CONTOUR: A Framework for Evaluating LLM-Generated and Human-Written Knowledge on Controversial Topics

Anonymous ACL submission

Abstract

Large language models increasingly mediate how users understand controversial social and political issues, yet their representations of contested knowledge remain underexplored. We introduce CONTOUR, a corpus and framework for comparing LLM-generated and human-written knowledge on controversial topics. The corpus includes 1,224 parallel articles across 153 politically sensitive topics, covering Wikipedia, Encyclopaedia Britannica, and six contemporary LLMs. CONTOUR evaluates these texts with 35 metrics spanning linguistic features, cognition, factual grounding, narrative roles, pro–con framing, and semantic perspective structure. Our analysis shows that LLMs match human sources in surface-level linguistic quality but differ in framing, using more positive sentiment, stronger balanced framing, and higher apparent perspective complexity. These findings reveal a tension between broader viewpoint coverage and artificial or "false" balance, offering an explainable foundation for evaluating AI-mediated controversial knowledge.

1 Introduction

Large language models (LLMs) are increasingly used as gateways to public knowledge. Users now turn to LLMs not only for factual lookup, but also for explanations of controversial social and political issues such as wars, social movements, public policy debates, and historical conflicts (Zamaraeva et al., 2025; Zeleke et al., 2025). In these settings, LLMs do more than summarize information: they decide which stakeholders to mention, which evidence to foreground, how to assign responsibility, and how to present competing viewpoints. These choices can shape how users understand conflict, legitimacy, uncertainty, and civic discourse.

Controversial knowledge poses a distinct challenge for evaluation. Unlike neutral factual topics, controversial topics involve competing stakeholders, contested evidence, moral disagreement, and disputes over legitimacy. A good representation of controversy is therefore not simply fluent, factually plausible, or ideologically "neutral." It must also manage tensions between viewpoint diversity, factual grounding, narrative framing, and the risk of misleading equivalence. For instance, presenting "both sides" may broaden perspective coverage, but it may also create superficial or false balance when (the actual) fringe viewpoints are elevated alongside well-established evidence.

Prior work has examined political bias, ideological leaning, and linguistic differences between LLM-generated and human-written texts (Khan et al., 2024; Ghafouri et al., 2023; Taubenfeld et al., 2024). However, existing studies often rely on a single baseline or focus on surface-level bias detection (Burnham et al., 2024; Haller et al., 2025), leaving limited tools for analyzing how controversy is constructed through cognition, narrative roles, factual grounding, and perspective structure (Motoki et al., 2024).

RQ.1 What linguistic metrics potentially capture the multidimensional nature of textual controversy?

RQ.2 To what extent can LLMs represent the complex and competing perspectives of controversial topics when compared with high quality human-curated knowledge systems?

RQ.3 How effectively can the proposed metrics predict human judgments of controversy?

To address **RQ1**, we introduce CONTOUR, a corpus and framework for investigating CONtroversial TOpic Understanding and Representation. The corpus contains 1,224 parallel articles across 153 politically sensitive topics, drawn from Wikipedia (community-negotiated consensus), Encyclopaedia Britannica (expert editorial knowledge), and six contemporary LLMs. Building on this corpus, CONTOUR operationalizes controversy through 35 metrics spanning linguis-

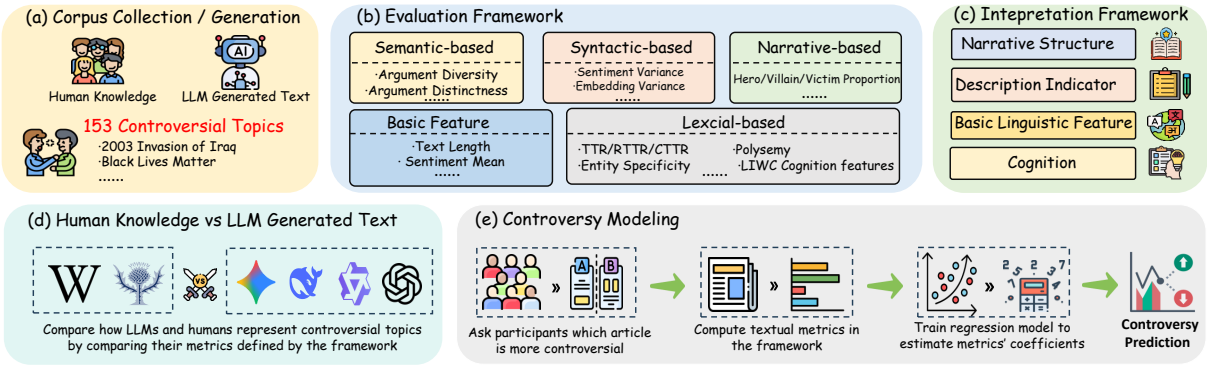


Figure 1: CONTOUR framework and experiment workflow illustration. (a) We collect a corpus of 153 controversial topics, each represented by both human-written articles (Wikipedia, Britannica) and LLM-generated articles. (b) For every article we compute a set of textual metrics in our proposed framework. (c) These metrics are mapped onto an interpretation framework. (d) We compare these metrics to analyze how LLM-generated and human-written texts differ in their representation of controversial topics. (e) We conduct human study to collect pairwise judgments of which article is more controversial, compute the textual metrics for each article, train a regression model to estimate metric coefficients and examine which metrics are associated with controversy.

tic, cognitive, descriptive, narrative, and semantic dimensions (**evaluation dimensions**), mapping onto surface features, LIWC-based cognitive indicators, factual grounding, stakeholder specificity, and argument-level perspective structure (**interpretation dimensions**). It provides an explainable and reproducible foundation for evaluating how AI systems mediate contested public knowledge.

To address **RQ2**, we apply CONTOUR framework to compare LLM- and human-generated controversial articles. LLMs achieve **comparable surface-level linguistic quality to human-curated sources**, yet differ substantially in controversy framing. This reveals a central tension—LLMs may broaden viewpoint coverage at surface-level while simultaneously **falling short in capturing deeper, argument-level controversy**, thus introducing an artificial balance.

Further, we address **RQ3** by using univariate and multivariate regression techniques and human-annotated controversy labels to identify which textual features are most predictive of perceived controversy. Results show that in general, **predicting perceived controversy from text is inherently difficult**. The group of "***sentence-level variance***" metrics are particularly diagnostic of perceived controversy, pointing to simple yet highly effective indicators in predicting controversy.

In summary, our key contributions include:

- **Controversial Topic Corpus**. Benchmark LLM-generated contents against two widely used, human-curated knowledge systems, Wikipedia and Encyclopaedia Britannica.
- **CONTOUR Evaluation Framework**. Develop a systematic and multidimensional evaluation framework for texts on controversial topics.
- **Comparing LLM vs. Human-generated Controversial Contents**. Use CONTOUR to compare what systematic linguistic patterns characterize LLMs and Human-curated knowledge systems on controversial topics.
- **Predictive Capability Exploration**. Associating the quantitative metrics with human annotated text controversy level.
- **Open-Sourced Python Library**. Release a Python package for implementing the evaluation framework.

2 Related Work

Controversial Topics and Political Bias in LLMs. Prior research has only partially explored how LLMs handle controversy. Studies have analyzed interpretations of scandalous news (Khan et al., 2024) and political leaning in discourse (Ghafouri et al., 2023). Broader literature documents systematic liberal leanings (Taubenfeld et al., 2024), Western-centric values (Motoki et al., 2024), and sensitivity to framing (Lunardi et al., 2024). Common methodologies include zero-shot stance classification (Burnham et al., 2024), political questionnaires (Haller et al., 2025), and persona-based role-playing (Motoki et al., 2024). To move beyond a single baseline, we evaluate LLM outputs against two distinct, human-curated knowledge systems:

Table 1: Basic linguistic feature performance across all models.

	Britannica	Wikipedia	DeepSeek-v3	Gemini-2.5	GPT-4o	GPT-4o-mini	Qwen-3-8B	Qwen-3-32B
1. Text Length	50,895	57,289	54,850	50,438	47,866	56,517	95,758	86,148
2. Sentiment Mean	-0.020	-0.087	0.3216	0.3665	0.4737	0.4665	0.2911	**0.1718**
3. POS Variability	0.01667	0.005	0.0017	0.0019	0.0089	0.0017	0.0075	0.0012
4. Vocabulary Complexity	1.0654	1.0971	1.0763	1.0685	1.0712	1.0745	1.0870	1.0902
5. Type-Token Ratio (TTR)	0.4078	0.2721	0.2496	0.2720	0.2739	0.2230	0.1923	0.2648
6. Root TTR (RTTR)	18.813	22.540	21.720	22.912	21.701	19.759	21.183	28.383
7. Corrected TTR (CTTR)	13.303	15.938	15.358	16.201	15.345	13.972	14.979	20.070
8. Normalized Entropy	0.8466	0.8053	0.8160	0.8181	0.8210	0.8086	0.7876	0.8149
9. Polysemy	4.6482	4.6256	4.3395	4.4218	4.4522	4.4835	4.3367	4.0305

Wikipedia and Encyclopedia Britannica.

Comparing LLM-Generated and Human-Written Texts. Prior work compares LLM- and human-authored texts using general linguistic features across domains such as news, narrative fiction, and online discussions (Muñoz-Ortiz et al., 2024; Zhang et al., 2024; Zamaraeva et al., 2025; Zeleke et al., 2025; Uchendu et al., 2023), as well as domain-specific analyses of persuasive and argumentative writing (Falk and Lapesa, 2023; Dönmez et al., 2025). Beyond NLP, public policy research has identified narrative strategies relevant to persuasive discourse in contested settings (Rupinsky et al., 2023). However, no existing work provides a systematic comparison framework that explicitly targets controversial, high-stakes knowledge.

3 Controversial Topic Corpus

We construct a corpus of controversial topics across human-curated and LLM-generated texts.

