Causal Factors of Trustworthiness Judgements in LLMs

Kiko Trevino UCLA kikotrevino@ucla.edu
Marie Yang UCLA marieyang@ucla.edu
Paul Talma UCLA paultalma@ucla.edu
Ramnath Kumar UCLA ramnathk@ucla.edu

Abstract

Large Language Models (LLMs) are increasingly used to evaluate the reliability of news content, yet the mechanisms driving these judgments remain poorly understood. In this work, we investigate whether LLMs assess news trustworthiness based on substantive journalistic quality or superficial linguistic signals. We deploy a multi-stage pipeline combining an LLM-as-a-Judge classification baseline, causal feature perturbation via controlled counterfactual rewrites, and mechanistic attribution via Integrated Gradients. We find that models rely heavily on institutional-register tokens as proxies for trustworthiness, a shallow heuristic that mirrors wire-service style rather than factual accuracy.

1 Introduction

Large Language Models (LLMs) increasingly mediate our interactions with online content. Tools like Perplexity, SearchGPT, and Gemini synthesize vast amounts of data into singular authoritative summaries, increasingly serving as the "front page" of the internet. While these tools offer unprecedented efficiency, they also introduce a new layer of algorithmic gatekeeping. The criteria by which these models judge sources as "trustworthy," "reliable," or "neutral" remain largely opaque.

This lack of transparency raises at least two significant concerns. First, it makes it difficult to notice and correct biases and blind spots in LLM judgments: such biases could pollute the information chain, contributing to political polarization and degrading the broader information ecosystem. Second, if a model's perception of trustworthiness is influenced by superficial factors such as emotional tone or certain keywords, then its judgments may be easily manipulable. Understanding the mechanisms by which LLMs evaluate sources of information is essential to identifying and combating such vulnerabilities.

We address three specific problems in current LLM-based content curation:

- **Causal Ambiguity:** Does a model judge an article based on its substantive content, or is it swayed by a few "suspect tokens," stylistic markers, or political signaling?
- **Mechanistic Opacity:** Even in open-source models, intermediate attribution signals that lead to a reliability score are rarely inspected or reported.
- **Transferability:** Do different model families share a common "latent space" for trustworthiness, or is the definition of a trustworthy source model-dependent?

Our contributions are as follows: (1) a reproducible LLM-as-a-Judge evaluation across seven open-source models on a fake news detection benchmark; (2) token-level mechanistic attribution via Integrated Gradients, revealing specific linguistic markers behind predictions; and (3) a counterfactual intervention study demonstrating that purely stylistic rewrites systematically flip model predictions. All code can be found at `https://github.com/paul-talma/fake_news_interpretability`.

2 Data

We use the **FakeNewsDetection** dataset as our primary source of ground truth (Jikadara, 2023). It is split into `fake.csv` and `real.csv` and contains the title, text, subject, and date of various news articles. We operationalize trustworthiness as a binary label: real articles are labeled 1 (trustworthy) and fake articles are labeled 0 (untrustworthy). We sample 100 balanced articles (50 real, 50 fake) for our main evaluation experiments.

3 Methodology

We design a multi-stage pipeline (Figure 1) to analyze how linguistic features causally influence LLM trustworthiness judgments. The system combines LLM-as-a-Judge evaluation, controlled text interventions, and cross-model analysis.

3.1 LLM-as-a-Judge Baseline

We prompt a suite of models to evaluate these articles. We prompt a suite of open-source models to classify each article as trustworthy (1) or untrustworthy (0). Our final prompt design is:

> *You are a news trustworthiness classifier. Classify an ARTICLE as: 1 = trustworthy (neutral, fact-focused reporting with attribution; wire-service style; avoids conspiratorial framing) or 0 = untrustworthy (opinion/propaganda, sensational or conspiratorial claims without evidence, ad hominem insults, heavy partisan framing, or manipulative rhetoric). Do NOT judge by whether the topic is controversial. Judge by reporting quality and style. If it reads like a blog/opinion piece or uses loaded language, it is likely 0. If it reads like Reuters/AP-style reporting with named sources and neutral tone, it is likely 1. Output exactly one token: 0 or 1.*

To improve instruction adherence, we add $k = 5$ labeled few-shot examples covering both trustworthy and untrustworthy articles. Constraining outputs to 0,1 reduces format errors compared to string labels.

