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

Anonymous ACL submission

Abstract

As Large Language Models (LLMs) increasingly serve as primary information-seeking tools, their role in shaping public perception of complex social issues becomes a critical concern for civic discourse. This study investigates the representation of controversial topics in LLMs. We propose a comparative framework that benchmarks LLM-generated content against two human-curated knowledge systems: Wikipedia, representing community-driven consensus, and Encyclopedia Britannica, representing elite expert viewpoints. Through a multi-dimensional linguistic analysis across 153 politically sensitive topics, we quantify the "intensity of controversy" using a novel taxonomy.

1 Introduction and Related Work

Large language models (LLMs) are increasingly used as general-purpose knowledge sources, raising concerns about how they present controversial topics that are inherently contested, value-laden, and socially consequential. Understanding how LLMs represent such topics is therefore critical for assessing their impact on public knowledge formation and democratic discourse. Prior research has only partially explored this problem, examining LLM interpretations of scandalous or sensitive news (Khan et al., 2024), political leaning in generated discourse (Ghafouri et al., 2023), and broader patterns of ideological bias. This literature documents systematic liberal leanings (Taubenfeld et al., 2024), Western-centric value orientations (Motoki et al., 2024), and sensitivity to framing effects (Lunardi et al., 2024), often using methodologies such as zero-shot stance classification (Burnham et al., 2024), political questionnaires (Haller et al., 2025), and persona-based role-playing (Motoki et al., 2024). In parallel, a growing body of work compares LLM- and human-authored texts using general linguistic or stylistic features across domains including 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 and communication research has identified narrative strategies relevant to persuasion and legitimacy in contested political settings (Rupinsky et al., 2023). However, existing approaches typically rely on a single normative baseline and do not provide a systematic comparison framework that explicitly targets controversial, high-stakes knowledge. To address this gap, we ask: (1) how can LLM-generated texts on controversial topics be systematically evaluated, and (2) how do LLM-generated texts on controversial topics differ from human-written texts? We address these questions by introducing the CONTOUR framework, which examines **CON**troversial **TO**pic **U**nderstanding and **R**epresentation in both LLM-generated and human-written knowledge, and by comparing LLM outputs against two distinct human-curated reference systems—Wikipedia and Encyclopedia Britannica. Our contributions are four folds:

- **CONTOUR Evaluation Framework**. develop a systematic evaluation framework for texts on controversial topics.
- **Empirical Interpretation Framework**. interpretation method to analyze the empirical findings from the CONTOUR outcomes.
- **Controversial Topic Dataset**. benchmark LLM-generated contents against two widely used, human-curated knowledge systems, Wikipedia and Encyclopaedia Britannica, to pinpoint the key linguistic distinctions
- **Open-Sourced Python Library**. release a Python package for implementing the evaluation framework.

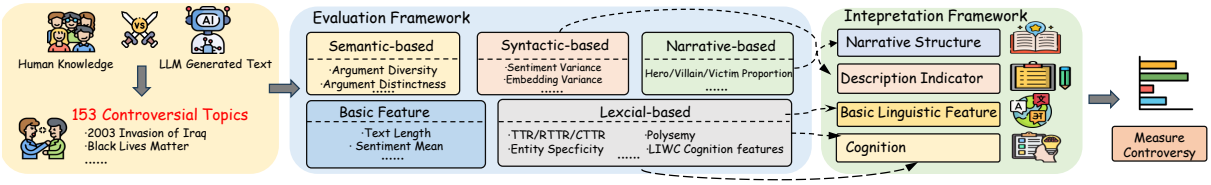


Figure 1: CONTOUR Framework and Experiment Workflow Illustration

2 CONTOUR Framework

As shown in Table 1, CONTOUR combines an **Evaluation Dimension** (§2.1), which operationalizes discourse properties through 35 linguistically grounded metrics, and an **Interpretation Dimension** (§2.2), which maps these metrics to theoretically motivated dimensions of controversy writing.

2.1 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.

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, and the `Angel-Devil Shift`, 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`.

2.2 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

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.

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*,

3 Experimental Setups

Corpus Collection. We drew an initial sampling frame from **Wikipedia**'s curated list[1] of highly controversial articles. These topics were then matched to corresponding entries in **Encyclopedia Britannica**. This process yielded **153 political topics** with aligned Wikipedia and Britannica articles.