3.1 Corpus Collection

Human-generated Controversial Articles. Data collection began in June 2025 using Wikipedia's official curated list[1] of contested articles as a sampling frame. Topics were manually matched to corresponding **Encyclopedia Britannica** entries, yielding **153 political topics** with aligned articles from the two sources, which served as the human-authored reference corpora.

LLM-Generated Controversial Articles. For each topic, we generated parallel articles using six large language models spanning closed-source (GPT-4o, GPT-4o-mini (Achiam et al., 2023), Gemini 2.5 Flash-Lite (Comanici et al., 2025)) and open-source variants (DeepSeek-V3 (DeepSeek-AI, 2025), Qwen3-32B, Qwen3-8B (Yang et al., 2025)), reflecting diverse institutional and cultural development contexts.

[1] List of Controversial Issues in Wikipedia.

The final corpus includes 1,224 articles in total, documenting how 153 controversial topics are presented from 8 different sources. See more experimental setup details in Appendix A.2.

3.2 LLMs Comparable to Human Sources in Basic Features

A brief comparison of basic linguistic features are shown in Table 1. LLMs are generally more positive in sentiment than human writers, but there is no significant differences across a set of basic metrics related to vocabulary richness and diversity. Overall, LLMs are achieving highly comparable performance at the lexical level with human-curated knowledge systems.

3.3 Human Annotation on Controversy Level

To obtain ground-truth controversy labels for validating the CONTOUR, we collected human annotations on article pairs from four sources: Wikipedia, Britannica, DeepSeek-V3, and GPT-4o, covering 153 topics. For each topic, annotators identified the more controversial article between two randomly selected sources. Annotators could evaluate up to three pairs and reported confidence on a 5-point Likert scale. Participants were recruited via Prolific and restricted to US-based respondents. More details can be found in Appendix A.5.

4 CONTOUR Framework

We introduce CONTOUR, a unified framework for systematically evaluating and interpreting how LLMs and humans represent controversial topics.

4.1 Framework Construction

To capture the abstract and multidimensional nature of controversy, we conduct a systematic, multidisciplinary survey spanning social science, cognition, writing pedagogy, and political science literature. From this survey, we identify a list of theoretically motivated constructs underlying controversy, beyond basic linguistic features, and map each

Evaluation	Metrics	Interpretation	Evaluation	Metrics	Interpretation
Basic Features	1. Text Length	Basic Linguistic Features	Syntactic-based	19. Sentiment Variance	Descriptive Indicators (Sentence Variance)
	2. Sentiment Mean			20. Embedding Variance	
Lexical-based	3. POS Variability		Narrative-based	21. Hero Proportion	Description Indicators (Role Distribution)
	4. Vocabulary Complexity			22. Villain Proportion	
	5. Type-Token Ratio (TTR)			23. Victim Proportion	
	6. Root TTR (RTTR)			24. Stakeholders' claims	Description Indicators (Factual Evidence)
	7. Corrected TTR (CTTR)		Lexical-based	25. Entity specificity	
	8. Entropy			26. Country Name	
	9. Polysemy			27. Date Time	
	10. Cognition Class	Cognition	Syntactic-based	28. Balanced Pro and Con	Narrative Structure (Contradition)
	11. Memory		Narrative-based	29. Angel-Devil Shift	
	12. All-or-None		Semantic-based	30. Main vs Fringe Perspective Ratio	
	13. Insight			31. Main Characters Pro & Con Ratio	
	14. Causation			32. Argument Diversity	Narrative Structure (Perspective Complexity)
	15. Discrepancy			33. Argument Distinctness	
	16. Tentativity			34. Argumentativeness	
	17. Certitude			35. Deliberation Intensity	
	18. Differentiation				

Figure 2: CONTOUR Framework.

construct to quantitative metrics drawn from both domain-specific and NLP literature. For instance, argument-level perspective complexity has been reviewed by Németh (2023); lexical and syntactic complexity by Irani et al. (2025) and Ghafouri et al. (2023); and narrative strategy, such as role differentiation, has been discussed by Rupinsky et al. (2023).

As shown in Table 2, CONTOUR combines an **Evaluation Dimension** (§4.2), which operationalizes discourse properties through 35 linguistically grounded metrics, and an **Interpretation Dimension** (§4.3), which maps these metrics to theoretically motivated dimensions of controversy writing. The framework is neither exhaustive nor intended to be applied wholesale. Its multidimensional structure and the clustering of multiple metrics onto shared theoretical constructs afford flexible, selective use.

4.2 Evaluation Dimension

The Evaluation Dimension captures how texts construct and organize competing perspectives across **five linguistic levels**: *basic*, *lexical*, *syntactic*, *narrative*, and *semantic* based metrics. With details of all features elaborated in Appendix A.1.1, we briefly introduce each metric group next.

Basic Features describe overall textual properties, primarily including `Text Length`, and `Sentiment Mean`.

Lexical-based metrics capture cognitive, stylistic, and informational properties expressed through word choice. They include **cognitive process indicators**, extracted via LIWC categories, such as `insight`, `cause`, `tentativeness`, `certainty`, `discrepancy`, `differentiation`, and absolutist language, which together reflect the depth and complexity of reasoning. **Lexical diversity and complexity** are quantified through measures like `part-of-speech variability`, `vocabulary richness (TTR, RTTR, CTTR)`, `Shannon entropy`, and WordNet-based `polysemy`. Finally, `entity specificity` is measured by the proportion of unique named entities relative to total tokens.

Syntactic-based metrics capture surface-level patterns and structural variation in texts that are largely independent of semantic content. These features include `balanced pro-con percentage`, which measures the extent to which sentences explicitly juxtapose opposing viewpoints using contrastive constructions, and variance-based metrics, which quantify sentence-level heterogeneity such as fluctuations in sentiment or semantic orientation. For instance, given a document of n sentences, let s_i denote the sentiment score of sentence i. `Sentiment Variance` is

$$\text{SentVar} = \frac{1}{n}\sum_{i=1}^{n}(s_i - \bar{s})^2, \bar{s} = \frac{1}{n}\sum_{i=1}^{n} s_i,$$

capturing how much the emotional tone fluctuates across the document.

Narrative-based metrics capture higher-level discourse strategies that shape role attribution and moral framing across an article. They include **narrative roles**, which classify entities into archetypal categories such as `hero`, `villain`, or `victim`, and

compute their normalized proportions. Another key metric is the `Angel-Devil Shift`, defined as

$$\text{AngelDevilShift} = \frac{\#\text{Hero} - \#\text{Villain}}{\#\text{Hero} + \#\text{Villain} + 1},$$

which quantifies narrative asymmetry by measuring the balance between heroic and blame-centric framing.

Semantic-based metrics capture how multiple viewpoints are represented and organized at the level of meaning in texts on controversial topic. Key measures include the `Main vs. Fringe Perspective Ratio`, which quantifies the dominance of central viewpoints relative to marginal ones, and the `Main Character Pro-Con Ratio`, which evaluates the polarity and consistency of positions expressed by primary actors. **Argument structure** is assessed through a set of argument mining techniques, measuring `argument diversity`, `distinctness`, and overall `argumentativeness`.

4.3 Interpretation Dimension

Building on evaluation measurements, the Interpretation Dimension provides an explainable lens that links observed linguistic patterns to **four higher-level constructs** central to understanding the complexity of controversial discourse. In addition to *basic linguistic features*, *we highlight three domain-specific elements: cognition, descriptive indicators, and narrative structure*. Details are elaborated in Appendix A.1.1

Cognition. Processing complex social information requires analytical depth and context-rich explanations to support informed decision-making. To quantify this cognitive depth, we utilize the established LIWC 22 cognitive processes taxonomy (Boyd et al., 2022). This framework employs nine metrics to identify lexical indicators of in-depth processing, such as expressions of discrepancy, certainty, and differentiation.

Descriptive Indicators. Descriptive indicators encompass the manifest textual attributes that constitute the foundational components of controversy discourse. This dimension focuses on quantifiable, surface-level features such as *sentiment* and *embedding variance*, which reflect the emotional and semantic consistency of the text. Furthermore, it incorporates *role distribution* of heroes, villains, and victims role labels. This dimension also provides a granular map of the texts, capturing the mentioned entities (*country names*) and factual evidence (*stakeholders' claims*, and the *specificity of entities*).

Narrative Structure. Narrative structure refers to the latent structural mechanisms used to curate, sequence and re-organize the complexity of multiple viewpoints. This dimension points to the contradiction and perspective complexity inherent in the discourse. By measuring metrics such as the *Angel-Devil Shift*, *Main vs. Fringe Perspective Ratios*, this dimension captures how a narrative navigates binary oppositions. It also assesses the depth of writing through metrics like *deliberation intensity* and *argumentative diversity*,

4.4 Human Annotation for Metric Validation

To assess the metric reliability, we conduct human-annotation studies with two experts. The study evaluates LLM-based judgments on several metrics.

Stakeholder Claims and Stance (LLM-based). We further validate LLM-based classification for *stakeholder claims* and sentence-level stance. Annotators label whether a sentence explicitly attributes a claim to a stakeholder and, when applicable, whether the sentence expresses a *pro*, *con*, or *neutral* stance toward that stakeholder.

Balanced Pro–Con Patterns and People Extraction (LLM-based). Annotators additionally label (i) whether a sentence exhibits a *balanced pro–con* construction, and (ii) the extraction of *narrative roles*, including `human`, `villain`, and `victim`.

Results. We report accuracy as the primary evaluation metric. **Narrative role extraction** achieves an accuracy of **0.86** (105 annotated instances). **Stakeholder's Claim**, evaluated on 120 annotated sentences, achieves an accuracy of **0.84**. **Balanced pro–con detection**, evaluated on 133 sentences, achieves an accuracy of **0.9**.

5 Comparing LLM vs. Human-generated Controversial Texts

We make several key observations of the LLM vs Human-generated controversial texts comparison.