3.2 Causal Interventions

To test whether surface-level linguistic features causally influence model judgments, we apply controlled counterfactual rewrites to articles that were initially classified as untrustworthy. A second LLM (GPT-4o) is instructed to rephrase each article while preserving its core factual content and named entities, modifying only stylistic and rhetorical dimensions. We track six intervention axes per article: *hedging* (epistemic qualifiers added or removed), *citations* (named source attributions), *inflammatory language* (loaded or emotionally charged words), *exclamations*, *rhetorical questions*, and *all-caps tokens*. We also record the change in total word count as a proxy for the extent of the rewrite.

Each intervention is designed to shift the article from one of the following rhetorical modes:

- **Accusation → Attribution:** Direct accusations are replaced by attributed claims (e.g., "Critics argue that…").
- **Loaded language → Neutral vocabulary:** Emotionally charged terms are replaced by standard institutional terminology (e.g., "Obama regime" → "Obama administration").
- **Rhetorical questions → Analytical statements:** Rhetorical questions implying known answers are replaced by declarative analytical statements.
- **Advocacy → Reporting:** First-person or partisan framing is replaced by detached, third-person institutional register.

3.3 Mechanistic Attribution via Integrated Gradients

To understand which specific tokens drive trustworthiness predictions, we apply Integrated Gradients (Sundararajan et al., 2017, IG) to the Qwen2-1.5B-Instruct model -our best-performing judge. IG computes the attribution of each input token to the output prediction by integrating the model's gradients along a straight path from a baseline input (PAD tokens) to the actual input. For each article, we use 300 integration steps and record the attribution score assigned to each token.

We note a completeness gap of 0.596 in our IG computation, meaning the attributions do not fully sum to the difference between the model's prediction at the baseline and at the actual input. Consequently, we treat attribution *magnitudes* cautiously and primarily interpret the rank ordering and sign patterns of the top-attributed tokens.

3.4 Cross-Model Comparison

To assess whether trustworthiness representations transfer across model families, we evaluate seven models spanning two families (Qwen and Llama) and a range of sizes (0.5B–3.2B parameters). We additionally leverage chain-of-thought (CoT) reasoning from models that expose thinking tokens (Llama-3.2 and Qwen3), using the generated rationales as qualitative evidence for the features each model attends to.

4 Experiments and Results

4.1 Classification Accuracy

Table 1 reports few-shot classification accuracy across all seven models on our 100-article evaluation set.

Several findings stand out. First, performance does *not* scale monotonically with model size:

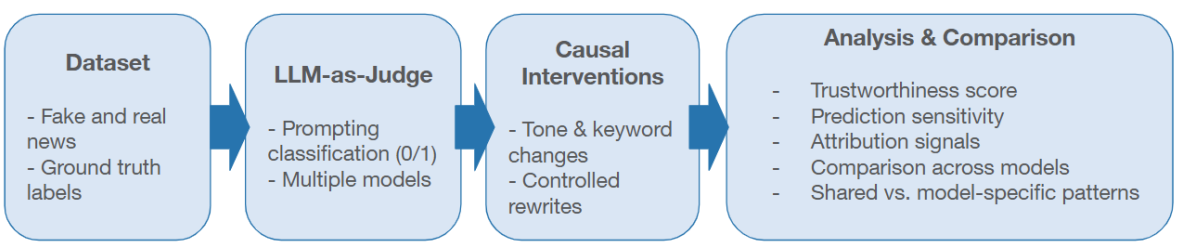

Figure 1: Overview of our pipeline. Articles from the FakeNewsDetection dataset are passed to a suite of LLM judges producing binary trustworthiness labels. We then apply causal interventions (counterfactual rewrites) and mechanistic attribution (Integrated Gradients) to understand what drives the predictions, and compare patterns across model families.

