

Discourse-Structural Noise in LLM-Generated Explanations

Raghvi Baloni
University of Trier
Trier, Germany
baloni@uni-trier.de

Achim Rettinger
University of Trier
Trier, Germany
rettinger@uni-trier.de

Abstract

Large language models (LLMs) are increasingly utilized to generate explanations of complex phenomena, yet existing evaluation frameworks assess quality extensively through semantic similarity to human references, a measure blind to discourse organization. We conduct a computational study of 300 ELI5 question pairs, comparing Llama-3.1 8B, Gemma-2 9B, and Mistral-7B against human-written responses using PDTB-grounded connective analysis across five relation types. All three models exhibit elaborative dominance, i.e. Elaborative connectives account for 56-66% of discourse markers versus 38% in human writing, with Causal, Contrastive, and Conditional relations systematically suppressed. This 18-28 percentage point gap persists after connective sense disambiguation (11-20 pp), ruling out measurement artifact. Across three semantic similarity metrics, PDTB connective noise explains less than 8% of shared variance ($r^2 < 0.08$), confirming it captures evaluation signal invisible to existing approaches. We establish discourse-structural noise as a distinct, measurable evaluation dimension for LLM-generated explanations.¹

1 Introduction

The usage of large language models as explanation generators has accelerated across domains ranging from medical question answering to educational tutoring (Fan et al., 2019). Evaluation of these explanations has primarily focused on semantic fidelity (does the model convey the right content?) as measured by reference-based metrics such as ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2020), or by LLM-as-judge frameworks such as G-Eval (Liu et al., 2023).

These frameworks do not measure *discourse structure*, the rhetorical organization of an expla-

nation, explaining how claims relate to evidence, how contrasts are drawn, how conditions are introduced. Discourse theory has long established that the same propositional content can be organized in structurally very different ways, and that structural variation affects how readers process and evaluate text (Mann and Thompson, 1988; Prasad et al., 2008). If LLMs systematically produce explanations with a different discourse structure from human writers, this divergence is latent to current evaluation practice.

We introduce the concept of **discourse-structural noise** as the systematic deviation of LLM explanation structure from the discourse organization patterns present in human-authored explanations from the same communicative task. The human connective distribution constitutes the expected structural signal for explanation text in communicative context. Systematic deviation from it, however internally consistent across models, constitutes noise relative to that reference standard. This framing does not imply randomness, rather, the cross-model consistency of the deviation makes it diagnostically meaningful and systematically measurable. We operationalize this through PDTB-grounded connective analysis, which decomposes markers into five relation types (Temporal, Causal, Contrastive, Elaborative, Conditional) and measures the distributional distance between LLM and human profiles.

Our main findings are: (i) all three LLMs tested exhibit elaborative dominance, an 18-28 pp overuse of Elaborative connectives paired with systematic suppression of argumentative relation types, robust to connective sense disambiguation; (ii) this pattern is consistent across distinct model families, indicating it is a property of LLM explanation generation broadly rather than any individual model; (iii) PDTB connective noise explains less than 8% of shared variance with three semantic similarity metrics ($r^2 < 0.08$ across ROUGE-L, DistilBERT,

¹<https://github.com/RaghviBaloni/discourse-structural-noise>

and MiniLM), confirming it constitutes a distinct and currently unmeasured evaluation dimension.

2 Related Works

Discourse coherence evaluation The Penn Discourse Treebank (PDTB) (Prasad et al., 2008) provides the canonical framework for annotating discourse relations via connective expressions. Entity grid models (Barzilay and Lapata, 2008) capture local coherence through entity role transition across adjacent sentences. More recent neural approaches (Liu and Strube, 2025) extend coherence modeling with discourse relation supervision. These frameworks were developed and validated on human-authored newswire and academic text while their behavior on LLM-generated conversational explanation text remain uncharacterized.