LLMs Exhibit Distinct Cognitive Preferences from Humans. Figure 3 reveals that LLMs and humans have distinct and nuanced preferences over the nine items related to cognitive processes. Humans tend to use more words related to *Insight, Causation*, *Certitude* and *Discrepancy*, suggesting

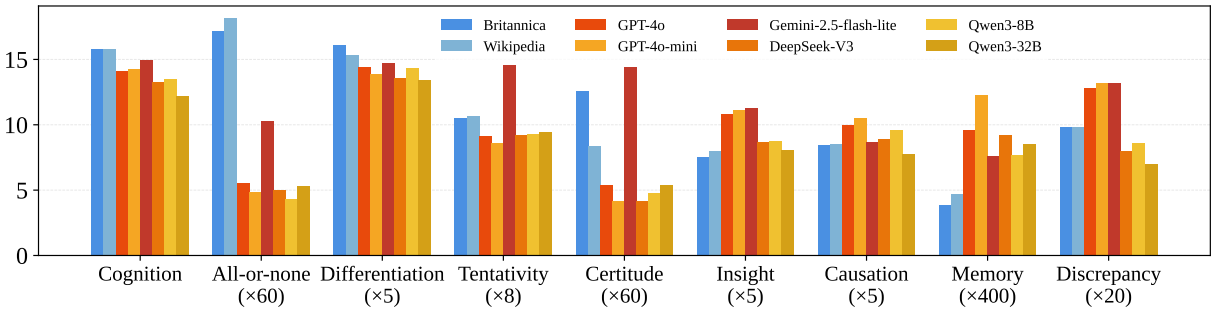


Figure 3: LIWC cognition features. Some metrics are rescaled for better visualization.

an emphasis on explaining causal mechanisms and internal contradictions. LLMs outscore humans in *Memory, Differentiation*, and *All-or-none* thinking, reflecting LLMs tendency toward binary or exhaustive classification. Overall, humans and machines exhibit distinct cognitive preferences in how they present complex issues.

LLMs has Lower Diversity in Sentence-level Variance Metrics. Human systems consistently achieved higher diversity than LLMs in sentence-level variance metrics (see Figure 2). This means human writers exhibit greater differentiation in sentiment and opinion across individual sentences within an article.

Humans Prefer Explicit Factual Evidence. Humans also prefer explicit citation of factual evidence more than models, as shown in Appendix A.1.1 Figure 6b. Human writers more frequently reference specific stakeholders and entities.

LLMs More Balanced in Narrative-level Metrics. However, distinct patterns between humans and LLMs emerge in stakeholders' *role distribution*, as shown in Appendix A.1.1 Figure 6a. Human systems emphasize negative roles of "Villains" and "Victims" more, while LLMs show a preference for depicting "Heroes." This is also consistent with LLMs overall higher *sentiment* scores in Table 2 and more pronounced *angel-devil shift* in Appendix A.1.1 Figure 7a.

Narrative structures points to a clear discrepancy between the inherent perspective complexity and manifest linguistic features above. As demonstrated in Appendix A.1.1 Figure 7a, human-curated systems are less balanced than LLMs when juxtaposing opposing viewpoints, particularly regarding the distribution of "*pro and con*" statements of main characters, and the weights between *mainstream vs. fringe perspectives*. Human systems also generally exhibit lower perspective complexity in Appendix A.1.1 Figure 7b, scoring lower on three out of four metrics related to argumentative diversity and the inclusion of distinct viewpoints.

6 Exploring Predictive Capability of CONTOUR Metrics

To assess how strongly each of the theoretically-identified metrics in CONTOUR predicts human controversy judgments, we adopt a conditional logistic regression (CLR)(Kuo et al., 2018) over 153 article pairs.

6.1 Modeling Setup

Each pair consists of two articles that share the same topic but come from different sources. The ground truth controversy labels were obtained in a human study (§3.3).

Formally, for a pair containing the more-controversial article with feature vector x_w and the less-controversial article with feature vector x_l, the model estimates

$$P(x_w \succ x_l) = \sigma\big(\beta^\top (x_w - x_l)\big)$$

where β is a vector of metric coefficients and $\sigma(\cdot)$ is the logistic function.

6.2 Conditional Logistic Regression (Single-Variable)

We first fit a separate CLR for each metric in isolation, as shown in Figure 4.

Positive Association. Controversial content is positively associated with greater lexical diversity (*Cognition Class, Differentiation, Type–Token Ratio*), higher affective and semantic variability (*Sentiment Variance, Embedding Variance*), richer temporal and factual grounding (*Date Time, Entity Specificity*), and greater epistemic hedging (*All-or-None, Tentativity, Discrepancy*).

Negative Association. Higher *Sentiment Mean* and longer *Text Length* are associated with lower controversy, suggesting that more overtly opinionated

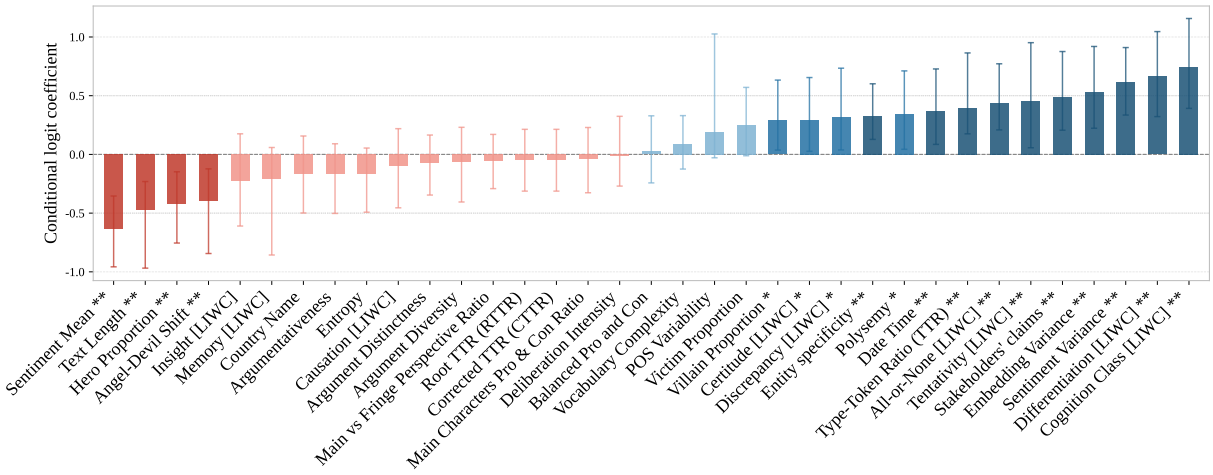


Figure 4: Conditional logistic regression coefficient (single variable). FDR corrected, $^{*}p < .05$, $^{**}q < .05$. Bar color indicates significance and direction: dark blue/red = FDR-significant ($q < .05$); medium blue/red = nominally significant ($p < .05$); light blue/red = non-significant. Error bars represent 95% bootstrap confidence intervals.

or elaborated texts are judged as less controversial. Notably, a strong *Angel-Devil Shift* and high *Hero Proportion* also reduce perceived controversy, indicating that narratives with clear moral polarization or a dominant heroic frame are seen as less ambiguous and thus less controversial.

Table 2: Basic Linguistic Features VIF Screening Results. The results for the remaining groups can be found in the appendix.

Group	Metric	VIF	Status
Basic Linguistic	RTTR	∞	×
	Entropy	17.99	×
	POS Variability	9.92	×
	Text Length	3.94	✓
	Sentiment Mean	1.12	✓
	Vocabulary Complexity	1.35	✓
	TTR	3.39	✓
	CTTR	1.57	✓
	Polysemy	1.10	✓

6.3 Multivariate Regression

We then performed a multivariate CLR for each interpretation group separately.

VIF filtering. To address multicollinearity, we applied iterative Variance Inflation Factor (VIF) filtering, removing any feature with VIF > 5. This step filtered out 5 metrics and remain 30 metrics.This step filtered out 5 metrics, leaving 30 metrics remaining. For example, as Table 5 shows, three metrics in the Basic Linguistic feature group were filtered out due to high VIF values. The rest results are in Appendix A.6 Table 5.

Multivariate Results. As Figure 5 shows, the multivariate results largely converge with the univariate findings, but with notable attenuation. *Sentiment Variance* and *Stakeholders' claims* remain significant positive predictors, confirming that sentence-level affective variability and the presence of multiple stakeholder voices are robust signals of controversy independent of other features within their respective groups.

Interpretation Group	Accuracy	AUC
Basic Linguistic Features	0.630	0.621
Role Distribution	0.630	0.593
Sentence Variance	**0.696**	**0.640**
Contradiction	0.565	0.545
Factual Evidence	0.609	0.602
Cognition	0.652	0.619
Perspective Complexity	0.543	0.505

Table 3: Multivariate conditional logistic regression model classification accuracy and AUC on the held-out test set, grouped by interpretation framework.

