

Are LLMs More Skeptical of Entertainment News?

Huiqian Lai

School of Information Studies, Syracuse University
hlai12@syr.edu

Abstract

Large language models (LLMs) are increasingly used for automated news credibility assessment, yet it remains unclear whether they apply even-handed standards across journalistic genres. We examine whether zero-shot LLMs are more likely to misclassify legitimate entertainment news as fake than legitimate hard news, using a within-dataset design on GossipCop from FakeNewsNet. Across four frontier models, we find a clear but model-specific genre asymmetry: DeepSeek-V3.2 and GPT-5.2 show false-positive-rate gaps of 10.1 and 8.8 percentage points, respectively (both $p < .001$), whereas Claude Opus 4.6 and Gemini 3 Flash show no comparable difference. A style-swap experiment yields only limited and inconsistent changes, suggesting that the asymmetry is not reducible to stylistic register alone. Prompt-based mitigation is likewise possible but not generic: framing the model as an entertainment-news fact-checker reduces false positives for DeepSeek-V3.2 by about 50% without detectable recall loss, but offers little improvement for GPT-5.2. Exploratory qualitative coding further suggests two recurring error patterns in sampled false positives: treating private-life claims as inherently unverifiable and discounting entertainment journalism as an epistemically weaker genre. Taken together, these findings show that aggregate performance metrics can obscure structured false positives within legitimate journalism. We argue that LLM-based credibility assessment may not only evaluate truth claims but also differentially recognize the legitimacy of journalistic genres, and that evaluation should therefore include genre-stratified false-positive analysis alongside overall accuracy.

Introduction

Entertainment journalism often reports on events that are later confirmed by mainstream outlets. For example, TMZ's 2009 report of Michael Jackson's death preceded network confirmation by nearly an hour (The Guardian 2009), and the National Enquirer's coverage of John Edwards's extramarital affair broke a story that the mainstream press initially declined to pursue (CNN 2008). These cases illustrate that entertainment reporting can produce factually accurate information, even when it operates outside traditional journalistic hierarchies.

At the same time, its stylistic conventions—such as a focus on private life and more narrative or emotionally vivid forms of presentation—overlap with surface-level features that stylometric and LLM-based detectors have been shown to rely on when assessing credibility (Rashkin et al. 2017; Potthast et al. 2018; Horne and Adali 2017; Wu, Guo, and Hooi 2023). As large language models (LLMs) are increasingly used to evaluate news veracity at scale, this overlap raises a concrete concern: systems may misclassify legitimate entertainment journalism as unreliable. We use “entertainment news” and “hard news” in the sense established in journalism studies, and discuss this distinction in detail in *Related Work*.

This concern points to a broader, structured risk in AI-mediated credibility assessment. Because automated systems often rely on stylistic and linguistic cues as proxies for truthfulness, they may disproportionately flag content that conforms to the conventions of particular journalistic genres (Rashkin et al. 2017; Horne and Adali 2017). In the case of entertainment news, this creates the possibility of systematic false positives against legitimate reporting. Given the scale and visibility of entertainment news within contemporary media ecosystems, such asymmetries are not merely marginal but potentially consequential (Reinemann et al. 2012; Turner 2016). They can bias platform moderation systems by disproportionately downranking certain types of content (Gillespie 2018; Roberts 2019), reinforce uneven recognition of journalistic legitimacy across genres (Carlson 2015; Tong 2018), and ultimately erode user trust in AI-assisted information systems (Sundar 2020; Flanagan and Metzger 2013). Understanding whether and how such systematic misclassification arises is therefore critical for the responsible deployment of LLM-based credibility assessment (Barocas, Hardt, and Narayanan 2023; Mehrabi et al. 2021).

Prior work on LLM-based misinformation detection has focused primarily on aggregate benchmark performance, prompting strategies, and domain generalization (Chen et al. 2025; Su, Cardie, and Nakov 2024; Gupta et al. 2025). Related studies have also examined topic variation, domain-specific detection settings, and model vulnerability to stylistic perturbations (Francis 2024; Cao et al. 2024; Wu, Guo, and Hooi 2023; Tahmasebi et al. 2026). Yet this literature has not directly asked whether LLMs apply uneven credibility

standards across journalistic genres, particularly by misclassifying legitimate entertainment news as fake. That omission matters because aggregate performance can mask systematic false positives against specific forms of legitimate journalism. We therefore still know little about whether entertainment news is disproportionately penalized, or whether any such pattern is driven by style, prompting, or broader assumptions about credibility.

We therefore ask the following research questions:

RQ1: Are zero-shot LLMs more likely to misclassify legitimate entertainment news as fake, compared to legitimate hard news?

RQ2: If such an asymmetry exists, is it driven primarily by stylistic cues, or by other aspects of how models evaluate credibility?

RQ3: Which prompt-based strategies can reduce these false positives, and what do recurring error patterns reveal about why they occur?

To answer these questions, we combine four steps: a within-dataset comparison, a style-swap test, a prompt-mitigation study, and qualitative error analysis, all based on GossipCop within FakeNewsNet (Shu et al. 2020). This paper makes three contributions. *First*, it identifies a model-specific genre asymmetry in zero-shot veracity classification: some frontier LLMs are substantially more likely to misclassify legitimate entertainment news as fake than legitimate hard news, even within the same dataset. *Second*, it shows that this asymmetry is not reducible to stylistic register alone, as rewriting entertainment articles into a hard-news style produces only limited and inconsistent changes in model judgments. *Third*, it demonstrates that mitigation is possible but not generic: an entertainment-domain expert prompt substantially reduces false positives for DeepSeek-V3.2 without detectable recall loss, but offers little improvement for GPT-5.2. More broadly, the study shows that aggregate performance metrics can obscure structured false positives within legitimate journalism, and that LLM-based credibility assessment may therefore not only evaluate truth claims but also differentially recognize which genres of journalism are presumptively trustworthy, effectively reproducing existing hierarchies of journalistic legitimacy.

Related Work

Hard News, Soft News, and Entertainment News in Journalism Studies

The distinction between hard news and soft news is one of the most enduring analytic categories in journalism studies (Reinemann et al. 2012; Lehman-Wilzig and Seletzky 2010). Hard news traditionally refers to reporting on public-affairs topics—politics, economics, and crime—characterized by high news value and temporal urgency (Tuchman 1978; Reinemann et al. 2012). Soft news, by contrast, foregrounds lifestyle, personal experience, and emotional narrative, and includes entertainment and celebrity reporting as prominent examples (Lehman-Wilzig and Seletzky 2010).

Scholarship has increasingly emphasized that this distinction is multidimensional rather than purely topical. Reine-

mann et al. (2012) synthesize prior work into three key dimensions—topic, focus, and style—arguing that news varies not only in subject matter but also in whether it emphasizes societal or individual relevance and whether it adopts impersonal or narrative-driven forms. Lehman-Wilzig and Seletzky (2010, p. 51) similarly critique the rigidity of the binary and propose an intermediate “general news” category for hybrid items that combine public-affairs content with softer presentation. Empirical work further shows that practitioners’ judgments are shaped by presentational choices and role conceptions as much as by topic (Glogger 2019). Beyond description, the distinction is embedded in a hierarchy of journalistic legitimacy, in which hard news has historically been treated as the prestige form of the profession, while soft and entertainment news are framed as lighter, more subjective, and less consequential (Sjøvaag 2015).

Despite this well-developed conceptual vocabulary, automated misinformation detection research has generally not engaged with it directly. Major surveys organize the field around content, style, propagation, and source credibility (Zhou and Zafarani 2020), and existing datasets typically treat entertainment as one topical domain among others rather than as a genre with distinctive evidentiary and stylistic conventions (Pérez-Rosas et al. 2018; Silva et al. 2021b). This leaves open a question that is central to the present study: whether the genre-level features used in journalism studies to characterize entertainment reporting—such as greater emphasis on individual relevance and more narrative or affective forms of presentation—are systematically interpreted by LLM-based credibility assessment systems as signals of unreliability.

LLM-based Veracity Classification and Prompt Sensitivity

Recent work on LLM-based misinformation detection suggests two related conclusions. First, advanced LLMs can serve as effective veracity classifiers and, in some settings, approach or exceed fine-tuned baselines (Pelrine et al. 2023; Vergho et al. 2024). Second, LLM judgments are not procedurally stable: semantically similar prompts, alternative evaluation formats, and other contextual framing cues can produce meaningful variation in outputs (Sclar et al. 2024; Zhuo et al. 2024; Germani and Spitale 2025; Errica et al. 2025). Taken together, this literature suggests that LLM-based credibility assessment is promising, but still highly sensitive to how the evaluative task is posed (Sclar et al. 2024; Zhuo et al. 2024; Germani and Spitale 2025; Vergho et al. 2024).

