metrics may inadvertently disenfranchise non-digitized knowledge systems—including indigenous epistemologies, oral histories, and low-resource languages—that lack dense citation networks. There is a specific risk that users may conflate "unverified" (lack of digital evidence) with "untrue" (falsehood). Consequently, we advocate that PDI be deployed strictly as a verification tool for digitized consensus, rather than a universal arbiter of truth, to prevent the exacerbation of existing epistemic exclusions.

Acknowledgments

This work was supported in part by JSPS KAKENHI Grant Numbers 25KK0001 (Fund for the Promotion of Joint International Research (Fostering Joint International Research)) and 25K21241 (Grant-in-Aid for Early-Career Scientists), and by JST Moonshot R&D Grant JPMJMS2012.

References

[Buder and Unfried, 2024] Fabian Buder and Matthias Unfried. Transparency without trust: The impact of consumer skepticism of AI-generated marketing content. *NIM Insights*, 7:36–41, 2024.

[Chen *et al.*, 2025] Shan Chen, Mingye Gao, Kuleen Sasse, Thomas Hartvigsen, Brian Anthony, Lizhou Fan, Hugo Aerts, Jack Gallifant, and Danielle S Bitterman. When helpfulness backfires: Llms and the risk of false medical information due to sycophantic behavior. *npj Digital Medicine*, 8(1):605, 2025.

[Cheong *et al.*, 2025] Inyoung Cheong, Alicia Guo, Mina Lee, Zhehui Liao, Kowe Kadoma, Dongyoung Go, Joseph Chee Chang, Peter Henderson, Mor Naaman, and Amy X Zhang. Penalizing transparency? how ai disclosure and author demographics shape human and ai judgments about writing. *arXiv preprint arXiv:2507.01418*, 2025.

[Chirayath *et al.*, 2025] Ginto Chirayath, K Premamalini, and Jeena Joseph. Cognitive offloading or cognitive overload? how ai alters the mental architecture of coping. *Frontiers in Psychology*, 16:1699320, 2025.

[Clark and Chalmers, 1998] Andy Clark and David Chalmers. The extended mind. *analysis*, 58(1):7–19, 1998.

[Cummings, 2017] Mary L Cummings. Automation bias in intelligent time critical decision support systems. In *Decision making in aviation*, pages 289–294. Routledge, 2017.

[Dechêne *et al.*, 2010] Alice Dechêne, Christoph Stahl, Jochim Hansen, and Michaela Wänke. The truth about the truth: A meta-analytic review of the truth effect. *Personality and Social Psychology Review*, 14(2):238–257, 2010.

[Farquhar *et al.*, 2024] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

[Galdin and Silbert, 2025] Anais Galdin and Jesse Silbert. Making talk cheap: Generative ai and labor market signaling. *arXiv preprint arXiv:2511.08785*, 2025.

[Gintis *et al.*, 2001] Herbert Gintis, Eric Alden Smith, and Samuel Bowles. Costly signaling and cooperation. *Journal of theoretical biology*, 213(1):103–119, 2001.

[Hasher *et al.*, 1977] Lynn Hasher, David Goldstein, and Thomas Toppino. Frequency and the conference of referential validity. *Journal of verbal learning and verbal behavior*, 16(1):107–112, 1977.

[Hertwig *et al.*, 2008] Ralph Hertwig, Stefan M Herzog, Lael J Schooler, and Torsten Reimer. Fluency heuristic: A model of how the mind exploits a by-product of information retrieval. *Journal of Experimental Psychology: Learning, memory, and cognition*, 34(5):1191, 2008.

[Hosseini *et al.*, 2025] Mohammad Hosseini, Bert Gordijn, Gregory E Kaebnick, and Kristi Holmes. Disclosing generative ai use for writing assistance should be voluntary. *Research Ethics*, 21(4):728–735, 2025.

[Jackson *et al.*, 2012] Jonathan Jackson, Ben Bradford, Mike Hough, Andy Myhill, Paul Quinton, and Tom R Tyler. Why do people comply with the law? legitimacy and the influence of legal institutions. *British journal of criminology*, 52(6):1051–1071, 2012.

[Jakesch *et al.*, 2023] Maurice Jakesch, Jeffrey T Hancock, and Mor Naaman. Human heuristics for ai-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11):e2208839120, 2023.

[Kahneman, 2011] Daniel Kahneman. *Thinking, fast and slow*. macmillan, 2011.

[Kaur, 2025] Avneet Kaur. Echoes of agreement: Argument driven sycophancy in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22803–22812, 2025.