Predictive Performance. To further evaluate each group's predictive capacity, we assessed within-pair classification accuracy and AUC on a held-out test set. As shown in Table 3, predicting perceived controversy from text is inherently difficult. Yet several groups consistently exceed the chance baseline of 0.5 suggests that controversy is not entirely idiosyncratic but can be recoverable from quantifiable linguistic and narrative features. *Sentence Variance* achieves the highest test accuracy (0.696) and AUC (0.640), suggest-

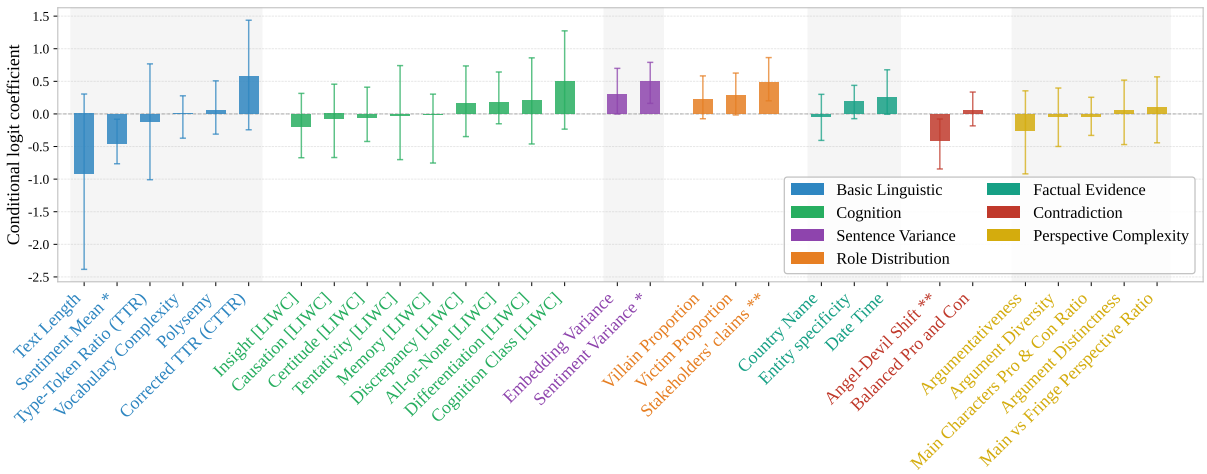


Figure 5: Multivariate conditional logistic regression coefficients for grouped metrics after VIF filtering. FDR corrected, $^{*}p < .05$, $^{**}q < .05$.

ing that higher sentiment variance and embedding variance within a text are particularly diagnostic of perceived controversy. These results support the view that controversy, while difficult to predict in absolute terms, can be meaningfully associated with measurable textual properties.

7 Discussion and Implication

Our findings show that controversy cannot be reduced to surface-level neutrality or balance; instead, it emerges through the interaction of linguistic, factual, narrative, and semantic structures, etc.

RQ1: Multidimensional Metrics for Controversy. Our multidisciplinary review yields 35 theoretically grounded metrics spanning five linguistic levels. We argue that controversy is not reducible to surface-level features: the framework distinguishes between *manifest* linguistic properties (e.g., sentiment, lexical diversity) and *latent* structural properties (e.g., argument-level perspective complexity, narrative role attribution). This separation proves consequential, as **LLMs and human writers diverge most sharply at the latent level** rather than at the surface level.

RQ2: LLMs vs. Human Representations of Controversy. LLMs achieve comparable surface-level linguistic quality to human-curated sources yet differ substantially in controversy framing. LLMs favor more positive sentiment, stronger balanced framing, and higher apparent perspective complexity, while human writers show greater sentence-level variance and richer factual grounding in specific stakeholders and entities. Critically, LLMs' higher perspective complexity does not straightforwardly indicate better controversy representation. It may reflect genuine breadth enabled by comprehensive training data, or a form of "false balance" that artificially tries to equate fringe and mainstream viewpoints. Distinguishing these interpretations requires qualitative expert validation and remains an important direction for future work.

RQ3: Predicting Human Controversy Judgments. Although predicting perceived controversy from text is inherently difficult, several metric groups consistently exceed chance, suggesting controversy is partially recoverable from linguistic features. Sentence Variance achieves the highest held-out accuracy (0.696) and AUC (0.640), identifying affective and semantic fluctuation across sentences as the most diagnostic signal. In multivariate analysis, Sentiment Variance and Stakeholders' Claims remain robust predictors, pointing to affective instability and attributional grounding as core controversy dimensions. Conversely, surface-level balance indicators—Angel-Devil Shift and Hero Proportion—are *negatively* associated with perceived controversy, confirming that narratively "cleaner" texts are judged as less contested.

8 Conclusion

We proposed CONTOUR, a systematic framework for unpacking the multi-layered nature of controversial discourse. Rather than evaluating LLMs against a single notion of "neutrality" or "bias," our framework makes explicit the specific goals, narrative strategies, and cognitive processes involved in writing about contested topics. It offer model developers a nuanced benchmark and policymakers a diagnostic tool for assessing AI's impact on democratic discourse.

Limitations

This study has several limitations. First, our analysis focuses on English-language content and a fixed set of 153 controversial topics, which may limit the generalizability of our findings to other languages, cultural contexts, or topics. Second, we benchmark LLM outputs against Wikipedia and Encyclopaedia Britannica as representative human-curated knowledge systems; while widely used, these sources do not exhaust the diversity of human perspectives and may themselves reflect institutional or cultural biases. Third, our evaluation relies on human-in-the-loop judgment, which may also introduce human bias. Finally, we study a snapshot of LLM behavior under specific prompting conditions and model versions; as models and deployment practices evolve, the observed patterns may change over time.

Ethical Considerations

This work analyzes how large language models (LLMs) and human-curated knowledge sources represent controversial topics. The study relies exclusively on publicly available texts, including Wikipedia, Encyclopedia Britannica, and LLM-generated outputs produced via standard prompts; no private, proprietary, or personally identifiable data are used. Because controversial topics inherently involve normative and contested perspectives, our framework does not assume a single correct or neutral representation. Instead, it provides descriptive metrics for comparing framing, argument structure, and perspective diversity across sources. A potential risk is that these metrics could be misinterpreted as definitive indicators of political bias, ideological correctness, or normative quality, or be used to rank systems along a single evaluative dimension. We caution against such uses and emphasize that the proposed measures are intended for comparative and analytical purposes rather than prescriptive judgments.

AI Usage

Large Language Model was used solely for grammar correction, language polishing, and coding assistance during the preparation of this work. The research ideas, study design, analysis, interpretations, and conclusions were entirely developed by the authors and were not generated by AI.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin*, 10:1–47.

Michael Burnham, Kayla Kahn, Ryan Yang Wang, and Rachel X Peng. 2024. Political debate: Efficient zero-shot and few-shot classifiers for political text. *Political Analysis*, pages 1–15.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Esra Dönmez, Maximilian Maurer, Gabriella Lapesa, and Agnieszka Falenska. 2025. AI Argues Differently: Distinct Argumentative and Linguistic Patterns of LLMs in Persuasive Contexts. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34583–34614, Suzhou, China. Association for Computational Linguistics.

Neele Falk and Gabriella Lapesa. 2023. Bridging Argument Quality and Deliberative Quality Annotations with Adapters. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2469–2488, Dubrovnik, Croatia. Association for Computational Linguistics.

Vahid Ghafouri, Vibhor Agarwal, Yong Zhang, Nishanth Sastry, Jose Such, and Guillermo Suarez-Tangil. 2023. AI in the Gray: Exploring Moderation Policies in Dialogic Large Language Models vs. Human Answers in Controversial Topics. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 556–565. ArXiv:2308.14608 [cs].

Patrick Haller, Jannis Vamvas, Rico Sennrich, and Lena Ann Jäger. 2025. Leveraging In-Context Learning for Political Bias Testing of LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24718–24738, Vienna, Austria. Association for Computational Linguistics.

Arman Irani, Jinyoung Park, Kevin Esterling, and Michalis Faloutsos. 2025. Wiba: What is being argued? a comprehensive approach to argument mining. In *Social Networks Analysis and Mining*, volume

15211 of *Lecture Notes in Computer Science*, page 21. Springer, Cham.

Arman Irani, Ju Yeon Park, Kevin Esterling, and Michalis Faloutsos. 2024. What Is Being Argued (WIBA)? An Application to Legislative Deliberation in the US Congress. *arXiv preprint arXiv:2407.06149*.

Awais Hameed Khan, Hiruni Kegalle, Rhea D'Silva, Ned Watt, Daniel Whelan-Shamy, Lida Ghahremanlou, and Liam Magee. 2024. Automating thematic analysis: how llms analyse controversial topics. *arXiv preprint arXiv:2405.06919*.

Chia-Ling Kuo, Yuhua Duan, and James Grady. 2018. Unconditional or conditional logistic regression model for age-matched case-control data? *Frontiers in Public Health*, 6:57.

Riccardo Lunardi, David La Barbera, and Kevin Roitero. 2024. The Elusiveness of Detecting Political Bias in Language Models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, pages 3922–3926, New York, NY, USA. Association for Computing Machinery.

George A. Miller. 1995. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41.

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

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2024. Contrasting Linguistic Patterns in Human and LLM-Generated News Text. *Artificial Intelligence Review*, 57(10):265.

Renáta Németh. 2023. A scoping review on the use of natural language processing in research on political polarization: trends and research prospects. *Journal of computational social science*, 6(1):289–313. Publisher: Springer.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108. Association for Computational Linguistics.

Shae Rupinsky, Madeline Schomburg, Gabriel Chandler, and Carrington Gelardi. 2023. Shifting narrative strategies: How monument advocates change their stories in response to conflict over time. *Policy Studies Journal*, 51(1):101–122. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/psj.12480

Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. 2024. Systematic Biases in LLM Simulations of Debates. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 251–267. ArXiv:2402.04049 [cs].

Adaku Uchendu, Jooyoung Lee, Hua Shen, Thai Le, Dongwon Lee, and 1 others. 2023. Does human collaboration enhance the accuracy of identifying llm-generated deepfake texts? In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 11, pages 163–174.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Olga Zamaraeva, Dan Flickinger, Francis Bond, and Carlos Gómez-Rodríguez. 2025. Comparing LLM-generated and human-authored news text using formal syntactic theory. *arXiv preprint*. ArXiv:2506.01407 [cs].

Brook Zeleke, Amish Soni, and Lydia Manikonda. 2025. Human or GenAI? Characterizing the Linguistic Differences between Human-Written and LLM-Generated Text. In *Companion Publication of the 17th ACM Web Science Conference 2025*, pages 34–37, New Brunswick NJ USA. ACM.

Chengming Zhang, Florian Hofmann, Lea Plößl, and Michaela Gläser-Zikuda. 2024. Classification of reflective writing: A comparative analysis with shallow machine learning and pre-trained language models. *Education and Information Technologies*, 29(16):21593–21619.