On the capability side, Pelrine et al. (2023) show that GPT-4 can outperform RoBERTa-large on LIAR and CT-FAN-22 in zero-shot settings, while exhibiting different failure modes from earlier supervised detectors. Vergho et al. (2024) similarly find that performance varies substantially across models and prompt formulations, with notable instability across GPT-3.5 versions. Hu et al. (2024) add an important qualification: although GPT-3.5 underperforms fine-tuned BERT on direct veracity judgments, it can still provide informative multi-perspective rationales that improve downstream classification when incorporated into hybrid

pipelines. At the same time, survey work positions LLMs as an increasingly important part of the fake-news detection landscape, especially for tasks such as classification, verification, and contextual analysis (Chen and Shu 2024; Papageorgiou et al. 2024). Relatedly, Leite et al. (2025) show that LLM-extracted credibility signals can substantially improve article-level veracity classification, particularly in cross-domain settings, suggesting that LLM-based evaluation can be useful even when direct zero-shot judgments remain limited.

A growing body of research also shows that LLM evaluation is highly prompt-sensitive (Sclar et al. 2024; Zhuo et al. 2024; Germani and Spitale 2025; Errica et al. 2025). Sclar et al. (2024) demonstrate that meaning-preserving changes in prompt formatting can produce large swings in model performance. Zhuo et al. (2024) similarly show that prompt sensitivity varies across tasks and becomes especially salient in subjective evaluation settings. Extending this concern beyond prompt wording alone, Germani and Spitale (2025) find that LLM judgments shift systematically when source attribution is manipulated, indicating that evaluative outputs may depend not only on textual content but also on contextual framing cues surrounding that content.

What remains unclear, however, is whether these systems apply uneven credibility standards across different genres of legitimate journalism. Most existing studies evaluate overall benchmark performance, prompt robustness, or general detection capability (Chen and Shu 2024; Papageorgiou et al. 2024; Errica et al. 2025; Jin et al. 2025). Even when they identify instability or bias, they do not directly test whether LLMs are more likely to misclassify legitimate articles from one journalistic genre as fake than equally legitimate articles from another. This omission matters because strong aggregate performance can obscure systematic false positives on particular forms of real journalism. We therefore still know little about whether zero-shot LLM credibility assessment is equally trustworthy across genres of legitimate news.

Topic Domain Variation and the Neglect of Entertainment News

Prior work suggests that entertainment news is not merely another topical category within fake-news detection, but a distinct detection domain with different linguistic cues, weaker cross-domain transfer, and different verification emphases (Pérez-Rosas et al. 2018; Silva et al. 2021a; Mosallanezhad, Karimi, and Tang 2022; Shrestha and Spezzano 2021). Pérez-Rosas et al. (2018) introduced multi-domain fake news datasets spanning several topics and found clear domain variation in cross-domain evaluation. In their leave-one-domain-out experiments, politics was among the most robust domains, whereas entertainment was among the least generalizable, suggesting that signals learned in one domain do not transfer cleanly to another. Silva et al. (2021a) similarly compared PolitiFact and GossipCop and showed that differences in word usage and propagation patterns contribute to substantial performance degradation in cross-domain fake news detection. Mosallanezhad, Karimi, and Tang (2022) further showed that domain-adaptive methods can improve target-domain performance even with limited

target-domain supervision, confirming that the political-to-entertainment transfer gap is both real and learnable.

At the dataset level, Shu et al. (2020) introduced Fake-NewsNet, which includes PolitiFact and GossipCop as political and entertainment subsets, respectively, and explicitly highlighted the value of studying fake news across different news domains. Shrestha and Spezzano (2021) performed systematic linguistic analysis across PolitiFact, BuzzFeed-News, and GossipCop and found that political and gossip news differ in their textual indicators of falsity. In particular, they show that fake political news contains more religion-related language, whereas gossip news relies more heavily on psychological and affective cues, and that feature usefulness varies across domains. Most directly, Li et al. (2024) evaluated an agentic LLM detection system on both PolitiFact and GossipCop and designed its workflow so that a politics-oriented standing/bias tool is used only when the article is identified as political, indicating that verification priorities may need to vary by domain. Taken together, these studies show that entertainment news has long been present in benchmark design and evaluation, but mostly as a transfer domain or robustness test rather than as a primary object of analysis in its own right.

Together, these findings establish that entertainment news constitutes a genuinely distinct detection domain with its own linguistic norms, verification emphases, and failure modes. Yet, to our knowledge, the literature has rarely asked what this distinctiveness means for the treatment of legitimate entertainment journalism. Prior work uses GossipCop as a benchmark or transfer target, but it does not directly examine whether automated systems, and LLMs in particular, are more likely to distrust legitimate entertainment reporting than other forms of legitimate news. In other words, the missing question is not whether entertainment is a different domain, but whether its genre conventions are systematically misread as signs of falsity.

A plausible explanation, suggested by prior work on domain-specific linguistic cues and domain-sensitive verification workflows, is that automated detectors may over-rely on stylistic and tonal cues associated with entertainment news, treating them as signs of low credibility rather than as routine conventions of the genre.

Stylistic Shortcuts and Their Genre-Conditioned Consequences

A growing body of work suggests that misinformation detectors, whether fine-tuned neural models or prompted LLMs, often rely on linguistic form as a proxy for truthfulness rather than evaluating factual content directly (Potthast et al. 2018; Schuster et al. 2020; Wu, Guo, and Hooi 2023; Wan et al. 2025; Su et al. 2023). Potthast et al. (2018) found that while stylometric features reliably distinguish hyperpartisan from mainstream content ($F1 = 0.78$), style-based fake news classification itself performed poorly ($F1 = 0.46$), suggesting that stylistic separability does not map cleanly onto veracity. Schuster et al. (2020) extended this critique to machine-generated text, showing that stylometry can identify text provenance but cannot distinguish legitimate from misleading uses of language models. Wu, Guo, and Hooi (2023) fur-

ther demonstrated that LLMs can camouflage fake news by rewriting it in the style of reputable outlets, causing up to a 38% decline in F1 score in state-of-the-art detectors and confirming their vulnerability to stylistic surface features. More recently, Wan et al. (2025) introduced a systematic taxonomy of detector shortcuts spanning sentiment, style, topic, and perplexity, and showed that models degrade sharply under shortcut induction and injection, while also noting that real content may adopt subjective styles to increase engagement. A closely related finding comes from Su et al. (2023), who show that fake-news detectors are biased against LLM-generated text: they flag LLM-generated fake news more readily than human-written fake news, and they also misclassify LLM-paraphrased real news as fake, which the authors interpret as evidence of detector reliance on shortcut-like linguistic cues. Related work on LLM-based credibility judgment further suggests that these evaluative systems may rely on lexical associations and statistical priors rather than contextual reasoning, and may confuse linguistic form with epistemic reliability (Loru et al. 2025).

These findings make style a plausible explanation for false positives on legitimate entertainment news, but not yet a sufficient one. Prior work has shown that detectors are vulnerable to stylistic manipulation, yet it has not directly tested whether rewriting legitimate entertainment news into a more conventional hard-news register reduces false positives, nor whether any remaining asymmetry reflects broader genre-level assumptions about credibility. The present study addresses this unresolved question by using a style-swap design to test whether style alone can account for LLM skepticism toward real entertainment news.

Taken together, these three strands of work motivate the present study. Prior research shows that LLM-based veracity assessment is increasingly capable but evaluatively unstable, that entertainment news constitutes a distinct detection domain, and that stylistic shortcuts may distort credibility judgments. What remains untested is whether these dynamics combine to produce systematic false positives on legitimate entertainment journalism, and whether any such asymmetry can be explained by style alone.

Method

Study Overview

We study whether zero-shot large language models (LLMs) treat legitimate entertainment news differently from legitimate hard-news reporting within the same dataset when asked to judge article veracity. Our design comprises three components. First, we compare false-positive rates on two groups of real articles within the same dataset to test for genre-conditioned bias while reducing dataset-level confounds. Second, we run a style-swap experiment in which real entertainment articles are rewritten into a hard-news style while keeping their core claims intact (Toshevska and Gievska 2021; Zhao et al. 2024). Third, we test whether prompt-based mitigation can reduce false positives using an exploratory-to-confirmatory workflow (Sclar et al. 2024). We additionally conduct a qualitative error analysis to complement the quantitative findings. The unit of analysis in

this study is the full news article, reflecting how credibility assessments—particularly in LLM-based settings—are typically conducted in practice. This approach also avoids the need to segment articles into discrete claims, a step that could introduce additional sources of variation (Lazer et al. 2018).

This design is intentionally conservative. The evaluation is conducted on article text alone, without access to external verification signals such as sources, metadata, or retrieval. The primary comparison is conducted within GossipCop, a benchmark in the FakeNewsNet repository, so that the focal article groups are drawn from a shared collection environment rather than from different benchmark datasets (Shu et al. 2020). As such, the design should be interpreted as a diagnostic test of model behavior rather than a fully identified causal estimate of the effect of stylistic features.

Data

We draw our data from FakeNewsNet, a benchmark collection of news articles labeled as real or fake (Shu et al. 2020). Our primary analysis uses the GossipCop portion of FakeNewsNet as a controlled testbed in which articles of different genres coexist under a shared annotation and collection framework. Within GossipCop, we focus on real articles and compare two subsets: *entertainment gossip* and *hard news*.