[Kay *et al.*, 2024] Jackie Kay, Atoosa Kasirzadeh, and Shakir Mohamed. Epistemic injustice in generative ai. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 684–697, 2024.

[Lewis *et al.*, 2020] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[Li and Yang, 2024] Fan Li and Ya Yang. Impact of artificial intelligence–generated content labels on perceived accuracy, message credibility, and sharing intentions for misinformation: Web-based, randomized, controlled experiment. *JMIR Formative Research*, 8(1):e60024, 2024.

[Liang *et al.*, 2023] Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. Gpt detectors are biased against non-native english writers. *Patterns*, 4(7), 2023.

[Liang *et al.*, 2025] Kaiqu Liang, Haimin Hu, Xuandong Zhao, Dawn Song, Thomas L Griffiths, and Jaime Fernández Fisac. Machine bullshit: Characterizing the emergent disregard for truth in large language models. *arXiv preprint arXiv:2507.07484*, 2025.

[Lin *et al.*, 2022] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252. Association for Computational Linguistics, 2022.

[Marewski and Schooler, 2011] Julian N Marewski and Lael J Schooler. Cognitive niches: An ecological model of strategy selection. *Psychological review*, 118(3):393, 2011.

[Milgram, 1963] Stanley Milgram. Behavioral study of obedience. *The Journal of abnormal and social psychology*, 67(4):371, 1963.

[Nakano *et al.*, 2026] Hiroki Nakano, Jo Takezawa, Fabrice Matulic, Chi-Lan Yang, and Koji Yatani. Understanding reader perception shifts upon disclosure of ai authorship. In *Proceedings of the 31st International Conference on Intelligent User Interfaces*, pages 2131–2146, 2026.

[Pennycook *et al.*, 2015] Gordon Pennycook, James Allan Cheyne, Nathaniel Barr, Derek J Koehler, and Jonathan A Fugelsang. On the reception and detection of pseudo-profound bullshit. *Judgment and Decision making*, 10(6):549–563, 2015.

[Reber and Schwarz, 1999] Rolf Reber and Norbert Schwarz. Effects of perceptual fluency on judgments of truth. *Consciousness and cognition*, 8(3):338–342, 1999.

[Reber and Unkelbach, 2010] Rolf Reber and Christian Unkelbach. The epistemic status of processing fluency as source for judgments of truth. *Review of philosophy and psychology*, 1(4):563–581, 2010.

[Rheu and Cho, 2025] Minjin Rheu and Janghee Cho. The trap of ai literacy: The paradoxical relationships between college students' use of llms, ai literacy, and fact-checking behavior. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2025.

[Sharma *et al.*, 2024] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[Shumailov *et al.*, 2024] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.

[Simon *et al.*, 2023] Felix M Simon, Sacha Altay, and Hugo Mercier. Misinformation reloaded? fears about the impact of generative ai on misinformation are overblown. *Harvard Kennedy School Misinformation Review*, 4(5), 2023.

[Spearing *et al.*, 2025] Emily R Spearing, Constantina I Gile, Amy L Fogwill, Toby Prike, Briony Swire-Thompson, Stephan Lewandowsky, and Ullrich KH Ecker. Countering ai-generated misinformation with pre-emptive source discreditation and debunking. *Royal Society Open Science*, 12(6):242148, 2025.

[Spence, 1978] Michael Spence. Job market signaling. In *Uncertainty in economics*, pages 281–306. Elsevier, 1978.

[Sperber *et al.*, 2010] Dan Sperber, Fabrice Clément, Christophe Heintz, Olivier Mascaro, Hugo Mercier, Gloria Origgi, and Deirdre Wilson. Epistemic vigilance. *Mind & language*, 25(4):359–393, 2010.

[Unkelbach *et al.*, 2011] Christian Unkelbach, Myriam Bayer, Hans Alves, Alex Koch, and Christoph Stahl. Fluency and positivity as possible causes of the truth effect. *Consciousness and cognition*, 20(3):594–602, 2011.

[Vu *et al.*, 2024] Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. FreshLLMs: Refreshing large language models with search engine augmentation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13697–13720, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[Wittenberg *et al.*, 2025] Chloe Wittenberg, Ziv Epstein, Gabrielle Péloquin-Skulski, Adam J Berinsky, and David G Rand. Labeling ai-generated media online. *PNAS nexus*, 4(6):pgaf170, 2025.

[Zahavi, 1975] Amotz Zahavi. Mate selection—a selection for a handicap. *Journal of theoretical Biology*, 53(1):205–214, 1975.