LLM output quality evaluation BERTScore and ROUGE (Zhang et al., 2020; Lin, 2004) measure reference-based semantic similarity. G-Eval and GPT-Score (Liu et al., 2023; Fu et al., 2023) leverage LLM as judges of holistic quality dimensions including coherence, fluency, and relevance. (Wiegreffe and Marasović, 2021) survey explanation evaluation broadly and identify faithfulness and plausibility as dominant criteria, neither of which targets discourse-structural organization. Our work fills this gap by introducing a computationally grounded, theoretically motivated measure of structural divergence.

LLM-generated text characteristics Prior work has characterized LLM outputs in terms of lexical diversity, factual consistency, and stylistic properties (Shaib et al., 2024; Laban et al., 2021; Muñoz-Ortiz et al., 2024). Discourse-structural properties of LLM-generated explanations have not been systematically audited. Existing evaluation efforts either rely on costly human annotation (Wiegreffe and Marasović, 2021) or holistic LLM-as-judge frameworks, neither of which provides systematic, reproducible discourse-structural analysis at scale.

The work of Wang et al. (Wang et al., 2022) is directly complementary to ours, they study exemplification as a discourse role in long-form QA and show that model-generated answers under-use certain discourse functions relative to human writers, consistent with our finding that LLMs suppress argumentative relation types. We extend this observation from the document-level perspective to the connective-level distributional perspective, pro-

viding a finer-grained and automatically scalable measure of structural divergence.

3 Methodology

3.1 Dataset and Sampling

We sample 300 question-explanation pairs from the ELI5 dataset (Fan et al., 2019), a long-form question-answering benchmark sourced from r/explainlikeimfive subreddit. Each item consists of a community-posed question and one or more human-authored explanations scored by community upvotes. We retain the highest-scored answer per question and apply three quality filters: a minimum of 50 and maximum of 500 words, at least three sentences, and a minimum community score of 2. These thresholds ensure explanations are discourse-rich enough to support structural analysis while avoiding multi-topic responses that confound cross-explanation comparison.

3.2 Models

We evaluate three open-weight instruction-tuned language models spanning a range of parameter scales, Gemma-2 9B, Llama-3.1 8B, and Mistral-7B v0.3 (Team, 2024; Grattafiori et al., 2024; Jiang et al., 2023). For each of the 300 questions, we generate one explanation per model using vLLM (Kwon et al., 2023) with temperature 0.7, top-p 0.9, and a maximum of 512 output tokens. All models receive an identical system prompt instructing them to provide a clear, structured explanation, without specifying any discourse organization. This ensures that any structural patterns we observe are emergent properties of the models’ generation behavior, not artifacts of explicit instruction.

3.3 Discourse-Structural Noise

We define **discourse-structural noise** as the systematic deviation of LLM-generated explanations from the discourse organization patterns present in human-written explanations for the same communicative task. We operationalize this along two dimensions:

PDTB Connective Analysis Based on the PDTB 2.0 sense hierarchy, we categorize discourse connectives into five relation types: Temporal, Causal, Contrastive, Elaborative, and Conditional. We compile lexicon of 120+ connective expressions across these five types, with multi-word expressions matched before their sub-strings to prevent spurious overlap. We categorize connectives

following the PDTB 2.0 coarse-grained sense hierarchy: TEMPORAL (eg. *then, after, when*), CONTRASTIVE (eg. *however, although, yet*), CONDITIONAL (e.g. *if, unless, provided that*). Multi-word expressions (e.g. *even though*) are matched before their constituent sub-strings (e.g. *though*) to prevent spurious double-counting. For each text, we extract all connective instances and compute a normalised distribution over the five types. We then measure PDTB connective noise as the symmetric KL divergence between each LLM’s connective distribution and the human distribution across the 300 matched samples.