Operationalizing the two focal genres. Building on the hard/soft news distinction reviewed in Related Work, we operationalize the two focal categories along the topic, focus, and style dimensions identified by Reinemann et al. (2012). An article is assigned to *entertainment gossip* if it centers on celebrity, lifestyle, or personal-life matters and is presented in an emotionally vivid or narrative-driven register (e.g., affective language, dramatized framing, or anecdotal storytelling). An article is assigned to *hard news* if it centers on public-affairs topics (e.g., politics, crime, major institutional events) and is presented in an event-centered, source-attributed register. Illustrative examples are provided in Appendix A.5 (Table 2); detailed coding definitions are reported in Appendix A.1–A.4.

Labeling procedure and validation. Because GossipCop also contains articles that are neither pure entertainment gossip nor hard news—notably opinion commentary and promotional or PR-driven coverage—we use a four-way labeling scheme during the filtering stage so that these non-focal items can be separated out rather than misassigned to one of the two focal categories. Each article is classified into exactly one of four categories: *entertainment gossip*, *hard news*, *opinion editorial*, or *promotional*. We implement this step with DeepSeek-V3.2 (temperature = 0.0), using the article text as input and a single-label JSON output (DeepSeek 2026). To validate the labels, we drew a stratified 10% sample ($n = 400$) and had two human coders independently assign categories using the same coding instructions. Inter-coder agreement was high (Cohen’s $\kappa = 0.86$), and agreement between the consensus human labels and the DeepSeek labels was 93%. Three of the four evaluation models used in the main analysis (GPT-5.2, Claude Opus 4.6, Gemini 3 Flash) did not participate in this labeling step, which further

reduces concerns about label circularity between genre assignment and veracity evaluation, as the labeling model is distinct from the evaluation models and the task of genre classification is orthogonal to veracity prediction.

Evaluation subsets. After filtering, we retain only articles labeled as *entertainment gossip* or *hard news*; articles labeled as opinion or promotional are excluded from the main analysis. Focusing on real articles allows us to directly measure false-positive errors, which are central to our research question. Because all articles originate from GossipCop, the hard-news subset should be understood as a within-dataset reference group rather than a representative sample of institutional journalism. The main analysis uses 1,421 real entertainment-gossip articles and 379 real hard-news articles. The size imbalance between subsets reflects GossipCop’s entertainment-focused construction. This limits the precision of the hard-news false-positive-rate estimate, but does not affect the direction of the comparison, which is the primary quantity of interest. The within-dataset design reduces the cross-dataset confounds that would arise in a direct comparison across different benchmarks. Additional evaluation-set details are reported in Appendix A.6.

Models and Inference Setup

We evaluate four frontier models accessed through their official API identifiers: DeepSeek-V3.2 (DeepSeek 2026), GPT-5.2 (OpenAI 2026), Claude Opus 4.6 (Anthropic 2026), and Gemini 3 Flash (Google 2026). Temperature was fixed at 0.0 in all conditions, and maximum output length was capped at 512 tokens. For each article, the model was asked to return a JSON object containing a binary veracity judgment (*real* or *fake*), a credibility score between 0 and 1, and a one-sentence rationale. The baseline prompt was intentionally neutral and did not provide external evidence, retrieval support, or chain-of-thought instructions, so the setup approximates a zero-shot moderation or credibility-screening scenario; the full prompt wording is reported in Appendix B.1.

Experiment 1: Within-Dataset False-Positive Comparison

The first experiment tests whether LLMs are more likely to produce false positives for legitimate entertainment news than for legitimate hard news. Using the same baseline prompt for all articles, we evaluate the two real-article subsets drawn from GossipCop under identical prompting conditions. For each model, we compute the false-positive rate (FPR) separately for real entertainment articles and real hard-news articles, where FPR is defined as the proportion of real articles predicted as *fake*. We also record the model-generated credibility score for supplementary analysis.

Our main outcome is the difference in false-positive rates between the two groups:

$$\Delta\text{FPR} = \text{FPR}_{\text{entertainment}} - \text{FPR}_{\text{hard-news}}.$$

A positive value indicates that the model is more likely to misclassify legitimate entertainment articles as fake than legitimate hard-news articles. We compare false-positive rates using two-proportion z -tests and report the difference in percentage points together with the associated p -value.

Experiment 2: Style-Swap Test

The second experiment examines whether the observed genre difference can be accounted for by stylistic form alone. We sample 50 real entertainment-gossip articles and use DeepSeek-V3.2 to rewrite each article into a more conventional hard-news style while keeping the core factual content as stable as possible. We then submit both the original article and its rewrite to all four evaluation models using the same baseline prompt.

Rewrite-fidelity checks indicated that the rewrites were adequate for diagnostic use; detailed automatic and manual fidelity results are reported in Appendix A.7. For each model, we report aggregate fake rates before and after rewriting, together with paired correction and degradation rates. This experiment is not intended as a fully identified causal test of style, but as a diagnostic test of whether rewriting alone can systematically reduce false positives.

Experiment 3: Prompt-Based Mitigation

The third experiment tests whether prompting can reduce false positives on entertainment-news articles. We use an exploratory-to-confirmatory design with five prompt variants: a neutral baseline prompt ($P0$), two lightweight heuristic prompts centered on verifiability and claim focus ($P1$ – $P3$), and an expert-role prompt that frames the model as an entertainment-news fact-checker ($P4$). Prompt definitions are summarized in Appendix A.8 (Table 4); the full wording of the baseline and mitigation prompts is provided in Appendix B.1–B.2.

The pilot phase used a pilot sample of 37 articles drawn from the false-positive cases identified for DeepSeek-V3.2 in Experiment 1 (approximately 15% of the full DeepSeek-V3.2 false-positive pool of 249), together with a GPT-5.2 pilot set of 44 articles. These pilot analyses were used only to screen prompt variants and identify the most promising mitigation strategy; they are not treated as confirmatory evidence.

Based on the pilot results, we select $P4$ for confirmatory evaluation. In the confirmatory phase, we re-run all 249 DeepSeek false-positive cases identified in Experiment 1 using only $P0$ and $P4$. Both prompts are re-run within the same evaluation session to reduce the influence of API drift and session-level output instability. Because outputs may vary across sessions, the confirmatory correction rate is defined with respect to the subset of cases that remain false positives under the same-session $P0$ rerun:

$$\text{CorrectionRate} = \frac{N(P0 = \text{fake}, P4 = \text{real})}{N(P0 = \text{fake})}.$$

Under this definition, the denominator is 155 same-session baseline false positives rather than the original false-positive set from Experiment 1.

We also conduct a trade-off analysis on 347 fake entertainment articles to test whether $P4$ reduces false positives at the cost of lower recall on fake articles. We report correction rates with 95% Wilson confidence intervals and use McNemar-style paired tests, with exact binomial variants where appropriate.

Qualitative Error Analysis

To complement the quantitative results, we conduct a qualitative error analysis on 85 articles: 30 DeepSeek-V3.2 cases corrected by P4, 30 DeepSeek-V3.2 cases not corrected by P4, and 25 persistent GPT-5.2 false positives. For each case, we examine the article content, the model’s rationale under the baseline prompt, and, where applicable, its rationale under mitigation prompting.

This analysis is not intended to estimate the prevalence of particular error types. Instead, it is used to interpret recurring forms of entertainment-news skepticism, including cases involving private-life unverifiability and cases in which entertainment journalism appears to be treated as an epistemically weaker genre.

Results

Finding 1: Legitimate entertainment news is more likely to be misclassified as fake, but this pattern is model-specific.

Our main within-dataset comparison shows that legitimate entertainment news is more likely to be misclassified as fake than legitimate hard news, but this pattern is not universal across models. As shown in Figure 1, the effect is concentrated in DeepSeek-V3.2 and GPT-5.2, whereas Claude Opus 4.6 and Gemini 3 Flash show little or no genre difference.

For DeepSeek-V3.2, the false-positive rate (FPR) on real entertainment-gossip articles is 17.5% (249/1,421), compared with 7.4% (28/379) on real hard-news articles, a difference of 10.1 percentage points (95% CI [6.84, 13.43], $p < 0.001$). GPT-5.2 shows a similarly large asymmetry, with an FPR of 20.9% (297/1,421) on real entertainment articles and 12.1% (46/379) on real hard-news articles, a difference of 8.8 percentage points (95% CI [4.85, 12.67], $p < 0.001$). By contrast, Claude Opus 4.6 and Gemini 3 Flash produce very low false-positive rates overall and show no statistically meaningful genre difference.

These null results matter substantively because they indicate that the entertainment-news penalty is not an inevitable property of zero-shot veracity classification, but a model-specific fairness concern in LLM-based credibility assessment. Exact test statistics are reported in Table 5, and supplementary overlap and credibility-score analyses are reported in Appendix C, including Table 6.

Finding 2: Style rewriting produces limited and inconsistent changes, which does not support a strong style-only explanation of the false-positive asymmetry.

The style-swap experiment provides little evidence that writing style alone explains the elevated false-positive rates observed for legitimate entertainment news. Across 50 paired cases, rewriting entertainment-gossip articles into a harder-news register produced only limited and inconsistent changes in model judgments. As shown in Figure 2, the dominant pattern is not systematic correction, but relative stability.