A Appendix

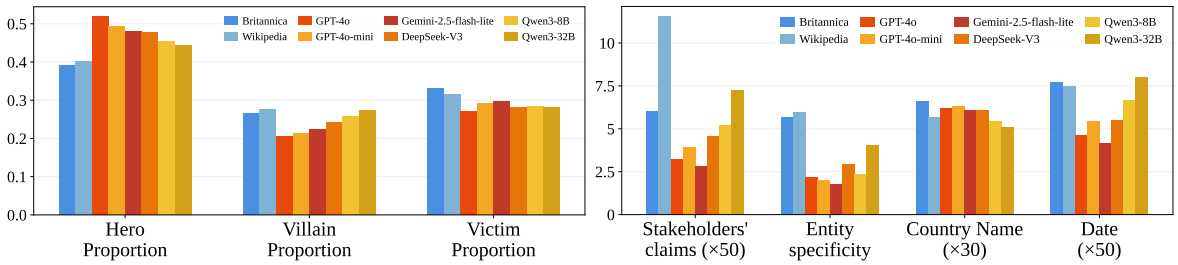
A.1 CONTOUR Framework Details

A.1.1 Evaluating Textual Controversial Topics with 35 Metrics

Semantic-based Features. Semantic-based features capture how multiple viewpoints are represented and organized at the level of meaning in texts on controversial topics. Following the feature taxonomy in Figure 2, these features focus on perspective structure, argumentative organization, and deliberative reasoning rather than surface form. They address a central question: to what extent does a text explicitly construct, contrast, and sustain multiple perspectives at the semantic level?

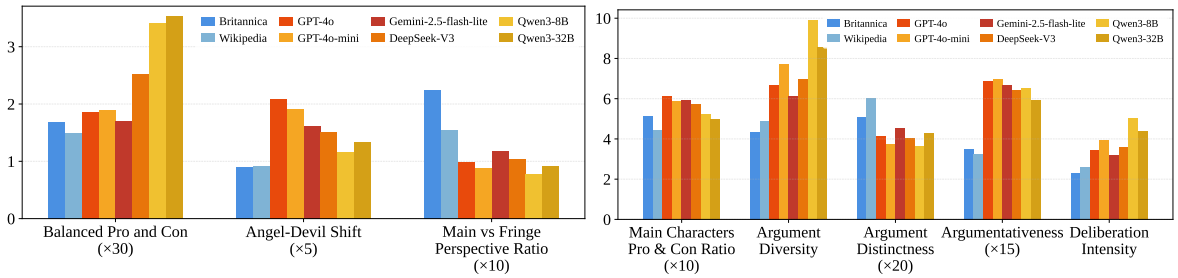
Main vs. Fringe Perspective Ratio. This metric measures the dominance of central viewpoints relative to marginal or fringe perspectives. Perspectives are identified via clustering over sentence-level semantic embeddings. Let n_{main} denote the number of sentences associated with the dominant cluster and n_{fringe} the number associated with all remaining clusters. The ratio is defined as

$$\text{MainFringeRatio} = \frac{n_{\text{main}}}{n_{\text{fringe}} + 1}.$$



(a) Role distribution (b) Factual Evidence

Figure 6: Quantitative analysis of Description Indicators dimensions



(a) Contradiction (b) Perspective Complexity

Figure 7: Quantitative analysis of Narrative Structure dimensions

Main Character Pro–Con Ratio. We focus on the most salient stakeholders to avoid noise from marginal or sparsely mentioned actors. Specifically, we first identify all stakeholders mentioned in the text and rank them by frequency of occurrence, retaining only the top 5 stakeholders.

For each of these stakeholders, we compute evaluative polarity across sentences associated with that stakeholder. Let

$$p_i \in \{+1, -1, 0\}$$

denote a pro, con, or neutral stance expressed toward the stakeholder in sentence s_i. For each stakeholder, we report both the mean and variance:

$$\mu_{\text{procon}} = \frac{1}{N}\sum_{i=1}^{N} p_i, \qquad \sigma^2_{\text{procon}} = \text{Var}(p_i),$$

where N is the number of sentences referring to the stakeholder. These statistics capture overall directional bias and internal evaluative inconsistency. Aggregate results are reported over the top 5 stakeholders.

Argument Structure and Deliberation. Argumentative units are identified using a pretrained argument mining model. We employed WIBA (What Is Being Argued) (Irani et al., 2024), a powerful argument mining tool that classifies a piece of text as argumentative, achieving an "F1 score between 79% and 86% on three different benchmark datasets", according to the developers of the tool. Argument diversity is measured as the number of semantically distinct argument clusters normalized by argument count. Argument distinctness captures the semantic separation between clusters based on inter-centroid cosine distance.

Argument diversity measures the breadth of distinct argumentative perspectives. Let K denote the number of argument clusters and N_{arg} the total number of arguments. Diversity is defined as

$$\text{ArgDiversity} = \frac{K}{\log(1 + N_{\text{arg}})},$$

and is further normalized by $\log_2(N_{\text{arg}} + 1)$ to improve comparability across texts. We additionally report the entropy of the cluster size distribution to capture the balance of argumentative perspectives.

Argument distinctness captures the semantic separation between argumentative perspectives. For each cluster, we compute a centroid embedding. Let $\mathcal{D}$ denote the set of pairwise cosine distances between all cluster centroids. Argument distinctness is defined as

$$\text{AD} = \sqrt{\bar{d} \cdot d_{\min}},$$

where $\bar{d}$ and $d_{\min}$ are the mean and minimum values in $\mathcal{D}$, respectively. This formulation empha-

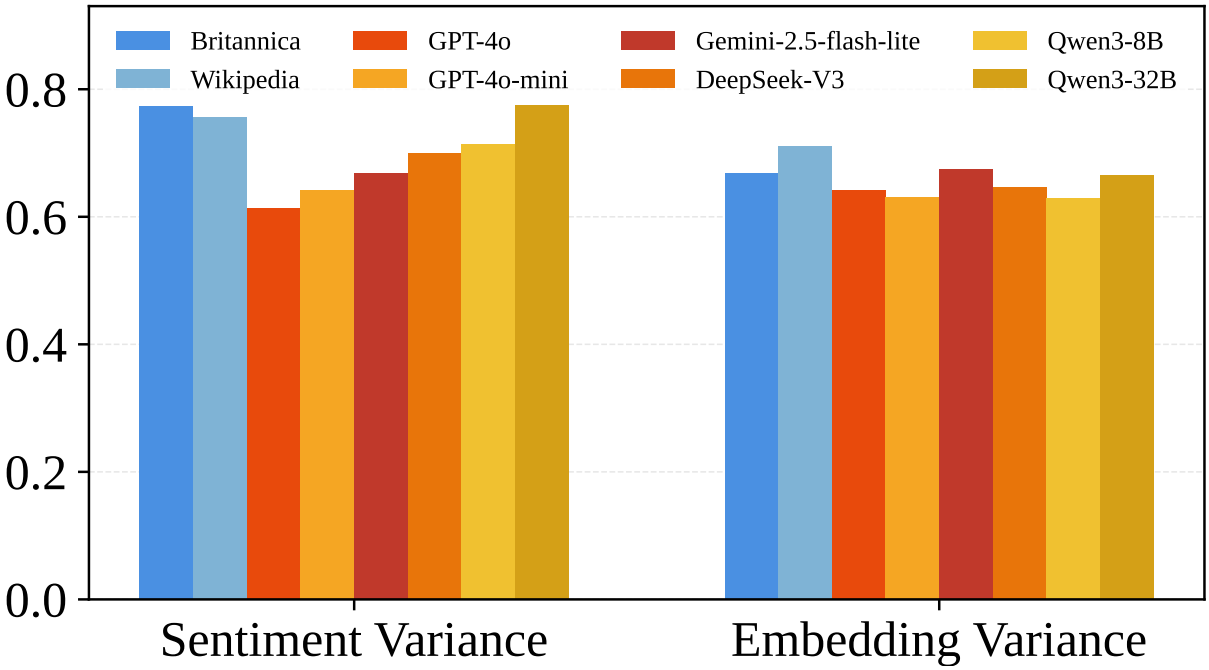


Figure 8: Quantitative Analysis of Description Indicator (Sentence Variance) dimension

sizes both global separation and the closest competing perspectives.

Argumentativeness. Argumentativeness measures the extent to which a text is organized around explicit argumentative reasoning. Using WIBA predictions over sliding windows, let $\mathbb{1}_{\text{arg}}(s_i)$ indicate whether window s_i is classified as argumentative above a confidence threshold. The metric is defined as

$$\text{Argumentativeness} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}_{\text{arg}}(s_i),$$

where N is the total number of windows in the text.

Deliberation Intensity. Deliberation intensity reflects the degree to which a text engages multiple, distinct argumentative perspectives. It is computed as a composite score combining normalized argument diversity (D_{div}) and narrative distinctness (D_{nd}):

$$\text{DelibIntensity} = \frac{D_{\text{div}} + D_{\text{nd}}}{2}.$$

Higher values indicate broader coverage of perspectives and clearer semantic separation between competing arguments, corresponding to stronger deliberative structure.

Syntactic-based Features. Syntactic-based features capture surface-level realization patterns and internal variation that are largely independent of semantic content. These features reflect stylistic regularities and structural balance within texts.

Balanced Pro and Con (LLM-based detection). Balanced framing measures whether a sentence explicitly juxtaposes opposing viewpoints using contrastive constructions. Let $\mathbb{1}_{\text{bal}}(s_i)$ denote whether sentence s_i is balanced. The balanced ratio is defined as

$$\text{BalancedRatio} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}_{\text{bal}}(s_i).$$

Variance-based Features. To capture internal heterogeneity and instability within a document, we compute variance-based features at the sentence level. These measures quantify the extent to which different parts of a text diverge in affective tone or semantic content.

Sentiment variance measures fluctuations in emotional polarity across sentences. Sentence-level sentiment scores are computed using a pre-trained DistilBERT sentiment classifier, which maps each sentence onto a continuous scale in $[-1, 1]$. Variance is then computed across all sentence scores:

$$\sigma_s^2 = \frac{1}{N}\sum_{i=1}^{N}(s_i - \mu_s)^2,$$

where s_i denotes the sentiment score of sentence i and μ_s is the document-level mean sentiment. Higher values indicate stronger oscillation between positive and negative emotional framing.