Let $\tilde{P}(t) = (P(t) + \varepsilon) / \sum_{t'} (P(t') + \varepsilon)$ denote the smoothed distribution with $\varepsilon = 10^{-8}$. We measure PDTB connective noise as the symmetric KL divergence between the smoothed LLM and human distributions:

$$D_{\text{sym}}(\tilde{P}_{\text{LLM}} \| \tilde{P}_{\text{H}}) = \frac{1}{2} \left[\sum_t \tilde{P}_{\text{LLM}}(t) \log \frac{\tilde{P}_{\text{LLM}}(t)}{\tilde{P}_{\text{H}}(t)} + \sum_t \tilde{P}_{\text{H}}(t) \log \frac{\tilde{P}_{\text{H}}(t)}{\tilde{P}_{\text{LLM}}(t)} \right] \quad (1)$$

where t indexes the five relation types.

Semantic similarity baseline To test whether discourse-structural noise captures evaluation signal orthogonal to existing semantic metrics, we compute a sentence-embedding cosine similarity score for each LLM response against its paired human explanation, using DistilBERT-base-uncased (Sanh et al., 2020) with mean pooling over non-padding tokens. We report Spearman rank correlation (ρ) between per-sample KL divergence scores and semantic similarity scores.

4 Results

4.1 LLMs Exhibit Consistent Elaborative Dominance

Figure 1 shows the mean normalised connective type distribution for human explanations and all three LLMs. Human explanations exhibit a relatively balanced discourse profile, with Causal connectives accounting for the largest share (25.0%), followed by Elaborative (37.9%), Temporal (14.7%), Contrastive (13.0%), and Conditional (9.1%).

LLM-generated explanations show a significantly different structural profile. Across all three models, Elaborative connectives dominate, accounting for 56.2% (Mistral) to 66.3% (Llama-3.1)

of all connective instances, an increase of 18-28 percentage points over the human baseline. This shift is accompanied by systematic suppression of all argumentative relation types - Causal connectives drop from 25.0% in humans to 14.0-17.9% across models, Contrastive from 13.0% to 5.6-7.7% and Conditional from 9.1% to 2.2-3.2%.

Source	Total	TEMP	CAUS	CONT	ELAB	COND	KL $\uparrow$
Human	11.2	14.7	25.0	13.0	37.9	9.1	—
Gemma-2 9B	18.9	14.7	15.6	5.6	60.8	3.2	2.39
Llama-3.1 8B	25.1	10.8	14.0	6.3	66.3	2.2	2.54
Mistral-7B	20.6	11.4	17.9	7.7	56.2	3.2	2.34

Table 1: Mean total connective count and normalised proportion (%) per relation type. KL = symmetric KL divergence from human profile; $\uparrow$ = higher means more structural noise.

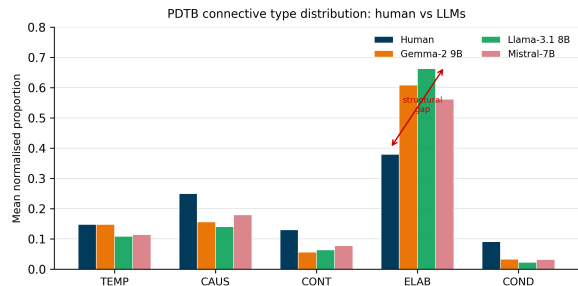


Figure 1: Mean normalised PDTB connective type distribution across 300 matched ELI5 pairs. All three LLMs show elaborative dominance (ELAB: 56-66%) versus a balanced human profile. Error bars show one standard deviation.

This pattern is consistent across all three model families tested. Gemma-2, Llama-3.1, and Mistral-7B, architecturally distinct models trained on different data mixtures, converge on the same discourse structure where elaborative addition replaces argumentative reasoning as the dominant connective mode. This consistency between models indicates that elaborative dominance is broadly a structural property of LLM explanation generation, rather than an idiosyncrasy of any particular model or training procedure.