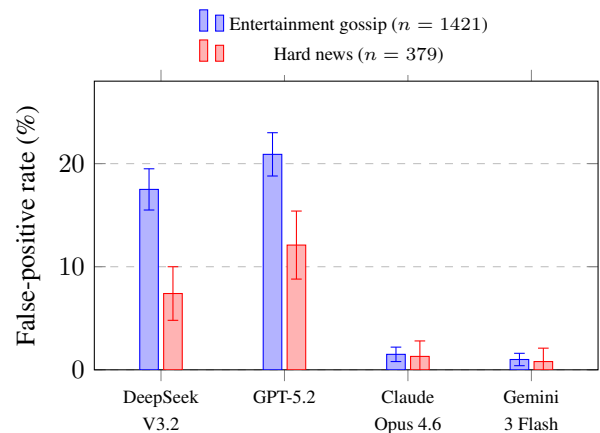


Figure 1: False-positive rates for real entertainment-gossip and real hard-news articles within GossipCop. Error bars indicate approximate 95% confidence intervals. DeepSeek-V3.2 and GPT-5.2 show substantial genre-conditioned false-positive gaps, whereas Claude Opus 4.6 and Gemini 3 Flash do not.

Rewrite-fidelity checks indicated that the rewritten articles were adequate for diagnostic use. The paired prediction results do not support a strong style-only explanation. For DeepSeek-V3.2, the aggregate fake rate remained unchanged at 10.0% before and after rewriting, even though 3 of 5 original false positives were corrected and 3 originally correct judgments degraded. GPT-5.2 shows an even clearer mismatch with a style-only account: its fake rate increased from 22.0% to 32.0% after rewriting, with only 2 of 11 original false positives corrected and 7 of 39 originally correct judgments degrading into false positives.

The other two models again show that the pattern is not universal. Claude Opus 4.6 and Gemini 3 Flash remain low-error overall, and neither exhibits a systematic correction pattern after rewriting. Taken together, these results suggest that stylistic cues may matter at the margin, but they do not provide a sufficient explanation for the higher false-positive rates observed for legitimate entertainment news in Experiment 1. Full raw counts, confidence intervals, rewrite-fidelity checks, and illustrative paired examples are reported in Appendix D, including Table 7.

Finding 3: Prompt-based mitigation substantially reduces false positives for DeepSeek-V3.2, but remains weak for GPT-5.2.

The prompt-based mitigation results show that prompting can reduce false positives on legitimate entertainment news, but the effect is highly model-specific. Across the exploratory comparisons, simple verifiability reminders or claim-focus instructions do not reliably help. The only prompt variant that shows meaningful mitigation potential is P4, the expert-role prompt that frames the model as an entertainment-news fact-checker.

For GPT-5.2, the exploratory pilot provides little evidence that prompting offers a useful mitigation strategy. Across all

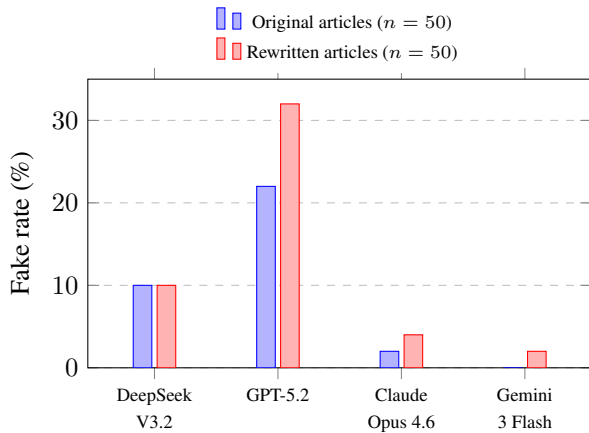


Figure 2: Fake rates before and after rewriting legitimate entertainment articles into a harder-news register. The dominant pattern is not systematic correction. DeepSeek-V3.2 shows no net change, GPT-5.2 becomes more skeptical after rewriting, and Claude Opus 4.6 and Gemini 3 Flash remain low-error overall. See Table 7 for raw counts and 95% confidence intervals.

four non-baseline prompts, fake-to-real flip rates remain low and the small pilot denominator yields wide confidence intervals. Even the best-performing variant, P4, corrects only 3 of 33 baseline false positives, while also introducing 7 real-to-fake reversals. Full GPT-5.2 pilot results are reported in Appendix E, including Table 8.

DeepSeek-V3.2 behaves differently. In exploratory screening, P4 clearly outperforms the other prompt variants and is therefore selected for confirmatory evaluation. The main confirmatory test uses a same-session paired design on all 249 DeepSeek-V3.2 entertainment false positives identified in Experiment 1. This design is necessary because the baseline itself shows meaningful session-level instability even at temperature 0.0: 94 of the 249 articles (37.8%) change their baseline label on rerun. We therefore treat same-session paired evaluation not only as a methodological control, but also as a more appropriate way to assess prompt effects under unstable API-based inference. For this reason, the confirmatory correction rate is defined over the 155 articles that remain false positives under the same-session P0 rerun.

A further trade-off analysis shows that this mitigation effect is not achieved simply by making the model more lenient overall. On a separate sample of 347 fake entertainment articles, DeepSeek-V3.2 attains a baseline fake-news recall of 51.3% under P0 (95% CI [46.0, 56.5]) and 53.9% under P4 (95% CI [48.6, 59.1]), a small positive difference of 2.6 percentage points. This change is not statistically significant ($p = 0.289$), and the overlapping confidence intervals are consistent with the absence of a detectable recall trade-off. In other words, P4 reduces false positives on real entertainment articles without evidence of degraded fake-article detection.

Taken together, these results show that prompt-based miti-

gation is possible, but not generic. The intervention succeeds for DeepSeek-V3.2, but the same family of prompts yields only minimal improvement for GPT-5.2. Prompting is therefore best understood not as a universal fix, but as a model-specific interaction between prompt design and veracity reasoning.

Finding 4: Exploratory coding suggests that many sampled entertainment-news false positives reflect private-life unverifiability and genre-level distrust, but these patterns are not exhaustive.

The qualitative error analysis helps explain why the false-positive asymmetry persists after style rewriting and only partially responds to prompting. We conducted a lightweight exploratory coding of 85 entertainment-news false positives at the model-article level, including 60 cases from DeepSeek-V3.2 and 25 from GPT-5.2. Each case was assigned one primary error pattern after keyword-assisted pre-labeling and manual review. Rationales that did not clearly invoke either private-life unverifiability or genre-level distrust were coded as *Other/unclear*.

Across the coded cases, 65.9% exhibited one or both of the two hypothesized patterns. Private-life unverifiability was the most common single pattern, accounting for 29 of 85 cases (34.1%), followed by genre-level distrust in 21 of 85 cases (24.7%); 6 additional cases (7.1%) showed both patterns. At the same time, 29 of 85 cases (34.1%) remained in the *Other/unclear* category, indicating that these two mechanisms explain a substantial share of false positives but do not exhaust the error space.

Substantively, these patterns help interpret the earlier experiments. Private-life unverifiability helps explain why style rewriting often yields limited gains, especially for stories whose evidentiary status is inherently indirect. Genre-level distrust helps explain why some models continue to reject entertainment reporting even when its presentation is rewritten into a harder-news register, and why prompting can reduce some errors without eliminating the asymmetry. Detailed pattern counts and illustrative cases are reported in Appendix E, including Table 9.

Discussion

Across four frontier models and three complementary experiments, this study shows that zero-shot LLM-based veracity classification can apply uneven credibility standards within legitimate journalism. More specifically, legitimate entertainment news is not judged neutrally by all models: DeepSeek-V3.2 and GPT-5.2 are significantly more likely to misclassify it as fake than legitimate hard news drawn from the same dataset, whereas Claude Opus 4.6 and Gemini 3 Flash show no comparable asymmetry. The paper also clarifies the scope of that asymmetry. It is not reducible to writing style alone, since rewriting entertainment articles into a more conventional hard-news register produces only limited and inconsistent changes, and in GPT-5.2 sometimes increases skepticism. Nor is it uniformly correctable through prompting: an expert-role prompt substantially reduces DeepSeek-V3.2 false positives without de-

Metric	Value	Interpretation
Same-session baseline false positives	155	Confirmatory denominator
P4 corrections (F→R)	77	Corrected by P4
Confirmatory correction rate	49.7% [41.9, 57.5]	77/155; 95% Wilson CI
Paired exact test	$p < 0.001$	Significant improvement
Fake-article sample size	347	Trade-off analysis
P0 fake recall	51.3% [46.0, 56.5]	Baseline recall on fake articles
P4 fake recall	53.9% [48.6, 59.1]	Recall under mitigation prompt
Recall difference	+2.6 pp	No evidence of degradation
Paired test on recall	$p = 0.289$	Not significant

Table 1: DeepSeek-V3.2 confirmatory mitigation and trade-off results for P4. The confirmatory correction rate is defined over the same-session P0 false positives only. Bracketed values indicate 95% Wilson confidence intervals.

grading fake-article recall, but does not generalize to GPT-5.2. Taken together, these results position entertainment-news skepticism as a model-specific failure mode in LLM credibility judgment, rather than an inherent property of zero-shot veracity classification itself.