Embedding variance captures semantic dispersion across sentences. Each sentence is encoded using the `all-MiniLM-L6-v2` SentenceTransformer model, producing dense semantic embeddings. Variance is computed as the average squared distance from the mean embedding:

$$\sigma^2_{\text{emb}} = \frac{1}{N}\sum_{i=1}^{N} \|e_i - \bar{e}\|^2 ,$$

where e_i is the embedding of sentence i and $\bar{e}$ is the mean embedding across the document.

Basic Features. Basic features provide coarse-grained controls describing the overall structural and stylistic properties of a document. These features are not designed to directly capture controversy, but instead serve as important baselines and normalization controls when interpreting higher-level discourse features.

Text length is measured as the total number of characters in the article. This feature captures document scale and verbosity, which may influence the prevalence of other linguistic phenomena simply through exposure effects.

Mean sentiment represents the average emotional polarity across all sentences:

$$\mu_s = \frac{1}{N}\sum_{i=1}^{N} s_i,$$

where s_i is the sentiment score assigned to sentence i. Positive values indicate overall positive framing, while negative values indicate more negative framing.

Lexical-based Features. Lexical-based features capture cognitive, stylistic, and informational properties realized through word choice.

Cognitive Process Indicators. Using the Linguistic Inquiry and Word Count (LIWC), which maps words to psychologically meaningful lexical categories, we extract normalized word-category frequencies as proxies for cognitive and epistemic stance. Specifically, we use: cognition (general analytic thinking), insight (reflective understanding), memory (references to past experience), cause (causal reasoning), tentativity (hedging and uncertainty), certitude (confident assertion), discrepancy (unmet expectations or contradictions), differentiation (drawing distinctions), and all-or-none (absolute, binary framing).

Lexical Diversity and Complexity. We compute several lexical diversity and complexity measures to characterize the richness, variability, and ambiguity of language use within a document. These include part-of-speech variability, vocabulary complexity, type–token ratio (TTR), root TTR (RTTR), corrected TTR (CTTR), Shannon entropy, and WordNet-based(Miller, 1995) polysemy.

Type–token ratio (TTR) measures the proportion of unique word types relative to the total number of tokens:

$$\text{TTR} = \frac{V}{N},$$

where V denotes the number of unique word types and N is the total number of tokens. Higher TTR values indicate greater lexical diversity and less repetition. However, standard TTR is sensitive to document length and tends to decrease as texts become longer.

To reduce this length sensitivity, we additionally compute *root TTR (RTTR)*:

$$\text{RTTR} = \frac{V}{\sqrt{N}},$$

which normalizes vocabulary size by the square root of the token count. RTTR provides a more stable estimate of lexical diversity across texts of varying lengths.

We also compute the *corrected TTR (CTTR)*:

$$\text{CTTR} = \frac{V}{\sqrt{2N}},$$

a further length-adjusted variant commonly used in computational linguistics and stylometric analysis. Compared to standard TTR, both RTTR and CTTR reduce the inflation or compression effects introduced by document size.

Shannon entropy measures the distributional unpredictability of lexical usage:

$$H = -\sum_{i=1}^{V} p(w_i)\log p(w_i),$$

where $p(w_i)$ denotes the empirical probability of word type w_i. Higher entropy indicates a broader and more evenly distributed vocabulary,

while lower entropy reflects concentration around a smaller set of recurring terms.

Polysemy measures lexical ambiguity using WordNet synsets. For each content word, we compute the number of associated senses in WordNet and average across the document. Higher polysemy values indicate greater semantic ambiguity and interpretive flexibility.

Finally, *part-of-speech variability* captures syntactic diversity by measuring variation in part-of-speech usage across the text, while *vocabulary complexity* estimates the prevalence of rarer or morphologically complex words.

Entity Specificity. Entity specificity reflects grounding in concrete actors, institutions, temporal references, and geographic locations. We used the Stanford Stanza package(Qi et al., 2020) for named entity recognition (NER) to extract entities from the text. It is defined as

$$\text{EntitySpecificity} = \frac{\#\text{unique named entities}}{T/100},$$

where T is the total number of tokens. We additionally track two sub-proportions of entity types relative to all named entities:

$$\text{DateProportion} = \frac{\#\text{DATE entities}}{\#\text{total named entities}},$$

$$\text{CountryProportion} = \frac{\#\text{GPE entities}}{\#\text{total named entities}},$$

Narrative-based Features. Narrative-based features capture higher-level discourse strategies that guide attribution, moral framing, and interpretive alignment.

Stakeholder Claims (LLM-based detection). Stakeholder claims measure the extent to which a text explicitly attributes positions, arguments, or viewpoints to identifiable social actors or groups. Let $\nVdash_{\text{stake}}(s_i)$ denote whether sentence s_i contains an explicit stakeholder-attributed claim. The metric is defined as

$$\text{StakeholderClaims} = \frac{1}{N}\sum_{i=1}^{N} \nVdash_{\text{stake}}(s_i),$$

where N is the total number of sentences in the text.

Narrative Roles (LLM-based detection). Entities are classified into three archetypal roles drawn from narrative theory: hero, villain, and victim. The *hero* is the protagonist who acts to resolve a conflict, protect others, or advance a valued cause—typically portrayed with agency and moral legitimacy. The *villain* is the antagonist whose actions are framed as harmful, threatening, or morally illegitimate, serving as the source of conflict. The *victim* is the entity that suffers harm, injustice, or loss, often used to elicit sympathy and justify the actions of heroes against villains. These roles are not mutually exclusive and reflect the framing of entities within the narrative rather than objective attributes. Let $\#\text{People}_r$ denote the number of entities assigned role r. The normalized role proportion is

$$P(r) = \frac{\#\text{People}_r}{\#\text{People}}, \quad r \in \{\text{Hero}, \text{Villain}, \text{Victim}\}.$$

Angel–Devil Shift. Narrative asymmetry is quantified as

$$\text{AngelDevilShift} = \frac{\#\text{Hero} - \#\text{Villain}}{\#\text{Hero} + \#\text{Villain} + 1}.$$

Positive values indicate heroic framing, while negative values indicate blame-centric narratives.

A.1.2 Interpretation Dimension

Drawing on social science research, we propose an explainable interpretation framework that ties the linguistic metrics above to key dimensions for evaluating writing on controversial topics. In addition to basic linguistic features, we highlight three domain-specific elements for analyzing such texts: **cognitive processes**, manifest **descriptive indicators**, and latent **narrative structure**.

Cognition. The processing of complex social information requires that texts engage with the topics in an analytical and interpretive manner. Readers also anticipate that such materials will provide rich, context-sensitive explanations to support informed decision-making. This requirement for thoroughness and cognitive depth can be captured through lexical indicators of cognitive processing, such as expressions of discrepancy, certainty, and differentiation. To quantify the degree of in-depth cognitive processes, we employ the established LIWC 22 cognitive processes taxonomy which includes nine metrics.

Descriptive Indicators. Descriptive indicators encompass the manifest textual attributes that constitute the foundational components of controversy

discourse. This dimension focuses on quantifiable, surface-level features such as sentiment and embedding variance, which reflect the emotional and semantic consistency of the text. Furthermore, it incorporates role distribution (the proportion of heroes, villains, and victims). These indicators provide a granular map of the texts, capturing the mentioned entities (geographic locations) and factual evidence (stakeholders' claims, and the specificity of entities). They could be understood as surface-level linguistic attributes of controversy contents.

Narrative Structure. Narrative structure refers to the latent structural mechanisms used to curate, sequence and re-organize the complexity of multiple viewpoints. This dimension goes beyond manifest attributes and points to evaluate the contradiction and perspective complexity inherent in the discourse. By measuring metrics such as the "*Angel-Devil Shift*," "*Main vs. Fringe Perspective Ratios*," this dimension captures how a narrative navigates binary oppositions. It also assesses the depth of the debate through deliberation intensity and argumentative diversity, quantifying the degree to which a text preserves the nuances and multidimensionality of an issue. Inherent and structural complexity is central to effective writing about controversies, as it could deepen readers' understanding of complex social issues.

A.2 Experimental Setup Details

A.2.1 Corpus Collection

Data collection started in June 2025, including the following steps.

First, we compiled a list of controversial topics. Controversiality could be directly reflected in the fact that many people hold different or even conflicting viewpoints, giving rise to sustained contestation, negotiation, and even debates among individuals and groups. As an initial sampling frame, we drew on Wikipedia's curated list of highly controversial topics, which contains articles that frequently become the focus of editorial disputes and therefore require particular attention from the community to ensure adherence to the Neutral Point of View (NPOV) policy. From this list, we identified all 262 politics-related topics, and then matched these topics with the Encyclopaedia Britannica database. In cases where an exact title match between Wikipedia and Britannica was not available, the research team manually identified the most closely related Britannica entry (for instance, the Wikipedia article "2003 invasion of Iraq" was matched to the Britannica entry "Iraq War"). Topics for which no adequate Britannica counterpart could be determined were excluded from the dataset. This procedure yielded a final corpus of 153 political topics for which content pages existed in both Wikipedia and Britannica. Second, we retrieved the full text of each selected article from both Wikipedia and Britannica. Third, we employed several prominent large language models (LLMs), developed in diverse national and cultural contexts, to generate articles on the same set of 153 topics, thereby producing a parallel corpus of model-generated texts for subsequent analysis.

A.2.2 Models

To generate comparative texts, we employ a diverse set of large language models spanning different model sizes and degrees of openness. Specifically, we use GPT-4o and GPT-4o-mini as representative closed-source models with strong general-purpose and lightweight capabilities, respectively. In addition, we include Gemini 2.5 Flash-Lite to extend the evaluation beyond the GPT family within the class of proprietary models. We further evaluate several open-source models, including DeepSeek-V3, Qwen3-32B, and Qwen3-8B. This model set covers a broad spectrum of contemporary LLM design choices, allowing us to examine whether observed differences in our framework are consistent across models. For metrics that require LLM-based detection, we use GPT-4o-mini to identify text content.