Table 1 further shows that LLMs generate substantially more total connective instances than humans (19-25 vs 11 per text), yet this increase is concentrated entirely in elaborative connectives. The models are discourse-monotone instead of discourse-sparse, defaulting to a single relational register regardless of the communicative demands of the question. Bootstrap 95% confidence intervals confirm that the human-LLM gap on all four

argumentative relation types is non-overlapping across all models (Table 2 in Appendix), establishing the finding as statistically robust. Furthermore, surface features including response length explain only 1.9% of KL divergence variance ($R^2=0.019$; Appendix E), ruling out length and complexity as confounds.

4.2 Robustness to Connective Sense Disambiguation

A potential confound is connective polysemy, high frequency markers such as *since*, *as*, *while*, and *and* carry multiple sense across PDTB relation types. To assess whether elaborative dominance is an artifact of mis-classification, we re-ran the connective analysis with dependency parse sense disambiguation using spaCy (en_core_web_trf), applying heuristics for the five most polysemous connectives.

The disambiguation reclassified 393, 568, 1112, and 819 connective instances for human, Gemma-2, Llama-3.1, and Mistral-7B responses respectively. After reclassification, Elaborative proportions shift to 30.9% for Human and 51.0%, 49.8%, 42.3% for Gemma-2, Llama-3.1, and Mistral-7B respectively. The human-LLM gap narrows from 18-28 to 11-20 percentage points but is not eliminated under any model. Notably, Llama-3.1 had the most connective instances reclassified yet maintained the highest elaborative proportion after disambiguation, indicating that its dominance is driven by unambiguously elaborative markers (*furthermore*, *additionally*, *also*) that are not subject to sense ambiguity. We conclude that polysemy partially explains but does not account for elaborative dominance.

4.3 Discourse-Structural Noise is Substantively Independent of Semantic Similarity

Figure 2 shows Spearman ρ between PDTB KL divergence and three semantic similarity metrics, ROUGE-L, DistilBERT cosine similarity, and MiniLM cosine similarity (Wang et al., 2020; Lin, 2004; Sanh et al., 2020), for each model. Full numerical results are reported in Appendix Table 3.

All correlations are negative and small in magnitude. The largest observed correlation (Llama-3.1 $\times$ ROUGE-L: $\rho=-0.27$, $r^2=0.07$) explains less than 8% of shared variance. Across all other model-metric combination, $r^2<0.05$, confirming that the two measures capture largely independent signal about explanation quality. Statistically significant

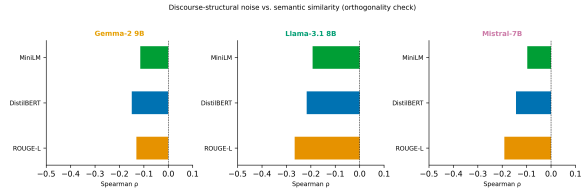


Figure 2: All bars extend leftward (negative ρ), confirming an inverse relationship; short bar lengths confirm that structural noise is largely independent of semantic similarity.

correlations ($p<0.05$) appear in the majority of combinations, as expected at $N=300$ for non-zero population correlations. We report effect sizes (r^2) as the relevant quantity.

The practical implication is direct. Mean semantic similarity scores are nearly identical across all three models (DistilBERT: 0.887-0.902, range <0.02), whereas KL divergence shows model-specific variation (2.34-2.54). An evaluation framework relying on semantic similarity alone would observe no meaningful difference between models and would systematically miss the structural deviation we document.

5 Conclusion

We introduce discourse-structural noise as a previously uncharacterized measurable failure mode of LLM explanation generation. Through a PDTB-grounded audit of 300 ELI5 question pairs across three open-weight model families, we identify consistent elaborative dominance. Dependency-parse sense disambiguation of the five most polysemous connectives reduces the elaborative gap to 11-20 pp but does not eliminate it. PDTB connective noise explains less than 8% of shared variance with three independent semantic similarity metrics, confirming it captures evaluation signal invisible to current metrics. Surface features including response length account for only 1.9% of KL divergence variance, ruling out length and lexical complexity as confounding explanations.