Empirical Implications

Existing work on LLM-based veracity classification has largely emphasized benchmark performance, prompt robustness, and cross-domain generalization (Papageorgiou et al. 2024; Errica et al. 2025; Leite et al. 2025). Our findings identify a different kind of failure: aggregate veracity performance does not guarantee even-handed credibility judgment across journalistic genres within legitimate news. In our within-corpus comparison of real articles, DeepSeek-V3.2 and GPT-5.2 were 10.1 and 8.8 percentage points more likely, respectively, to misclassify legitimate entertainment news as fake than legitimate hard news, whereas Claude Opus 4.6 and Gemini 3 Flash showed no comparable asymmetry. The empirical contribution, then, is to make visible a structured false-positive risk within legitimate journalism itself, one that conventional benchmark summaries are poorly equipped to capture.

Our experiments also narrow the plausible explanation for that failure mode. Prior work on prompt sensitivity and shortcut learning has shown that LLM judgments can vary under semantically equivalent prompts, contextual framing cues, and stylistic manipulation (Sclar et al. 2024; Zhuo et al. 2024; Germani and Spitale 2025; Wan et al. 2025; Wu, Guo, and Hooi 2023; Su et al. 2023). Our findings extend this line of work in two ways. First, they identify the complementary risk to style-based deception: not only can fake content be rewritten to look more credible, but legitimate journalism can also be misread as fake because its genre conventions resemble the surface features models associate with unreliability. Second, the experiments show that this asymmetry is neither reducible to stylistic register alone nor uniformly correctable through prompting. Rewriting entertainment articles into a more conventional hard-news register produces only limited and inconsistent changes, while expert-role prompting substantially reduces false positives for one model but fails to generalize to another. Taken together, these results suggest that the observed asymmetry re-

flects a deeper and model-specific form of skepticism, rather than prompt or register effects alone.

Media-Theoretical Implications: Reproducing Journalistic Legitimacy Hierarchies

Scholarship on journalistic legitimacy has generally treated legitimacy as the basis on which journalism secures cultural authority to produce valid knowledge of the world (Carlson 2015; Tong 2018; Skovsgaard and Bro 2011). Carlson’s key intervention is to distinguish validity from truth: the issue is not only whether an account corresponds to reality, but whether particular forms and practices of knowing are recognized as legitimate in the first place (Carlson 2015). Tong (2018) likewise emphasizes that journalistic legitimacy is dynamic and must be continually maintained, while Skovsgaard and Bro (2011) show why this problem is especially acute for journalism as a profession whose authority remains comparatively vulnerable and continuously in need of justification.

Our findings extend this literature by showing that journalistic legitimacy may be unevenly allocated not only between journalism and its external challengers, but also across genres within journalism itself. In our results, some LLMs appear to grant legitimate entertainment reporting a weaker presumption of credibility than legitimate hard news, even when both are real articles. The implication is that these systems are not simply making isolated errors of factual classification; they are differentially recognizing what counts as a legitimate object of truth evaluation. Seen this way, the entertainment-news penalty is better understood as a model-specific downgrading of journalistic legitimacy across genres. More broadly, this suggests that LLM-based credibility assessment does not merely evaluate truth claims, but also implicitly ranks the legitimacy of different forms of journalism.

A more specific way to interpret this uneven allocation of credibility is through the hard/soft news hierarchy. The observed pattern—a weaker presumption of truth for legitimate entertainment news—is consistent with a long-standing hierarchy in journalism, in which hard news is treated as the most legitimate and authoritative form, while soft and entertainment news are framed as more emotional, subjective,

and less consequential (North 2016; Sjøvaag 2015; Banjac and Hanusch 2023).

What is new in our findings is not the existence of this hierarchy, but its translation into AI-mediated credibility judgment. This suggests that LLMs are not neutral veracity evaluators, but systems that operationalize and reproduce existing genre hierarchies as epistemic criteria. In our data, this is most clearly reflected in the substantial false-positive gaps (approximately 10 percentage points) that DeepSeek-V3.2 and GPT-5.2 produce between legitimate entertainment and legitimate hard news—a gap that does not appear in Claude Opus 4.6 or Gemini 3 Flash. This cross-model variation is itself diagnostic. Because the asymmetry is not universal across models, it is better understood not as an inherent property of entertainment news itself, but as a model-specific form of credibility sorting that reflects how different systems operationalize implicit assumptions about journalistic legitimacy. In this sense, variation across models is not simply noise, but evidence that these assumptions are contingent rather than fixed.

Limitations

Several limitations bound the scope of these conclusions. The primary analysis is confined to GossipCop within FakeNewsNet, so whether the observed entertainment-news penalty generalizes to other corpora, languages, platforms, or entertainment subgenres remains an open question. Within GossipCop, the hard-news subset ($n = 379$) is much smaller than the entertainment-gossip subset ($n = 1,421$), which limits the precision of hard-news false-positive estimates and means that this reference group should be understood as a within-corpus comparison category rather than a representative sample of institutional journalism more broadly. The evaluations are also text-only and were conducted under a fixed prompting protocol, so the results speak to how these models judge article text in isolation, not to how fuller verification systems might behave when given retrieval support, source metadata, publisher information, or social-context signals. The style-swap experiment is informative as a diagnostic, but it is not a fully identified causal test. Rewriting reduced article length by approximately 40% and changed features beyond register alone, so the findings argue against a strong style-only explanation without isolating style as the sole causal factor. The qualitative error analysis was conducted by a single analyst on a convenience sample of 85 articles, and the coding categories were exploratory rather than theory-derived, so the reported proportions should be treated as suggestive patterns rather than estimates of the full false-positive population. Although all evaluations were conducted at temperature 0.0, API-based outputs were not perfectly stable across sessions: 94 of 249 DeepSeek-V3.2 baseline cases (37.8%) changed labels on rerun. This degree of cross-call variation is itself important for interpreting the prompt-mitigation results, because it indicates that article-level judgments may shift even under nominally fixed inference settings. We therefore used a same-session paired design in Experiment 3 to reduce this source of variation. Even so, the remaining findings should be interpreted as snapshot measurements under a fixed eval-

uation protocol rather than as immutable properties of the underlying models.

Because the four models examined here represent a rapidly evolving technical landscape, both the magnitude of the genre asymmetry and the apparent effectiveness of prompt-based mitigation may shift as models, APIs, and deployment settings change over time.

Conclusion

This study shows that zero-shot LLMs can misrecognize legitimate journalism in genre-specific ways. In a within-dataset comparison on GossipCop, DeepSeek-V3.2 and GPT-5.2 were substantially more likely to label real entertainment news as fake than real hard news, whereas Claude Opus 4.6 and Gemini 3 Flash showed no comparable asymmetry. Across a style-swap test, prompt-based mitigation, and qualitative error analysis, we find that this pattern is not well explained by stylistic register alone. Instead, for some models, legitimate entertainment reporting appears to be granted a weaker presumption of credibility, especially when private-life claims are treated as inherently unverifiable or when entertainment journalism is implicitly discounted as a less trustworthy genre.

The broader implication is methodological as much as substantive. Aggregate benchmark performance can obscure structured false positives within real journalism, meaning that strong average results do not necessarily imply even-handed credibility judgment. For LLM-based misinformation detection, moderation, and news-assistance systems, evaluation should therefore include genre-stratified false-positive analysis rather than relying on overall accuracy alone. More broadly, the challenge is not only whether models can detect fake news, but whether they can recognize different genres of legitimate news as legitimate in the first place—and in doing so, avoid reproducing existing hierarchies of journalistic legitimacy.

References

- Anthropic. 2026. Models Overview. Accessed March 10, 2026.
- Banjac, S.; and Hanusch, F. 2023. The struggle for authority and legitimacy: Lifestyle and political journalists' discursive boundary work. *Journalism*, 24(10): 2155–2173.
- Barocas, S.; Hardt, M.; and Narayanan, A. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.
- Cao, Y.; Nair, A. M.; Eyimife, E.; Soofi, N. J.; Subbalakshmi, K. P.; Wullert II, J. R.; Basu, C.; and Shallcross, D. 2024. Can Large Language Models Detect Misinformation in Scientific News Reporting? *arXiv preprint arXiv:2402.14268*.
- Carlson, M. 2015. Metajournalistic Discourse and the Meanings of Journalism: Definitional Control, Boundary Work, and Legitimation. *Communication Theory*, 26(4): 349–368.
- Chen, C.; and Shu, K. 2024. Combating Misinformation in the Age of LLMs: Opportunities and Challenges. *AI Magazine*, 45(3): 354–368.