LLM-Generated Controversial Topics. For each topic, we then generated parallel articles using a diverse set of large language models developed across different institutional and cultural contexts, including closed-source models (GPT-4o, GPT-4o-mini (Achiam et al., 2023), and Gemini 2.5 Flash-Lite (Comanici et al., 2025)) and open-source models (DeepSeek-V3 (DeepSeek-AI, 2025), Qwen3-32B, and Qwen3-8B (Yang et al., 2025)). This diverse model selection enables a systematic comparison of LLM-generated and human-curated representations of controversial topic knowledge. See more experimental setup details in Appendix A.2.

[1] List of Controversial Issues in Wikipedia.

4 Findings

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

Cognition. 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 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.

Descriptive Indicators. Across three subtypes of manifest descriptive indicators, LLMs

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

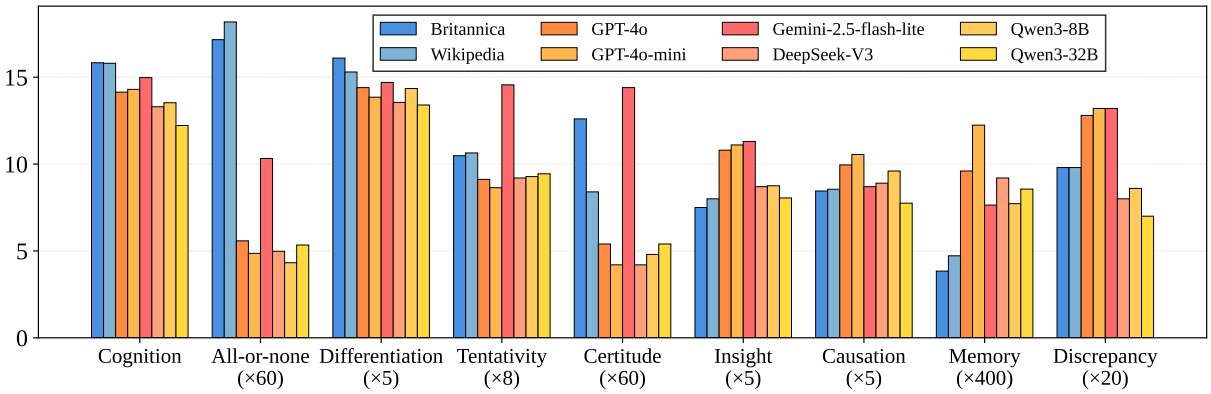


Figure 3: LIWC Cognition features. Some metrics are rescaled for visualization.

and human systems exhibit distinct linguistic patterns. First, human systems consistently achieved higher diversity than LLMs in sentence-level variance metrics (see Table 1). This means human writers exhibit greater differentiation in sentiment and opinion across individual sentences within an article. Humans also prefer explicit citation of factual evidence more than models, as shown in Appendix A.1.1 Fig 4b. Human writers more frequently reference specific stakeholders and entities. However, distinct patterns between humans and LLMs emerge in stakeholders' *role distribution*, as shown in Appendix A.1.1 Fig 4a. 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 1 and more pronounced *angel-devil shift* in Appendix A.1.1 Fig 5a.

Narrative structure. Narrative structures points to a clear discrepancy between the inherent perspective complexity and manifest linguistic features above. As demonstrated in Appendix A.1.1 Fig 5a, 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 Fig 5b, scoring lower on three out of four metrics related to argumentative diversity and the inclusion of distinct viewpoints.

5 Conclusion

As LLMs increasingly mediate access to information on contentious social issues, understanding how they represent controversy is critical for civic discourse. This paper introduced CONTOUR, a systematic framework for evaluating the linguistic presentation of controversial topics, and benchmarked LLM-generated texts against Wikipedia and Encyclopaedia Britannica across 153 politically sensitive topics. Our analysis shows that LLMs exhibit distinct controversy-writing patterns that differ from both community-driven consensus and expert-curated knowledge, reflecting implicit trade-offs between neutrality and false balance. By releasing an open-source implementation of CONTOUR, we aim to support reproducible evaluation and future research on designing LLMs that more responsibly mediate controversial knowledge.

6 Limitations

This study has several limitations. First, our analysis focuses on English-language content and a fixed set of 153 politically sensitive topics, which may

limit the generalizability of our findings to other languages, cultural contexts, or domains of controversy. 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 linguistic features to quantify controversy intensity, which captures how controversy is expressed but not necessarily how readers interpret or are influenced by such representations. Finally, we study a snapshot of LLM behavior under specific prompting 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.

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.

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*.

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.

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.