These findings have direct implications for LLM evaluation practice, discourse structure should be treated as an independent quality dimension alongside semantic fidelity and factual correctness. We release our PDTB connective evaluator, corpus of 300 human-LLM explanation pairs and code to support reproducibility.

Limitations

Connective lexicon and implicit relations Our rule-based lexicon may miss implicit discourse relations not signaled by surface connectives. Several high-frequency markers (*since, as, while, and*) carry multiple senses across PDTB relation types. We partially address this in §4.2 through dependency-parse sense disambiguation of the five most polysemous connectives (spaCy en_core_web_trf), which reduces but does not eliminate the elaborative dominance gap (11–20pp post-disambiguation vs. 18–28pp baseline). Implicit discourse relations, those not signaled by overt connectives, remain uncaptured. The observed suppression of Causal and Contrastive relations may partly reflect a shift toward implicit signaling rather than genuine argumentative impoverishment.

Prompt and length confounds Our system prompt requests a “clear, structured explanation,” phrasing that may encourage additive enumeration. However, surface features including response length ($\rho = -0.008$, $p = 0.810$), sentence count ($p = 0.947$), and model identity together explain only 1.9% of KL divergence variance ($R^2 = 0.019$; Appendix E), providing empirical evidence against a length or prompt artifact account.

Scope This study covers three English-language instruction-tuned models and one domain (ELI5 long-form QA). Generalization to other model families, languages, and explanation domains remains to be established.

References

- Regina Barzilay and Mirella Lapata. 2008. [Modeling local coherence: An entity-based approach](#). *Computational Linguistics*, 34(1):1–34.
- Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. [ELI5: Long form question answering](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.
- Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. [Gptscore: Evaluate as you desire](#). *Preprint*, arXiv:2302.04166.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *Preprint*, arXiv:2310.06825.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. [Efficient memory management for large language model serving with pagedattention](#). In *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP ’23*, page 611–626, New York, NY, USA. Association for Computing Machinery.
- Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2021. [Summac: Re-visiting nli-based models for inconsistency detection in summarization](#). *Preprint*, arXiv:2111.09525.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Wei Liu and Michael Strube. 2025. [Discourse relation-enhanced neural coherence modeling](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4748–4762, Vienna, Austria. Association for Computational Linguistics.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. [G-eval: NLG evaluation using gpt-4 with better human alignment](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- William Mann and Sandra Thompson. 1988. [Rhetorical structure theory: Toward a functional theory of text organization](#). *Text - Interdisciplinary Journal for the Study of Discourse*, 8:243–281.
- Maximilian Maurer. 2026. [elfen: A python package for efficient linguistic feature extraction for natural language datasets](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 61–74, Rabat, Morocco. Association for Computational Linguistics.
- Alberto Mu  oz-Ortiz, Carlos G  mez-Rodr  guez, and David Vilares. 2024. [Contrasting linguistic patterns in human and llm-generated news text](#). *Artificial Intelligence Review*, 57(10).

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Milt-sakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. *The Penn Discourse TreeBank 2.0*. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. *Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter*. *Preprint*, arXiv:1910.01108.

Chantal Shaib, Yanai Elazar, Junyi Jessy Li, and Byron C Wallace. 2024. *Detection and measurement of syntactic templates in generated text*. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6416–6431, Miami, Florida, USA. Association for Computational Linguistics.

Gemma Team. 2024. *Gemma*.

Shufan Wang, Fangyuan Xu, Laure Thompson, Eunsol Choi, and Mohit Iyyer. 2022. *Modeling exemplification in long-form question answering via retrieval*. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2079–2092, Seattle, United States. Association for Computational Linguistics.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. *Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers*. *Preprint*, arXiv:2002.10957.