- Chen, M.; Wei, L.; Cao, H.; Zhou, W.; and Hu, S. 2025. Explore the Potential of LLMs in Misinformation Detection: An Empirical Study. In *Proceedings of the AAAI 2025 Workshop on Preventing and Detecting LLM Misinformation*.
- CNN. 2008. Edwards Affair: Was Media Part of a ‘Conspiracy of Silence’? <https://edition.cnn.com/2008/POLITICS/08/10/edwards.coverage/index.html>. Accessed: 2026-04-20.
- DeepSeek. 2026. DeepSeek API Docs. Accessed March 10, 2026.
- Errica, F.; Sanvito, D.; Siracusano, G.; and Bifulco, R. 2025. What Did I Do Wrong? Quantifying LLMs’ Sensitivity and Consistency to Prompt Engineering. In Chiruzzo, L.; Ritter, A.; and Wang, L., eds., *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 1543–1558. Albuquerque, New Mexico: Association for Computational Linguistics. ISBN 979-8-89176-189-6.
- Flanagin, A. J.; and Metzger, M. J. 2013. Trusting expert-versus user-generated ratings online: The role of information volume, valence, and consumer characteristics. *Computers in Human Behavior*, 29(4): 1626–1634.
- Francis, E. M. C. 2024. Variation between Credible and Non-Credible News across Topics. In *Proceedings of the 1st International Conference on NLP & AI for Cyber Security*, 86–96.
- Germani, F.; and Spitale, G. 2025. Source framing triggers systematic bias in large language models. *Science Advances*, 11(45): eadz2924.
- Gillespie, T. 2018. *Custodians of the Internet : Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. New Haven, CT: Yale University Press. ISBN 9780300235029.
- Glogger, I. 2019. Soft Spot for Soft News? Influences of Journalistic Role Conceptions on Hard and Soft News Coverage. *Journalism Studies*, 20(16): 2293–2311.
- Google. 2026. Gemini 3 Flash Preview. Accessed March 10, 2026.
- Gupta, A.; Hanley, H.; Lechner, M.; Habibi, J.; et al. 2025. SoK: Machine Learning for Misinformation Detection. *arXiv preprint arXiv:2308.12215*.
- Horne, B.; and Adali, S. 2017. This Just In: Fake News Packs A Lot In Title, Uses Simpler, Repetitive Content in Text Body, More Similar To Satire Than Real News. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1): 759–766.
- Hu, B.; Sheng, Q.; Cao, J.; Shi, Y.; Li, Y.; Wang, D.; and Qi, P. 2024. Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 22105–22113.
- Jin, W.; Gao, Y.; Tao, T.; Wang, X.; Wang, N.; Wu, B.; and Zhao, B. 2025. Veracity-Oriented Context-Aware Large Language Models-Based Prompting Optimization for Fake News Detection. *International Journal of Intelligent Systems*, 2025(1): 5920142.
- Lazer, D. M. J.; Baum, M. A.; Benkler, Y.; Berinsky, A. J.; Greenhill, K. M.; Menczer, F.; Metzger, M. J.; Nyhan, B.; Pennycook, G.; Rothschild, D.; Schudson, M.; Sloman, S. A.; Sunstein, C. R.; Thorson, E. A.; Watts, D. J.; and Zittrain, J. L. 2018. The science of fake news. *Science*, 359(6380): pp. 1094–1096.
- Lehman-Wilzig, S. N.; and Seletzky, M. 2010. Hard news, soft news, ‘general’ news: The necessity and utility of an intermediate classification. *Journalism*, 11(1): 37–56.
- Leite, J. A.; Razuvayevskaya, O.; Bontcheva, K.; and Scarton, C. 2025. Weakly supervised veracity classification with LLM-predicted credibility signals. *EPJ data science*, 14(1): 16–23.
- Li, X.; Jang, Y.; Shi, C.; Sharma, Y.; Bhumbra, A.; and Singh, K. 2024. Large Language Model Agent for Fake News Detection. *arXiv preprint arXiv:2405.01593*.
- Loru, E.; Nudo, J.; Marco, N. D.; Santirocchi, A.; Atzeni, R.; Cinelli, M.; Cestari, V.; Rossi-Arnaud, C.; and Quattrocio, W. 2025. The simulation of judgment in LLMs. *Proceedings of the National Academy of Sciences*, 122(42): e2518443122.
- Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.*, 54(6).
- Mosallanezhad, A.; Karimi, H.; and Tang, J. 2022. Domain Adaptive Fake News Detection via Reinforcement Learning. In *Proceedings of the ACM Web Conference 2022*, 3632–3640.
- North, L. 2016. The Gender of “soft” and “hard” news. *Journalism Studies*, 17(3): 356–373.
- OpenAI. 2026. GPT-5.2 Model. Accessed March 10, 2026.
- Papageorgiou, E.; Chronis, C.; Varlamis, I.; and Himeur, Y. 2024. A Survey on the Use of Large Language Models (LLMs) in Fake News. *Future Internet*, 16(8).
- Pelrine, K.; Mosber, A.; Zheng, J.; Yang, J.-Y.; Peng, A.; Rabbany, R.; and Cheung, J. C. K. 2023. Towards Reliable Misinformation Mitigation: Generalization, Uncertainty, and GPT-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 6053–6068. Association for Computational Linguistics.
- Pérez-Rosas, V.; Kleinberg, B.; Lefevre, A.; and Mihalcea, R. 2018. Automatic Detection of Fake News. In *Proceedings of the 27th International Conference on Computational Linguistics*, 3391–3401.
- Potthast, M.; Kiesel, J.; Reinartz, K.; Bevendorff, J.; and Stein, B. 2018. A Stylometric Inquiry into Hyperpartisan and Fake News. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 231–240.
- Rashkin, H.; Choi, E.; Jang, J. Y.; Volkova, S.; and Choi, Y. 2017. Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking. In Palmer, M.; Hwa, R.; and Riedel, S., eds., *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2931–2937. Copenhagen, Denmark: Association for Computational Linguistics.

- Reinemann, C.; Stanyer, J.; Scherr, S.; and Legnante, G. 2012. Hard and Soft News: A Review of Concepts, Operationalizations and Key Findings. *Journalism*, 13(2): 221–239.
- Roberts, S. T. 2019. *Behind the screen : content moderation in the shadows of social media*. New Haven, CT: Yale University Press. ISBN 0-300-24531-9.
- Schuster, T.; Schuster, R.; Shah, D. J.; and Barzilay, R. 2020. The Limitations of Stylometry for Detecting Machine-Generated Fake News. *Computational Linguistics*, 46(2): 499–510.
- Sciar, M.; Choi, Y.; Tsvetkov, Y.; and Suhr, A. 2024. Quantifying Language Models’ Sensitivity to Spurious Features in Prompt Design or: How I Learned to Start Worrying About Prompt Formatting. In *The Twelfth International Conference on Learning Representations*.
- Shrestha, A.; and Spezzano, F. 2021. Textual Characteristics of News Title and Body to Detect Fake News: A Reproducibility Study. In *Proceedings of the 21st International Conference on Web Engineering*, 261–275.
- Shu, K.; Mahudeswaran, D.; Wang, S.; Lee, D.; and Liu, H. 2020. FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media. *Big Data*, 8(3): 171–188.
- Silva, A.; Han, L.; Karunasekera, S.; and Leckie, C. 2021a. Embracing Domain Differences in Fake News: Cross-domain Fake News Detection using Multi-modal Data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 557–565.
- Silva, A.; Luo, L.; Karunasekera, S.; and Leckie, C. 2021b. Embracing Domain Differences in Fake News: Cross-domain Fake News Detection using Multi-modal Data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(1): 557–565.
- Sjøvaag, H. 2015. Hard news/soft news: The hierarchy of genres and the boundaries of the profession. In Carlson, M.; and Lewis, S. C., eds., *Boundaries of Journalism: Professionalism, Practices and Participation*, 101–117. New York: Routledge. ISBN 9781138017849.
- Skovsgaard, M.; and Bro, P. 2011. PREFERENCE, PRINCIPLE AND PRACTICE. *Journalism Practice*, 5(3): 319–331.
- Su, J.; Cardie, C.; and Nakov, P. 2024. Adapting Fake News Detection to the Era of Large Language Models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, 1473–1490.
- Su, J.; Zhuo, T. Y.; Mansurov, J.; Wang, D.; and Nakov, P. 2023. Fake News Detectors are Biased against Texts Generated by Large Language Models. arXiv:2309.08674.
- Sundar, S. S. 2020. Rise of Machine Agency: A Framework for Studying the Psychology of Human–AI Interaction (HAI). *Journal of Computer-Mediated Communication*, 25(1): 74–88.
- Tahmasebi, S.; et al. 2026. Robust Fake News Detection using Large Language Models under Adversarial Sentiment Attacks. arXiv preprint arXiv:2601.15277.
- The Guardian. 2009. How TMZ broke the news of Michael Jackson’s death. *The Guardian*.
- Tong, J. 2018. Journalistic Legitimacy Revisited. *Digital Journalism*, 6(2): 256–273.
- Toshevskva, M.; and Gievska, S. 2021. A Review of Text Style Transfer using Deep Learning. *IEEE Access*, 9: 128479–128495.
- Tuchman, G. 1978. *Making news : a study in the construction of reality*. New York: Free Press. ISBN 0029329604.
- Turner, G. 2016. *Understanding celebrity*. Los Angeles: SAGE, second edition. ISBN 1-4462-9271-1.
- Vergho, T.; Godbout, J.-F.; Rabbany, R.; and Pelrine, K. 2024. Comparing GPT-4 and Open-Source Language Models in Misinformation Mitigation. arXiv:2401.06920.
- Wan, Y.; Wang, X.; Gao, W.; He, J.; and Huang, M. 2025. Truth over Tricks: Measuring and Mitigating Short-cut Learning in Misinformation Detection. In *Advances in Neural Information Processing Systems*, volume 38.
- Wu, J.; Guo, J.; and Hooi, B. 2023. Fake News in Sheep’s Clothing: Robust Fake News Detection Against LLM-Empowered Style Attacks. arXiv preprint arXiv:2310.10830.
- Zhao, J.; Guan, Z.; Xu, C.; Zhao, W.; and Jiang, Y. 2024. SC2: Towards Enhancing Content Preservation and Style Consistency in Long Text Style Transfer. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9890–9902. Bangkok, Thailand: Association for Computational Linguistics.
- Zhou, X.; and Zafarani, R. 2020. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Comput. Surv.*, 53(5).
- Zhuo, J.; Wen, S.; He, X.; Huang, X.; Li, Z.; Liu, Y.; and Qiu, X. 2024. ProSA: Assessing and Understanding the Prompt Sensitivity of LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Appendix A: Supplementary Materials for Method

A.1 LLM-Based Genre Classification

To construct the within-GossipCop comparison subsets used in the main analysis, we classified articles into four genre categories using DeepSeek-V3.2 with temperature set to 0.0. This step was used only to assign genre labels for dataset construction. It was not used to generate article text.