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

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

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:

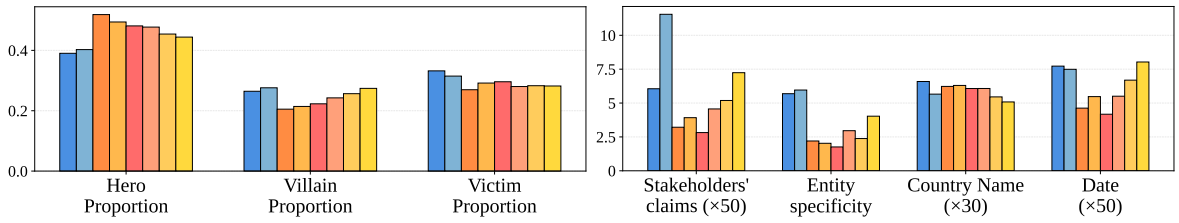
$$\mu_{\text{procon}} = \frac{1}{N}\sum_{i=1}^{N} p_i, \qquad \sigma^2_{\text{procon}} = \text{Var}(p_i),$$

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. Let $\mathbf{e}_i \in \mathbb{R}^d$ denote the embedding of argument i. 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.

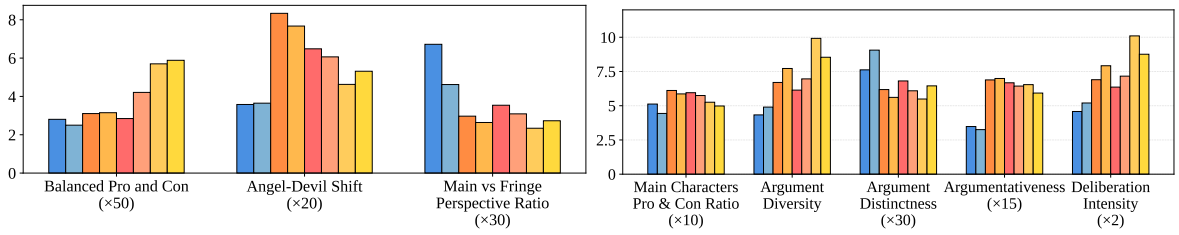
Argument diversity measures the breadth of distinct argumentative perspectives. Let K denote the number of argument clusters and N_{arg} the total number of arguments. Diversity is defined as

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



(a) Role distribution (b) Factual Evidence

Figure 4: Quantitative analysis of Description Indicators dimensions



(a) Contradiction (b) Perspective Complexity

Figure 5: Quantitative analysis of Narrative Structure dimensions

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. Narrative distinctness is defined as

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

where $\bar{d}$ and $d_{\min}$ are the mean and minimum values in $\mathcal{D}$, respectively. This formulation emphasizes 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–Con Percentage. 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. We compute sentence-level variance measures to capture internal heterogeneity, including sentiment variance and embedding variance. Higher values indicate oscillation across evaluative tones or semantic orientations.