Sarah Wiegrefe and Ana Marasović. 2021. *Teach me to explain: A review of datasets for explainable natural language processing*. *Preprint*, arXiv:2102.12060.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. *Bertscore: Evaluating text generation with bert*. *Preprint*, arXiv:1904.09675.

A Appendix

A.1 Connective Density by Type

B Bootstrap Confidence Intervals

Source	TEMP	CAUS	CONT	ELAB	COND
Human	0.147 $\pm$ 0.019	0.250 $\pm$ 0.020	0.130 $\pm$ 0.014	0.379 $\pm$ 0.022	0.091 $\pm$ 0.014
Gemma-2 9B	0.147 $\pm$ 0.011	0.156 $\pm$ 0.011	0.056 $\pm$ 0.007	0.608 $\pm$ 0.017	0.032 $\pm$ 0.006
Llama-3.1 8B	0.108 $\pm$ 0.011	0.140 $\pm$ 0.011	0.063 $\pm$ 0.009	0.663 $\pm$ 0.019	0.022 $\pm$ 0.005
Mistral-7B	0.114 $\pm$ 0.011	0.179 $\pm$ 0.013	0.077 $\pm$ 0.009	0.562 $\pm$ 0.024	0.032 $\pm$ 0.007

Table 2: Mean normalised connective type proportions with 95% bootstrap confidence intervals (2,000 resamples). Subscript $\pm$ shows the CI half-width. Non-overlapping intervals between human and all LLMs on ELAB and CAUS confirm the elaborative dominance finding is statistically robust.

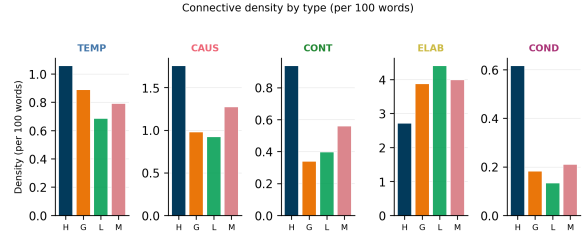


Figure 3: Mean connective density (instances per 100 words) by relation type for each source. Human explanations show higher density in Causal and Contrastive types relative to all LLMs. LLMs compensate with higher absolute Elaborative density, confirming the elaborative dominance finding at the rate level rather than only at the distributional level.

C Discourse-Structural Noise is Orthogonal to Semantic Similarity

Model	ROUGE-L		DistilBERT		MiniLM	
	ρ	p	ρ	p	ρ	p
Gemma-2 9B	-0.131	0.024*	-0.149	0.010*	-0.114	0.048*
Llama-3.1 8B	-0.266	0.000*	-0.217	0.000*	-0.193	0.001*
Mistral-7B	-0.191	0.001*	-0.143	0.013*	-0.097	0.092

Table 3: Spearman ρ between PDTB KL divergence and three semantic similarity metrics across 300 ELI5 pairs per model. Near-zero correlations across all three metrics confirm that discourse-structural noise captures evaluation signal orthogonal to existing semantic evaluation approaches. * $p < 0.05$.

D Qualitative Examples

Table 4 presents representative examples from the highest-KL quartile of Llama-3.1 responses, illustrating how elaborative dominance manifests in practice.

E Surface Feature Analysis and Confound Control

To test whether elaborative dominance is an artifact of response length, lexical complexity, or model identity rather than a genuine structural property, we extracted eight surface linguistic features for all 1,200 texts (300 questions $\times$ 4 sources: human, Gemma-2, Llama-3.1, Mistral-7B) using ELFEN (Maurer, 2026), a multilingual linguistic feature extraction library, with a spaCy backbone.

Features extracted Token count, sentence count, tokens per sentence, average word length, type-token ratio (TTR), MTLT, Flesch reading ease, and token entropy.