Model	Accuracy
Qwen2.5-0.5B-Instruct	0.45
Qwen3-1.7B	0.45
Phi-3-mini-4k-Instruct	0.55
Llama-3.2-3.2B-Instruct	0.63
Llama-3.2-1B-Instruct	0.67
Qwen3-0.6B	0.74
Qwen2-1.5B-Instruct	**0.91**

Table 1: Few-shot (k=5) classification accuracy on 100 balanced articles. Performance does not scale monotonically with model size.

the 1.5B Qwen2-1.5B-Instruct substantially outperforms both the 3.2B Llama model and Qwen3 variants. This suggests that instruction tuning recipe and training data composition may matter more than raw parameter count for this task. Second, the two smallest models (Qwen2.5-0.5B and Qwen3-1.7B) perform at chance level (0.45), likely due to poor instruction following on long-context news articles rather than a fundamental lack of semantic understanding. Third, few-shot prompting with k=5 provides modest improvements over zero-shot across most models, but does not uniformly close the gap for the worst performers.

4.2 Chain-of-Thought Reasoning

Models that expose chain-of-thought reasoning (Llama-3.2, Qwen3) provide qualitatively interpretable rationales. For a correctly classified trustworthy article, the model's internal reasoning highlights verifiable sourcing (e.g., references to the *Financial Times* and named government officials), neutral tone, and multi-perspective framing.

4.3 Mechanistic Attribution: Integrated Gradients

Figure 2 shows the top-attributed tokens for a representative article classified as trustworthy by Qwen2-1.5B-Instruct.

Positive (trustworthy-pushing) tokens. The highest-attributed positive tokens are overwhelmingly drawn from an institutional-news register: `government`, `agency`, `ruling`, `deal`, `warned`, and `Trump` (as a named-entity token). This reflects that articles that *read like Reuters* are treated as real by the model, regardless of their factual accuracy.

Negative (untrustworthy-pushing) tokens. The highest-attributed negative tokens fall into two groups: (1) meta-evaluative words associated with opinion or advocacy writing (`ratings`, `quality`, `National`), and (2) tokenization fragments of opinion-genre morphology (`atorial`, `ism`). Notably, competitive or rhetorical tokens like `win` and `winning` also carry strongly negative attributions. This suggests the model has internalized a " news vs. opinion" heuristic.

Completeness caveat. The completeness gap for our IG computation is 0.596. While increasing the number of integration steps (from 50 to 300) reduced numerical error, the gap persisted, suggesting that the PAD-token baseline is poorly suited to the transformer architecture. We therefore interpret sign patterns and rank orderings qualitatively, and treat precise attribution magnitudes with caution.

4.4 Counterfactual Intervention Study

We apply our counterfactual rewrite procedure to five articles initially classified as untrustworthy (label 0) by the Qwen2-1.5B-Instruct judge. Following the rewrite, **all five articles were reclassified as trustworthy (label 1)**, despite no changes to the core factual claims. Table 2 summarizes the linguistic changes made in each rewrite.

The most consistent intervention across articles is a reduction in inflammatory language (ΔInfl ≤ 0