A.3 Prompts for LLM Articles Generation

Because LLM generation has length limitations, we employed a multi-turn generation approach so that the resulting articles are more comparable in length to human-written articles. We first let the LLM generate an initial set of sections covering the topic, as shown in Figure 9, where the model is instructed to produce major sections in a single pass. Once the initial draft is produced, we apply a second turn for each section individually, as shown in Figure 10, in which the model is asked to greatly expand the section with more specific details, examples, and contextual depth while preserving the original heading and avoiding any summarizing statements. This two-turn pipeline allows the final article to reach a length that would otherwise exceed the output token limit of a single LLM call, making the generated articles more structurally and

quantitatively comparable to human-written counterparts.

A.4 Discussion

We develop a systematic evaluation and interpretation framework designed to unpack the multi-layered complexity of discourse on controversial issues. Prior research often oversimplifies political content evaluation by focusing on "unbiased" models, a goal that is neither practical nor even desirable in this context. Instead, we argue that preserving diverse viewpoints and the intricate balance between stakeholders first requires a framework that reflects the specific goals, narrative techniques, and cognitive processes involved in such sophisticated writing tasks. This framework provides model developers with a nuanced benchmark for performance evaluation and offers policymakers a critical tool for assessing AI's impact on democratic discourse and even election behaviors.

- Implication 1: surface-level complexity vs. latent structural complexity

There exists a clear discrepancy between manifest linguistic features that reflect surface-level complexity, and the inherent perspective complexity. Model developers are achieve comparable results or at least not bad performance in terms of the basic linguistic features and lexical level attributes. Future steps require researchers to more carefully analyze what these argument level complexity means for presenting comprehensive understanding on controversial issues.

- Implication 2: developing high quality contents on controversial topics

We do not argue that maximizing perspective complexity should be the ultimate goal for model development. Rather, the higher perspective complexity observed in LLMs may stem from two possible explanation that are yet to be confirmed. First, LLMs may be genuinely more balanced due to their comprehensive training data, which allows them to bypass the institutional logics that inherently bound human-curated systems. While AI research frequently focuses on identifying political and social biases, our results suggest that LLMs can indeed present controversial topics well, with reasonable depth and complexity that differs from the training benchmarks like Wikipedia or Britannica.

Alternatively, this higher diversity may represent "false balance" or "forced balance" achieved through specific narrative techniques. Distinguishing between true neutrality and artificial balancing requires collaboration with domain experts to qualitatively evaluate LLM-generated contents. This study establishes a necessary foundation for a systematic, explainable evaluation framework, but further research is required to enrich our qualitative understanding of these divergent narrative strategies.

A.5 Human Evaluation

We administered the task via a self-developed website. Participants first completed an attention check after reading the task instructions; those who failed were excluded from proceeding. In the main task, annotators were presented an article pair side by side. To control for article length effects, we extracted excerpts from the beginning, middle, and end of each article to roughly equalize article length. Each annotator can evaluate up to three article pairs and self-reported their confidence on a 5-point Likert scale after each judgment. Participants were recruited through Prolific, restricted to US-based respondents to ensure cultural background similarity, and compensated at approximately £0.70 GBP per task. Upon entering the study, participants completed an attention check after reading the task instructions. In the main task, each participant evaluated up to three article pairs. For each pair, two excerpts on the same topic drawn from different sources (Wikipedia, Britannica, or an LLM) were presented side by side. Participants dragged each article into one of two columns — *Less Controversial* or *More Controversial* — based on their personal impression. The interface is illustrated in Figures 14 and and 15. To assess reliability, 50 article pairs were re-annotated by at least three independent annotators, with majority vote used as the final label. Individual annotations achieved 89.3% accuracy for highest-confidence responses and 73.6% for lower-confidence responses, indicating reliable label quality.

A.6 VIF Screening Results

To address multicollinearity in the multivariate regression, we applied iterative Variance Inflation Factor (VIF) filtering, removing any feature with VIF > 5. This step filtered out 5 metrics — RTTR (∞), Entropy (17.99), POS Variability (9.92), Hero Proportion (6.10), and Deliberation Intensity (55.03), retaining 30 metrics for the multivariate analysis. Table 5 reports the full VIF screening results across all interpretation groups.

Prompt 1: Initial Comprehensive Article Generation

Write a comprehensive article about "`{topic}`".

Requirements for comprehensive coverage:

1. Start with the main title `# {topic}`
2. Include a substantial introduction that defines the topic and explains its significance
3. Create several major sections. **IMPORTANT**: Quality over quantity — choose the most important and relevant sections.
4. Use **ONLY** ## level headings for sections — no subsections
5. Include detailed content for each section
6. Do **NOT** include any conclusion, summary, or final thoughts sections
7. Do **NOT** add sections about future prospects or outlook
8. Focus on informative content

IMPORTANT:

This is the initial comprehensive coverage phase. Create as many relevant major sections as needed to thoroughly cover the topic from all important angles.

Topic: `{topic}`

Write the complete initial comprehensive article:

Figure 9: Prompt 1: Initial article generation (Turn 1)

Prompt 2: Section Expansion

You are expanding a section from an article about "`{topic}`".

Current section that needs expansion:

```
## {section_title}
{section_content}
```

Your task is to significantly expand this section with much more detail, depth, and comprehensive coverage.

Requirements:

1. Keep the same section heading: `## {section_title}`
2. Greatly expand the content with more specific details, examples, explanations, and context
3. Add more depth to the discussion of the topic
4. Include more specific information and details
5. Make the section much more comprehensive and informative
6. Write in detailed paragraphs — do **NOT** create subsections
7. Do **NOT** add any conclusion or summary statements at the end
8. Focus purely on expanding the informational content related to this specific aspect of `{topic}`

Write the complete expanded section with significantly more detail and depth:

Figure 10: Prompt 2: Section expansion (Turn 2)

Prompt 3: Balanced Pro/Con Detection

Determine whether the following sentence presents BOTH pro and con viewpoints on the same topic.

A balanced sentence should:

– Present arguments or perspectives from opposing sides

– Use contrastive connectives like "but", "however", "while", "although", "on the other hand"

– Show symmetric patterns like "Some argue X, others argue Y"

– NOT just mention two different topics without contrast

Examples of BALANCED sentences:

– "While supporters praise the policy's economic benefits, critics worry about environmental costs."

– "Some believe AI will create jobs, but others fear widespread unemployment."

Examples of NON-BALANCED sentences:

– "The policy has economic benefits." *(only one side)*

– "Critics oppose the plan." *(only one side)*

– "The company produces cars and also makes phones." *(no contrast, different topics)*

Sentence: `"{sent}"`

Output STRICT JSON:

```
{
  "is_balanced": true/false,
 "pattern_type": "symmetric_pattern" | "contrastive_connective" | "none"
}
```

Figure 11: Prompt 3: Balanced Pro/Con sentence detection

Prompt 4: Narrative Role Detection

Identify Hero / Villain / Victim roles in the text based on how entities are portrayed.

Definitions:

– **HEROES**: Entities portrayed as protagonists, saviors, or positive agents
– **VILLAINS**: Entities portrayed as antagonists, causing harm, or acting maliciously
– **VICTIMS**: Entities portrayed as suffering, harmed, or negatively affected

Instructions:

1. Identify specific entities (people, groups, organizations, countries)
2. Classify based on NARRATIVE ROLE in the text, not objective reality
3. An entity can appear in multiple roles
4. Use specific names/labels from the text

Text: `"{text}"`

Output STRICT JSON:

```
{
  "heroes":  ["entity1"],
  "villains": ["entity2"],
  "victims":  ["entity3"]
}
```

If no entities fit a category, use `[]`.

Figure 12: Prompt 4: Narrative role (Hero/Villain/Victim) detection

Prompt 5: Stakeholder Claims Detection

Does this sentence REFERENCE or QUOTE other people's viewpoints?

INTERACTIVE:

Attributions ("X says"), citations ("According to…"), references ("The CEO stated…")

NON-INTERACTIVE:

Plain facts with no source attribution.

Sentence: `"{sent}"`

Return JSON:

```
{
  "is_interactive": true/false,
  "citation_phrases": ["all citation phrases found, empty if none"]
}
```

