

A Mirage of Coherence: How Metaphor Impacts Language Models’ Discourse Coherence Assessment

Wanting Ning¹ Brian Choi² Eun Jin Paek³ Hyeju Jang¹

¹Indiana University Indianapolis ²Carnegie Mellon University

³The University of Tennessee Health Science Center

{ningw,hyejuj}@iu.edu bchoi3@andrew.cmu.edu epaek@uthsc.edu

Abstract

Recent approaches to discourse coherence assessment increasingly rely on language models, but it remains unclear whether they evaluate underlying discourse structure or merely surface-level lexical patterns. We use metaphor as a controlled lexical-semantic and stylistic probe. Using a paraphrase framework with automatic quality control and small-scale human validation, we generate metaphorical rewrites that hold propositional content fixed while varying local lexical realization. Across nine models spanning three architecture families, we find that models paradoxically reward figurative language, assigning higher coherence scores to texts with novel metaphors in nearly all model-domain conditions. Their internal representations appear to encode whether a rewrite changed the meaning, suggesting that the relevant information is present but is not what drives the judgment. Coherence judgments are thus influenced by local stylistic variation, indicating that current models often evaluate linguistic realization rather than relying on discourse-level reasoning alone.

1 Introduction

Discourse coherence assessment, which determines whether a text forms a coherent and comprehensible whole, is a fundamental problem in natural language processing, with applications ranging from essay scoring to mental health monitoring (Somasundaran et al., 2014; Iter et al., 2018; Just et al., 2019). Recent approaches increasingly leverage language models as a foundation for this evaluation, e.g., (Liu & Strube, 2025a;b). However, a deeper question remains unresolved: *when a model judges a text as coherent, is it responding to underlying semantic and discourse structure, or merely to surface-level lexical patterns that correlate with coherent text?* This distinction matters because coherence judgments may shift in response to local variation in linguistic realization even when the underlying discourse content and organization remain unchanged.

Metaphor provides a useful probe for disentangling these possibilities. By substituting literal vocabulary with figurative expressions from distinct semantic domains, metaphors introduce extreme variation in local linguistic realization while leaving the propositional content and discourse organization largely unchanged. Furthermore, recent work shows that language-model surprisal correlates with metaphor novelty (Jankowiak, 2020; Momen et al., 2026). This makes metaphor a natural test case for asking whether coherence models rely on deeper semantic and discourse information, or instead depend heavily on local lexical predictability and related surface cues.

Despite extensive research on discourse coherence and metaphor processing, to our knowledge, no prior work has examined how automated coherence judgments are affected by metaphorical language. Prior work shows that fine-tuned pretrained language models can achieve high accuracy on proxy coherence tasks, such as the shuffle test and sentence intrusion detection, without necessarily yielding robust or linguistically meaningful models of coherence (Laban et al., 2021; Shen et al., 2021). However, these limitations have primarily been demonstrated on such artificial tasks, leaving model behavior on human-annotated co-

Original text	Meaning-preserving metaphorical rewrite
... He helped the man who was <i>near death</i> , saved his life, and wanted no recognition for it. ...	... He helped the man who was <i>within a whisper of death</i> , saved his life, and wanted no recognition for it. ...
... It really <i>hurts</i> me when someone needs help, and no one around them will try to help because they are too busy with their own agenda to come to their aid. ...	... It really <i>gnaws at</i> me when someone needs help, and no one around them will try to help because they are too busy with their own agenda to come to their aid. ...

Table 1: Examples of controlled metaphorical rewrites. Only the targeted expression is changed, while the surrounding discourse context and propositional content are preserved.

herence judgments largely unexplored. At the same time, related work finds that metaphor can substantially impair performance on downstream NLP tasks (Li et al., 2024), suggesting that metaphorical variation may systematically affect model behavior. Yet its specific impact on discourse coherence evaluation remains unknown.