For each article, we provided the article text to the model and asked it to assign exactly one of four labels: `HARD_NEWS`, `ENTERTAINMENT_GOSSIP`, `OPINION_EDITORIAL`, or `PROMOTIONAL`. The model was instructed to return its output as a JSON object containing a genre label and a confidence score.

A.2 Classification Prompt

We used the following prompt:

```
Classify this news article into ONE of four genres:
1. HARD_NEWS - Factual, neutral tone; who/what/when/where/why structure; minimal
evaluative language
2. ENTERTAINMENT_GOSSIP - Exaggerated, emotionally vivid, hyperbolic; celebrity/personal
focus; designed to entertain
3. OPINION_EDITORIAL - Explicit subjective stance; passionate critique; persuasive intent
4. PROMOTIONAL - Reports on marketing/PR events; promotional language; brand
partnerships
Article: {text}
Respond with ONLY a JSON object: {"genre": "...", "confidence": 0.0-1.0}
```

A.3 Coding Instructions

The four genre categories were defined as follows, along with illustrative examples:

- **ENTERTAINMENT_GOSSIP**: Celebrity news, relationship updates, lifestyle coverage, and entertainment industry gossip. These articles typically foreground celebrity or personal matters and often use emotionally vivid, sensational, or hyperbolic language.
Example: “Jennifer Aniston FINALLY breaks silence on Brad Pitt reunion rumors!”
- **HARD_NEWS**: Articles oriented toward public-affairs topics such as politics, economics, science, crime, or major current events, and presented in a relatively more informational and event-centered register than celebrity gossip, opinion, or promotional content.
Example: “Federal Reserve raises interest rates by 0.25 percentage points.”
- **OPINION_EDITORIAL**: Opinion pieces, commentary, reviews, or analysis with an explicit subjective stance, evaluative framing, or persuasive intent.
Example: “Why Hollywood’s obsession with reboots is killing creativity.”
- **PROMOTIONAL**: Product announcements, sponsored content, press releases, or brand-partnership coverage characterized by promotional language or commercial intent.
Example: “Kim Kardashian launches new SKIMS collection in collaboration with Fendi.”

A.4 Validation Procedure

To assess the quality of the LLM-generated genre labels, we manually validated a 10% sample ($n = 400$) of the classified articles. The validation sample was stratified by veracity label and included 200 real and 200 fake articles randomly sampled from the classified dataset.

Two human annotators independently labeled each article using the coding instructions above and without access to the model predictions. Inter-annotator reliability between the two human coders was high (Cohen’s $\kappa = 0.86$). Disagreements were resolved through discussion to reach consensus. Agreement between the consensus human labels and the DeepSeek-generated labels was 93% (372/400).

The main analysis retains only articles labeled as `ENTERTAINMENT_GOSSIP` or `HARD_NEWS`.

A.5 Illustrative Examples of Focal Genre Categories

Genre	Example headline	Why it fits the label
ENTERTAINMENT_GOSSIP	Katy Perry and Orlando Bloom Spotted Vacationing Together in the Maldives	Celebrity relationship coverage with soft, lifestyle-oriented gossip framing
ENTERTAINMENT_GOSSIP	Brooke Burke Files for Divorce from David Charvet After Six Years	Celebrity private-life event framed in a tabloid-style register
ENTERTAINMENT_GOSSIP	Meghan Markle and Prince Harry to Spend Night Apart Before Royal Wedding	Royal-wedding lifestyle detail framed as soft entertainment news
HARD_NEWS	Oscars Debut New Rules To Avoid Another Envelope Mix-Up	Institutionally grounded coverage focused on industry rules and procedural reform
HARD_NEWS	Toxicology Report Reveals High Fentanyl Concentration in Prince’s Body	Investigative coverage centered on official forensic findings rather than gossip
HARD_NEWS	Rose McGowan Faces Felony Arrest Warrant for Controlled Substance	Legal-process reporting focused on law-enforcement action and judicial procedure

Table 2: Illustrative examples of the two focal genre categories used in the main analysis.

A.6 Evaluation Sets

Component	Article set	<i>n</i>	Use
Exp. 1 (main)	GossipCop real entertainment gossip	1,421	Main false-positive comparison
Exp. 1 (main)	GossipCop real hard news	379	Main false-positive comparison
Exp. 2	Real entertainment articles selected for style swap	50	Diagnostic style test
Exp. 3 pilot	DeepSeek pilot subset of baseline false positives	~15% of pool	Prompt screening
Exp. 3 pilot	GPT pilot articles	44	Prompt screening
Exp. 3 conf.	DeepSeek false-positive pool from Exp. 1	249	P0 vs. P4 evaluation
Exp. 3 recall	Entertainment fake articles for trade-off analysis	347	Recall trade-off
Qual. analysis	DeepSeek flipped, DeepSeek non-flipped, and GPT persistent false positives	85	Error/rationale analysis

Table 3: Evaluation sets used in the study. The main comparison is within GossipCop.

FP = false positive. In the confirmatory DeepSeek mitigation analysis, both P0 and P4 were re-run in the same session. The main correction-rate denominator is therefore the set of articles that remained false positives under the same-session P0 baseline, rather than the cached baseline from Experiment 1.

A.7 Rewrite-Fidelity Checks for Experiment 2

Because Experiment 2 relies on rewritten versions of legitimate entertainment articles, we conducted both automatic and manual fidelity checks. The automatic checks yielded a mean TF-IDF cosine similarity of 0.5682 and a mean named-entity retention score of 0.5249 between each original article and its rewrite. The average article length decreased from 230.8 words in the original texts to 138.8 words after rewriting.

We also manually reviewed 10 rewritten articles. In 9 of the 10 cases, the rewrite preserved the key factual content sufficiently for diagnostic comparison. One case showed partial simplification of the original content, reinforcing the interpretation of Experiment 2 as a diagnostic test rather than a fully identified causal test of style alone.

A.8 Prompt Strategy for Experiment 3

Prompt	Idea	Stage	Purpose
P0	Neutral zero-shot prompt	Pilot + conf.	Baseline reference
P1	Verifiability reminder	Pilot only	Encourage separation of low visibility from falsity
P2	Claim-focus prompt	Pilot only	Shift attention from genre tone to factual claims
P3	Combined prompt	Pilot only	Test whether combining P1 and P2 improves performance
P4	Expert entertainment fact-checker role	Pilot + conf.	Reframe the task with domain-specific expertise

Table 4: Prompt strategy design for Experiment 3.

Note. P1–P3 were tested in the pilot stage only. The full DeepSeek confirmatory run compares P0 and P4 only. GPT results are exploratory and remain pilot-only.

B: Veracity Detection Prompts

B.1 Baseline Zero-Shot Veracity Prompt

The baseline zero-shot prompt (P0) used for veracity detection across all models was:

Please read the following news article and determine whether it is real or fake.

News text:
{text}

Respond with ONLY:
{"result": "real" or "fake",
 "credibility": 0.0–1.0,
 "rationale": "one sentence"}

B.2 Mitigation Prompts

Building on the baseline zero-shot veracity prompt reported in Appendix B.1, we tested four prompt interventions designed to reduce genre-induced false positives on entertainment-news articles. Each variant preserves the same output format as P0 while modifying the framing of the evaluation task.

P1: Verifiability Reminder

Please read the following news article and determine whether it is real or fake.

Important: Entertainment and celebrity events are often verifiable through public records, official statements, social media, and credible news sources. Do not treat a topic as unverifiable simply because it concerns personal or celebrity matters.

News text:
{text}

Respond with ONLY:
{"result": "real" or "fake",
 "credibility": 0.0–1.0,
 "rationale": "one sentence"}

Design rationale: This prompt addresses the possibility that LLMs may implicitly treat entertainment topics as inherently less verifiable than other news topics.

P2: Explicit Claim Focus

Please read the following news article and determine whether it is real or fake.

First, identify the specific, verifiable factual claims made in the article (e.g., names, dates, events, and statements). Then assess only whether those claims are accurate.

News text:
{text}

Respond with ONLY:
{"result": "real" or "fake",
 "credibility": 0.0-1.0,
 "rationale": "one sentence"}

Design rationale: This prompt encourages claim-level evaluation rather than a holistic judgment based on genre cues or writing style.

P3: Combined (P1 + P2)

Please read the following news article and determine whether it is real or fake.