Figure 13: Prompt 5: Stakeholder Claims Detection

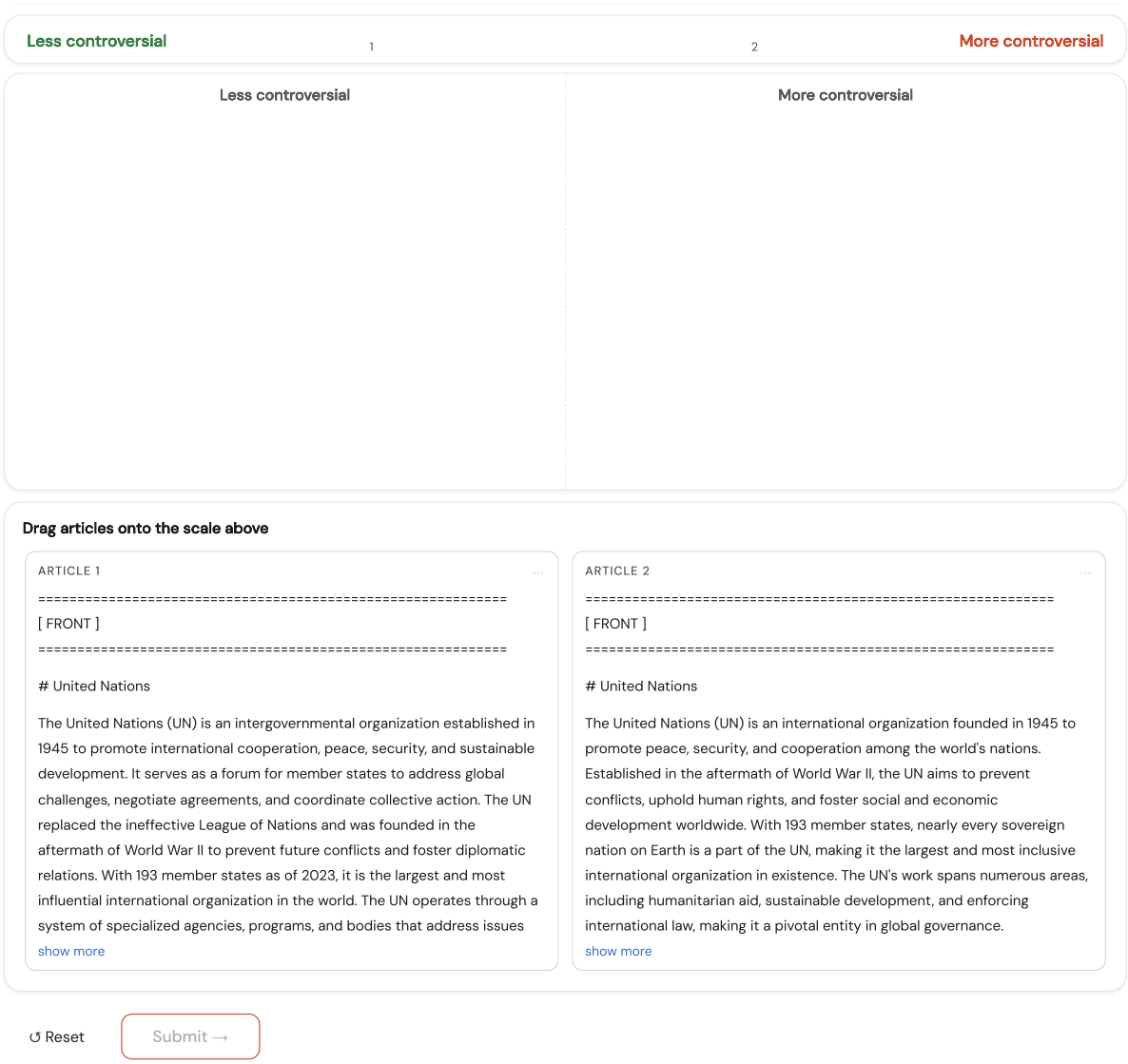


Figure 14: The interface for collecting human ground truth on comparing the controversy level of the same topic across two different sources.

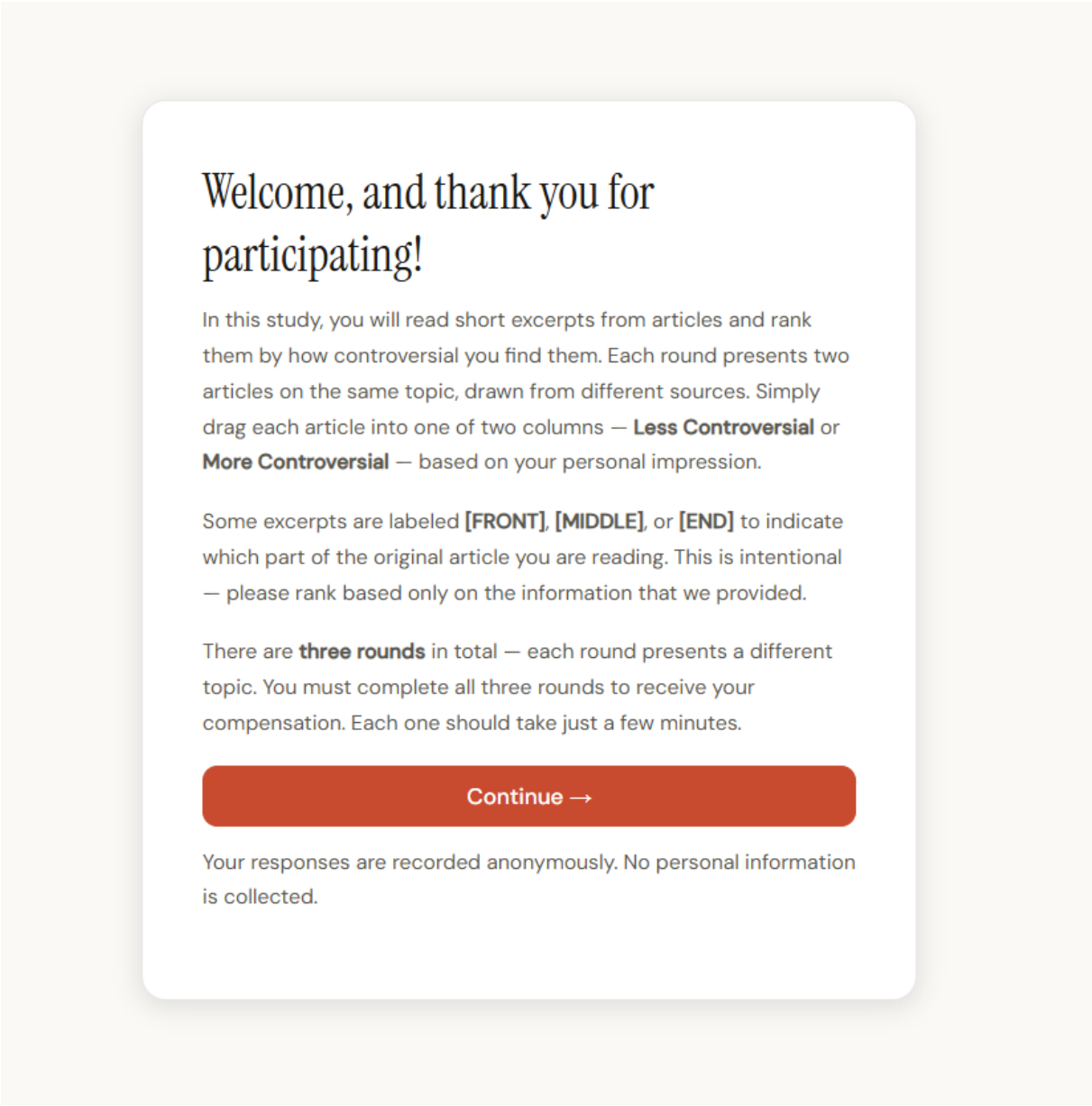


Figure 15: Instruction page of the human evaluation interface for controversy comparison tasks.

Table 4: Per-metric univariate conditional logistic regression performance ranked by test AUC.

Metric	Group	Accuracy	AUC
allnone	Cognition	0.622	0.639
pos_variability	Basic Linguistic	0.667	0.635
certitude	Cognition	0.600	0.634
differ	Cognition	0.674	0.625
sentiment_mean	Basic Linguistic	0.739	0.613
embedding_variance	Sentence Variance	0.578	0.606
text_length	Basic Linguistic	0.652	0.605
Cognition	Cognition	0.609	0.598
sentiment_variance	Sentence Variance	0.652	0.591
datetime_ratio	Factual Evidence	0.609	0.591
interactivity	Role Distribution	0.717	0.584
polysemy	Basic Linguistic	0.543	0.582
entity_specificity	Factual Evidence	0.609	0.578
ttr	Basic Linguistic	0.587	0.562
angel_devil_shift	Contradiction	0.609	0.546
cogproc	Other	0.587	0.539
discrep	Cognition	0.543	0.539
tentat	Cognition	0.478	0.539
victim_proportion	Role Distribution	0.522	0.535
hero_proportion	Role Distribution	0.543	0.531
memory	Cognition	0.478	0.527
villain_proportion	Role Distribution	0.457	0.524
insight	Cognition	0.413	0.520
main_vs_fringe_ratio	Perspective Complexity	0.435	0.516
argumentativeness	Perspective Complexity	0.457	0.514
main_characters_pro_con_ratio	Perspective Complexity	0.543	0.499
vocabulary_complexity	Basic Linguistic	0.565	0.497
entropy	Basic Linguistic	0.522	0.491
cause	Cognition	0.543	0.488
country_name_ratio	Factual Evidence	0.543	0.482
cttr	Basic Linguistic	0.500	0.474
rttr	Basic Linguistic	0.500	0.474
balanced_pro_con	Contradiction	0.457	0.469
argument_distinctness	Perspective Complexity	0.413	0.468
argument_diversity	Perspective Complexity	0.435	0.456
deliberation_intensity	Perspective Complexity	0.348	0.448

Table 5: Full VIF Screening Results.

Group	Metric	VIF	Status
Basic Linguistic	RTTR	∞	$\times$
	Entropy	17.99	$\times$
	POS Variability	9.92	$\times$
	Text Length	3.94	✓
	Sentiment Mean	1.12	✓
	Vocabulary Complexity	1.35	✓
	TTR	3.39	✓
	CTTR	1.57	✓
	Polysemy	1.10	✓
Cognition	Cognition Class	4.40	✓
	Memory	1.06	✓
	All-or-None	1.50	✓
	Insight	1.83	✓
	Causation	1.65	✓
	Discrepancy	1.84	✓
	Tentativity	2.34	✓
	Certitude	1.63	✓
	Differentiation	2.53	✓
Sentence Variance	Sentiment Variance	1.00	✓
	Embedding Variance	1.00	✓
Role Distribution	Hero Proportion	6.10	$\times$
	Villain Proportion	1.10	✓
	Victim Proportion	1.02	✓
	Stakeholders' Claims	1.08	✓
Factual Evidence	Entity Specificity	1.14	✓
	Country Name	1.23	✓
	Date Time	1.19	✓
Contradiction	Balanced Pro and Con	1.02	✓
	Angel-Devil Shift	1.02	✓
Perspective Complexity	Deliberation Intensity	55.03	$\times$
	Main vs Fringe Perspective Ratio	3.18	✓
	Main Characters Pro & Con Ratio	1.00	✓
	Argument Diversity	1.41	✓
	Argument Distinctness	2.89	✓
	Argumentativeness	4.28	✓