Question	Human explanation (excerpt)	Llama-3.1 explanation (excerpt)
what bleach (clorox) does to organic things.	Bleach is most commonly sodium hypochlorite. The hypochlorite ion ClO^- is a very strong oxidizing agent, meaning that it steals the electrons off of other things. This is why it is such a good disinfectant, because ionizing (stealing electrons) from an organism's body is deadly. For things with a tiny body like bacteria, this is even more quickly fatal. <i>Connectives: CAUSE - because</i>	Bleach, specifically Clorox, is a strong chemical disinfectant that contains sodium hypochlorite (NaOCl) as its active ingredient. <u>When</u> bleach comes into contact with organic matter, it undergoes a series of chemical reactions that break down the organic material. Here's a step-by-step explanation of what bleach does to organic things: 1. Penetration : When bleach is applied to an organic surface, it penetrates the material, allowing the sodium hypochlorite molecules to come into contact with the organic matter. 2. Oxidation : The sodium hypochlorite molecules in bleach are strong oxid... <i>Connectives: CONT - however, ELAB - specifically, also, TEMP - then, when</i>
Who was allowed to vote in 19th century Britain?	In the 1800s suffrage was expanded about every 2 or 3 decades. At the start of the 19th century, probably 5% of the British population could vote. This right was reserved to wealthy landowners. The Reform Act of 1832 changed this. Voting rights were now spread to bankers, merchants, etc. The Reform Act of 1867 was the next time voting rights were expanded. <u>If</u> you were a male, owned a home, or rented a home of at least a certain value you were permitted to vote. The Reform Act of 1884 extended suffrage to those owning or renting land in the countryside, outside of town limits. Around the ti... <i>Connectives: COND - if</i>	To understand who was allowed to vote in 19th century Britain, we need to break down the historical context and the various reforms that took place during this period. Early 19th century: Limited Voting Rights (1800s) In the early 19th century, voting rights in Britain were restricted to a small group of people known as the "Electoral Classes." These included: 1. Male property owners : Free men who owned property, such as land or a house, worth a certain amount, were eligible to vote. This meant that only those with a certain level of wealth and social status could participate in th... <i>Connectives: ELAB - in addition, also</i>

Table 4: Representative high-KL divergence examples from Llama-3.1 8B. Human explanations use causal and contrastive connectives to build logical arguments; LLM explanations substitute elaborative connectives (*furthermore, additionally, also*), adding information without establishing argumentative relations.

Regression We fit an OLS regression with KL divergence as the outcome and the eight surface features plus model identity dummy variables as predictors. Results are reported in Table 5.

Predictor	ρ with KL	p
Token count	-0.008	0.810
Sentence count	+0.002	0.947
Tokens per sentence	(collinear)	
Type-token ratio	-0.084	0.012*
Flesch reading ease	+0.098	0.004*
Token entropy	+0.025	0.464
Model: Gemma-2	+0.025	0.454
Model: Llama-3.1	+0.017	0.611
Model: Mistral-7B	-0.043	0.204
$R^2 = 0.019$, $R^2_{\text{adj}} = 0.010$, $N = 889$		

Table 5: OLS regression of PDTB KL divergence on surface features and model identity dummies. Spearman ρ between each predictor and KL divergence; R^2 for the full model. Response length and sentence count are non-significant ($p > 0.80$); the full model explains only 1.9% of KL variance. * $p < 0.05$.

Interpretation The full model explains only $R^2=0.019$ of KL divergence variance. Response length is not a significant predictor ($\rho=-0.008$, $p=0.810$), nor is sentence count ($p=0.947$). The two significant individual predictors, TTR ($\rho=-0.084$) and Flesch reading ease ($\rho=+0.098$), have trivially small effect sizes. Model identity dummy variables are non-significant after controlling for surface features (all $p>0.20$), consistent with the observation that all three models show high structural noise relative to the human baseline. These results confirm that elaborative dominance is not reducible to surface-level response characteristics and reflects genuine discourse-structural divergence.