Important: Entertainment and celebrity events are often verifiable through public records, official statements, social media, and credible news sources. Do not treat a topic as unverifiable simply because it concerns personal or celebrity matters.

First, identify the specific, verifiable factual claims made in the article (e.g., names, dates, events, and statements). Then assess only whether those claims are accurate.

News text:
{text}

Respond with ONLY:
{"result": "real" or "fake",
 "credibility": 0.0-1.0,
 "rationale": "one sentence"}

Design rationale: This prompt tests whether combining a verifiability reminder with explicit claim extraction produces stronger mitigation than either intervention alone.

P4: Expert Role (Entertainment Fact-Checker)

You are a fact-checker with expertise in entertainment journalism. Assess the factual accuracy of the following article based solely on its specific claims.

News text:
{text}

Respond with ONLY:
{"result": "real" or "fake",
 "credibility": 0.0-1.0,
 "rationale": "one sentence"}

Design rationale: This role-based prompt frames the model as a domain expert and tests whether an entertainment-specific fact-checking perspective reduces stylistic suspicion and improves veracity judgments.

C: Supplementary Statistics for Finding 1

Table 5 reports the full within-dataset statistics for Finding 1. The main result is that DeepSeek-V3.2 and GPT-5.2 show substantial false-positive asymmetries between real entertainment-gossip articles and real hard-news articles within GossipCop, whereas Claude Opus 4.6 and Gemini 3 Flash do not.

Model	Ent.	Hard	Δ (pp)	95% CI	z	p
DeepSeek-V3.2	17.5%	7.4%	+10.1	[6.84, 13.43]	4.86	< .001
GPT-5.2	20.9%	12.1%	+8.8	[4.85, 12.67]	3.86	< .001
Claude Opus 4.6	1.5%	1.3%	+0.2	[-1.15, 1.47]	0.23	.82
Gemini 3 Flash	1.0%	0.8%	+0.2	[-0.84, 1.22]	0.35	.73

Table 5: False-positive rates on two real-article subsets from GossipCop: 1,421 entertainment-gossip articles and 379 hard-news articles. Δ indicates the entertainment-minus-hard-news difference in false-positive rate, reported in percentage points, with 95% confidence intervals.

We also examined whether the two affected models failed on the same entertainment articles. DeepSeek-V3.2 produced 249 entertainment false positives and GPT-5.2 produced 297, with an overlap of 135 articles, corresponding to a Jaccard similarity of 0.33. This partial overlap suggests that the entertainment-news asymmetry is shared across the two affected models, while the exact boundary of suspicious entertainment content remains model-dependent.

A supplementary analysis of continuous credibility scores showed the same directional pattern. On real articles, DeepSeek-V3.2 assigned slightly lower average credibility to entertainment than to hard news (0.940 vs. 0.951), and GPT-5.2 showed a larger gap (0.785 vs. 0.848), as shown in Table 6. These supplementary results provide convergent support for the main finding that some models are systematically more skeptical of legitimate entertainment news than of legitimate hard news within the same dataset context.

Model	Entertainment	Hard news
DeepSeek-V3.2	0.940	0.951
GPT-5.2	0.785	0.848

Table 6: Average credibility scores on real entertainment-gossip and real hard-news articles for the two models that show a substantial false-positive asymmetry in the main analysis.

D: Supplementary Materials for Finding 2

Table 7 reports the full style-swap results for 50 paired real entertainment articles.

Model	Original fake rate	Rewrite fake rate	Correction rate	Degradation rate
DeepSeek-V3.2	10.0% (5/50) [4.3, 21.4]	10.0% (5/50) [4.3, 21.4]	60.0% (3/5) [23.1, 88.2]	6.7% (3/45) [2.3, 17.9]
GPT-5.2	22.0% (11/50) [12.7, 35.3]	32.0% (16/50) [20.8, 45.8]	18.2% (2/11) [5.2, 47.7]	17.9% (7/39) [8.9, 32.6]
Claude Opus 4.6	2.0% (1/50) [0.4, 10.5]	4.0% (2/50) [1.1, 13.5]	100.0% (1/1) [20.7, 100.0]	4.1% (2/49) [1.1, 13.7]
Gemini 3 Flash	0.0% (0/50) [0.0, 7.1]	2.0% (1/50) [0.4, 10.5]	N/A (0/0)	2.0% (1/50) [0.4, 10.5]

Table 7: Style-swap results for 50 paired real entertainment articles. Correction rate is the proportion of original false positives that flipped from *fake* to *real* after rewriting. Degradation rate is the proportion of original correct predictions that flipped from *real* to *fake*. Values are reported as percentages with raw counts in parentheses; bracketed values indicate 95% confidence intervals. Because some correction-rate denominators are very small, several interval estimates are necessarily wide.

We also evaluated whether the rewrites were sufficiently faithful for diagnostic comparison. On average, article length decreased from 230.8 words to 138.8 words, corresponding to a 39.9% reduction. The mean TF-IDF cosine similarity between each original article and its rewrite was 0.568, and mean named-entity retention was 0.525. In a manual review of 10 sampled pairs, 9 rewrites were judged to preserve the key factual content sufficiently for analysis. One partial case involved a fashion commentary article with minimal verifiable factual content, reinforcing the interpretation of Experiment 2 as a diagnostic test rather than a fully identified causal test.

Two illustrative paired cases help clarify the aggregate pattern. In *gossipcop-907697*, a report about legal filings involving the Kardashian family, GPT-5.2 labeled both the original entertainment-style version and the harder-news rewrite as *fake*. By contrast, in *gossipcop-855820*, DeepSeek-V3.2 changed from *fake* on the original article to *real* on the rewritten version. These examples are consistent with the broader quantitative result: rewriting can occasionally shift predictions, but it does not reliably eliminate the underlying asymmetry.

E: Supplementary Materials for Finding 3

Table 8 reports the exploratory GPT-5.2 pilot results across four mitigation prompts. Across all four non-baseline prompts, fake-to-real flip rates remain low and the net error profile does not improve.

Prompt	n	F→R flips	Flip rate (95% CI)	Notes
P1	33	2	6.1% [1.7, 19.6]	8 R→F; net worse
P2	33	0	0.0% [0.0, 10.4]	3 R→F; net worse
P3	33	1	3.0% [0.5, 15.3]	7 R→F; net worse
P4	33	3	9.1% [3.2, 23.6]	7 R→F; net worse

Table 8: Exploratory GPT-5.2 pilot results across four mitigation prompts. Flip rates are calculated over baseline false positives in the pilot set. Bracketed values indicate 95% Wilson confidence intervals.

E: Supplementary Materials for Finding 4

Table 9 reports the distribution of exploratory error-pattern labels across 85 entertainment-news false positives. The coding was conducted at the model–article level on a convenience sample, including 60 cases from DeepSeek-V3.2 and 25 from GPT-5.2. Each case was assigned one primary label after keyword-assisted pre-labeling and manual case-by-case review: *private-life unverifiability*, *genre-level distrust*, *both*, or *other/unclear*. Rationales that did not clearly invoke either of the two focal mechanisms were coded as *Other/unclear*.

Error Pattern	DeepSeek-V3.2		GPT-5.2	
	n	%	n	%
Private-life unverifiability	20	33.3%	9	36.0%
Genre-level distrust	16	26.7%	5	20.0%
Both patterns	5	8.3%	1	4.0%
Other/unclear	19	31.7%	10	40.0%
Total	60	100%	25	100%

Table 9: Distribution of coded error patterns in entertainment-news false positives ($N = 85$). Labels were assigned via keyword-assisted pre-labeling followed by manual case-by-case review.

The first recurring pattern, private-life unverifiability, is most visible in stories involving romantic relationships, family tensions, injuries, health, or other off-stage personal events. In these cases, the model often treats claims as suspicious because they are difficult to verify through public institutional evidence, even when such claims are commonly reported through interviews, representatives, court filings, televised appearances, or entertainment media outlets.

The second recurring pattern, genre-level distrust, is more structural. In these cases, the model appears to treat entertainment journalism itself as an epistemically weak domain. This pattern is especially visible when a change in stylistic presentation fails to alter the judgment.

Two illustrative paired cases help clarify the aggregate pattern. In *gossipcop-907697*, a report about legal filings involving the Kardashian family, GPT-5.2 labeled both the original entertainment-style version and the harder-news rewrite as *fake*. The rewrite removed much of the entertainment framing and restated the story in a more neutral reportorial style, yet the model’s judgment did not change. By contrast, in *gossipcop-855820*, DeepSeek-V3.2 changed from *fake* on the original entertainment-style article to *real* on the rewritten version. Together, these cases suggest that stylistic cues can occasionally shift decisions, but that some errors are tied to deeper assumptions about the verifiability of private-life claims or the epistemic status of entertainment reporting.

These patterns also help interpret the asymmetric prompt results in Experiment 3. The P4 mitigation prompt appears most effective when the error is driven by a relatively shallow heuristic and the model can be redirected toward evaluating a concrete claim. Prompting is less effective when the story concerns intimate personal life, ambiguous relationship dynamics, or events whose evidentiary status is intrinsically indirect. In such cases, the model’s skepticism appears to stem not only from genre cues, but from a deeper mismatch between entertainment reporting and the model’s implicit assumptions about reliable evidence.