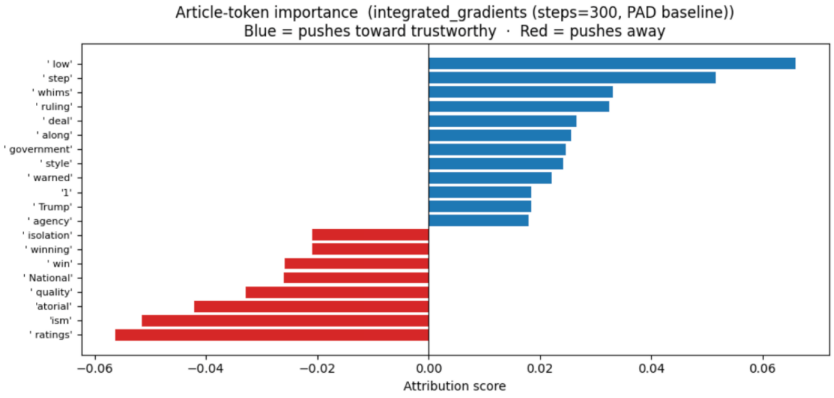

Figure 2: Token-level attribution scores (Integrated Gradients, 300 steps, PAD baseline). Blue bars indicate tokens that push the prediction toward *trustworthy*; red bars indicate tokens that push the prediction toward *untrustworthy*.

Idx	ΔHed	ΔCite	ΔInfl	ΔExl	ΔRhetQ	ΔCaps	ΔWords
0	−2	0	0	−1	0	−2	−387
1	+7	−1	−2	−1	0	−1	−139
2	+3	−1	−3	0	−3	+1	−230
3	+3	+1	−1	0	0	−1	−17
4	−7	−2	−1	0	−2	−1	−544

Table 2: Linguistic changes between original and counterfactual rewrites. All five articles flipped from 0 (untrustworthy) to 1 (trustworthy). ΔHed = change in hedging tokens; ΔCite = change in named-source attributions; ΔInfl = change in inflammatory/loaded language; ΔExl = change in exclamation tokens; ΔRhetQ = change in rhetorical questions; ΔCaps = change in all-caps tokens; ΔWords = change in total word count.

for all five articles). Rhetorical questions and all-caps tokens are also reduced in most cases. The pattern of changes is consistent with the IG attribution results: features that signal opinion or advocacy writing are the primary drivers of untrustworthy predictions, and removing them suffices to flip the model's judgment.

Qualitative examples. The rewrites illustrate several recurring transformation types:

- **Accusation → Attribution:** Direct accusatory framing is replaced by attributed claims:

  "It's yet another case of a Clinton cover up."

  ↓

  "Critics argue that repeated incidents involving sensitive government records raise concerns about accountability. However, the available email record does not establish intent."

- **Loaded language → Neutral vocabulary:** Emotionally charged terms are replaced with standard institutional terminology:

  "Obama regime", "Joe the Clown", "illegal aliens", "stinker"

  ↓

  "Obama administration", "former U.N. Ambassador John Bolton", "non-citizen voting debates"

- **Advocacy → Reporting:** Emotionally charged, mocking, and accusatory tone is replaced by a detached, structured, institutional register. Rhetorical questions are converted into analytical statements, and conspiracy-signaling language is removed entirely.

Implications. These results establish a concrete attack surface: a motivated actor could use a capable LLM to automatically rewrite partisan or misinformation content into a neutral institutional register, thereby inflating its perceived trustworthiness score without correcting any of the underlying factual claims.

5 Challenges and Lessons

Instruction following in long-context settings. Early experiments revealed frequent output-format failures (e.g., explanatory text or non-binary responses). Constraining the output vocabulary to 0,1 and adding few-shot examples substantially

improved compliance. However, instruction adherence degraded on longer articles, consistent with known limitations of smaller models in long-context settings.

Model size vs. accuracy. Classification accuracy does not scale linearly with model size within our evaluation range (0.5B–3.2B parameters). This non-monotonicity is practically important: it means that simply scaling up model size is not a reliable strategy for improving LLM-based content moderation, and that instruction tuning methodology plays an outsize role.

Integrated Gradients completeness. As noted in Section 4.3, the high completeness gap limits the reliability of IG magnitude estimates. Future work should explore alternative attribution baselines (e.g., Gaussian noise or learned baselines) or complementary methods such as attention rollout or LIME.

6 Discussion

Our results suggest that small LLMs rely primarily on surface-level register cues—such as institutional vocabulary, named sources, and neutral phrasing—rather than deeper assessments of factual accuracy. Both Integrated Gradients attribution and counterfactual interventions reveal a "reads like Reuters $\rightarrow$ must be real" heuristic.