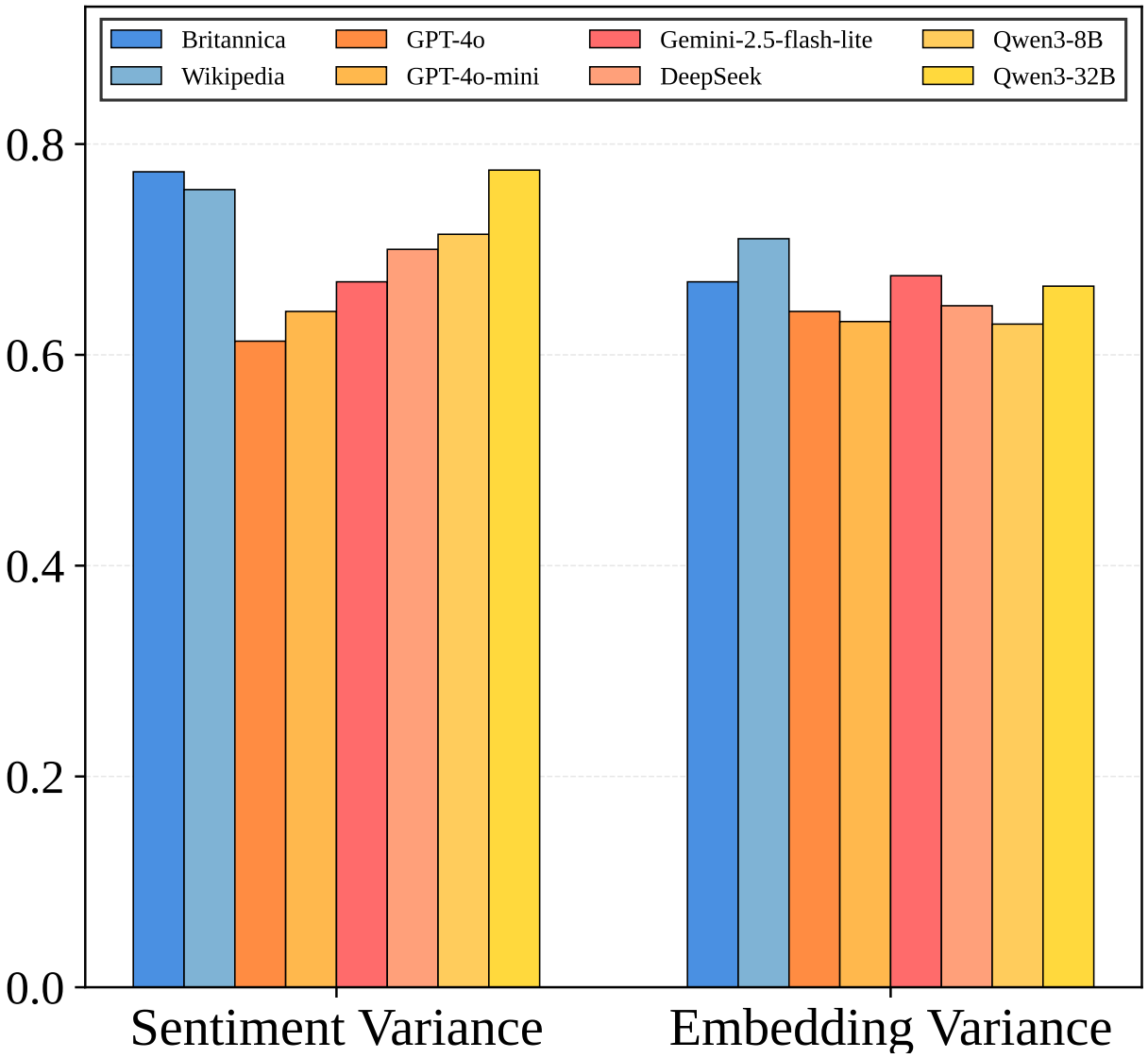


Figure 6: Quantitative Analysis of Description Indicator (Sentence Variance) dimension

Basic Features. Basic features provide coarse-grained controls describing overall textual properties. These include text length (measured in characters), mean sentiment score, and mean embedding magnitude. They are used primarily for normalization and robustness analysis.

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

Cognitive Process Indicators. Using LIWC lexical categories, we extract normalized frequencies of insight, cause, tentativity, certitude, discrepancy, differentiation, and all-or-none expressions. The all-or-none metric captures absolutist language, while discrepancy and differentiation reflect cognitive complexity.

Lexical Diversity and Complexity. We compute part-of-speech variability, vocabulary complexity, type–token ratio (TTR), root TTR (RTTR), corrected TTR (CTTR), Shannon entropy, and WordNet-based polysemy. These metrics quantify lexical richness, ambiguity, and functional flexibility.

Entity Specificity. Entity specificity reflects grounding in concrete actors and institutions. Using named entity recognition, it is defined as

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

where T is the total number of tokens.

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

Stakeholder Claims. 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. Entities are classified into archetypal roles: hero, villain, and victim. 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.

A.3 Human Annotation for Metric Validation

To assess the reliability of our metrics, we conduct a human-annotation study with two annotators. The study evaluates both (i) model-based components in the pipeline and (ii) LLM-based judgments required by several metrics.

Argument Detection (WIBA). We evaluate WIBA's argument detection by having annotators label whether each sliding window contains an argumentative unit, using the same windowing setup as in the automatic analysis. The evaluation is conducted on 76 annotated instances. WIBA achieves an accuracy of **0.6**.

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. These labels support evaluation of stakeholder-claim and pro–con–based metrics.

Balanced Pro–Con Patterns and People Extraction. Annotators additionally label (i) whether a sentence exhibits a *balanced pro–con* construction, and (ii) the extraction of *people roles*, including `human`, `villain`, and `victim`. Balanced pro–con patterns are annotated on 133 sentences, while people extraction is annotated on 78 instances.

Results. We report accuracy as the primary evaluation metric. **People extraction** achieves an accuracy of **0.92** (78 annotated instances). **Stakeholder-claim interactivity**, 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**. These results indicate that both the automatic and LLM-based components used in our framework achieve reliable performance for computing the proposed metrics.

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.