These findings have implications for both the design of LLM-based content evaluation systems and the threat model for adversarial manipulation. On the design side, they suggest that prompting alone may be insufficient to elicit principled journalistic quality judgments from small models. Fine-tuning on a richer annotation framework may be necessary. On the threat side, they demonstrate that the barrier to adversarial manipulation is low: a capable rewrite model can substantially inflate trustworthiness scores without altering factual content.

7 Conclusion

We investigated the mechanisms by which LLMs judge the trustworthiness of news articles, using a combination of few-shot classification, mechanistic attribution via Integrated Gradients, and controlled counterfactual intervention. Across seven open-source models, we find that trustworthiness judgments are driven primarily by surface-level register features rather than by principled assessment of factual accuracy. Critically, purely stylistic rewrites that neutralize emotional language and partisan framing, while preserving all content information, are sufficient to flip model predictions from untrustworthy to trustworthy in all five tested cases. Our results highlight a vulnerability in LLM-based information curation and motivate future work on more robust approaches to automated news evaluation.

8 Related Work

Political bias in LLMs. Rozado (2023) elicits a liberal political slant from ChatGPT using political alignment tests. Motoki et al. (2024) measure bias using responses to Political Compass questions, while Bang et al. (2024) defines three dimensions of political bias, finding that LLM-generated text exhibit varying degrees of ideological lean. They do not examine implications of model biases in information curation or attempt causal analyses of linguistic features behind biases.

LLMs as content evaluators. Several studies examine the use of LLMs for information curation. Minici et al. (2025) compare commercial LLMs with traditional news aggregators, finding differences in source diversity, ideology, and reliability judgments. Yang and Menczer (2025) compare LLM reliability ratings to human experts and show that prompt design can significantly influence model assessments.

Mechanistic interpretation of LLM judgments. Loru et al. (2025) analyze how LLMs judge news reliability using a multi-agent pipeline that aggregates signals from several model components. While effective, the agentic design introduces interpretability challenges. Our approach instead combines token-level attribution with controlled counterfactual rewrites to enable clearer causal analysis of the linguistic features driving model predictions.

9 Contribution Statement

Kiko - Worked on prompt engineering, designing model pipeline, evaluating LLM results

Marie - Datasets, running and writing results, pruning output vocabulary, paper writing

Paul - Motivation and ideation, designing model pipeline, paper writing

Ramnath - Designing and implementing model pipeline, model selection, visualizations, Integrated gradients, and causal analysis of our models

References

Yejin Bang, Delong Chen, Nayeon Lee, and Pascale Fung. 2024. Measuring Political Bias in Large Language Models: What Is Said and How It Is Said. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11142–11159, Bangkok, Thailand. Association for Computational Linguistics.

Bhavik Jikadara. 2023. Fake news detection dataset. https://www.kaggle.com/datasets/bhavikjikadara/fake-news-detection. Accessed: Feb 2026.

Edoardo Loru, Jacopo Nudo, Niccolò Di Marco, Alessandro Santirocchi, Roberto Atzeni, Matteo Cinelli, Vincenzo Cestari, Clelia Rossi-Arnaud, and Walter Quattrociocchi. 2025. The simulation of judgment in LLMs. *Proceedings of the National Academy of Sciences*, 122(42):e2518443122.

Marco Minici, Cristian Consonni, Federico Cinus, and Giuseppe Manco. 2025. Auditing LLM Editorial Bias in News Media Exposure. *Preprint*, arXiv:2510.27489.

Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2024. More human than human: Measuring ChatGPT political bias. *Public Choice*, 198(1):3–23.

David Rozado. 2023. The Political Biases of ChatGPT. *Social Sciences*, 12(3).

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Kai-Cheng Yang and Filippo Menczer. 2025. Accuracy and Political Bias of News Source Credibility Ratings by Large Language Models. In *Proceedings of the 17th ACM Web Science Conference 2025*, Websci '25, pages 127–137, New York, NY, USA. Association for Computing Machinery.