To address these gaps, we organize our investigation around the following research questions: **Main-RQ:** Do language models treat metaphor as a coherence signal or as a coherence disruption? **RQ1 (Judgment):** Does the presence of metaphor systematically alter language models’ coherence judgments? **RQ2 (Uncertainty):** Does metaphor affect the model’s predictive uncertainty in coherence evaluation? **RQ3 (Calibration):** When metaphor changes model confidence, do those confidence estimates remain trustworthy? **RQ4 (Modulation):** Are these effects modulated by metaphor density, novelty, and explicitness?

We investigate these questions on the Grammarly Corpus of Discourse Coherence (GCDC), which contains real-world texts from four discourse domains with expert ratings (Lai & Tetreault, 2018). We employ two complementary experimental settings. In **Experiment 1 (Observational Analysis of Metaphor Effects)**, we analyze the original texts to test whether naturally occurring metaphor use is associated with differences in model behavior across domains. In **Experiment 2 (Controlled Intervention on Coherence Judgments)**, we construct controlled metaphor rewrites that vary novelty (Table 1) while keeping the rest of the discourse as stable as possible. This process produces rewritten examples that remain closer in meaning and fluency to their source texts, allowing cleaner comparisons between original and rewritten versions. To quantify these effects, we analyze shifts in decision margin (i.e., the gap between the top two class probabilities), predictive entropy (Shannon, 1948), and probability calibration (Guo et al., 2017) derived from the models’ output distributions. We evaluate nine language models spanning three architecture families, together with two state-of-the-art (SOTA) coherence prediction models (Liu & Strube, 2025a;b). Beyond these behavioral evaluations, we further use representation-level analyses to examine whether models distinguish meaning-preserving metaphorical variation from changes that alter the underlying semantics.

Our main finding is simple: metaphor tends to make coherence look better than it is. When meaning and discourse organization are held fixed by design, novel metaphorical variation shifts model-assigned coherence upward in most baseline model–domain conditions ($d \approx 0.1\text{--}0.3$), with low- and medium-coherence texts more often absorbed into the *High* category. The effect is modest in magnitude but consistent in direction, which is what makes it consequential for borderline decisions and aggregated scores. Effects on uncertainty and calibration are smaller and more heterogeneous. Notably, the sensitivity persists in reproduced discourse-structured coherence models, though in a strongly domain-dependent form, suggesting that explicit discourse representations reshape rather than eliminate the influence of local linguistic realization. The four-condition representation analysis further suggests that semantic compatibility remains reflected in the model’s internal representations: in Flan-T5-Large, meaning-changing rewrites diverge most clearly from the original at the final encoder layer, whereas the meaning-preserving metaphor remains closer. Overall, these results indicate that coherence judgments are shaped not only by discourse-level organization, but also by the lexical and stylistic form through which the same content is expressed.

This paper makes three primary contributions: (1) We present the first systematic study of how metaphorical variation affects automated discourse coherence judgments on human-annotated data. (2) We introduce a controlled rewrite framework for varying metaphorical realization while holding propositional content and discourse organization fixed by design. (3) We combine behavioral, calibration, and representation-level analyses to characterize when and how coherence models respond to such variation.

2 Related Work

Discourse Coherence Assessment. Early computational approaches modeled coherence through entity continuity and lexical similarity (Barzilay & Lapata, 2008; Foltz et al., 1998). The Entity-Grid model (Barzilay & Lapata, 2008) became a foundational benchmark by representing grammatical role transitions of discourse entities across sentences. Neural approaches subsequently learned distributed representations over such grids (Nguyen & Joty, 2017; Joty et al., 2018), while graph-based methods extended coherence modeling to structural document representations (Guinaudeau & Strube, 2013; Liu et al., 2023). More recent work has shown that fine-tuned transformers can achieve high shuffle-test accuracy without necessarily modeling coherence in a linguistically meaningful sense (Laban et al., 2021), and that discourse relation features provide complementary signal beyond entity-based representations (Liu & Strube, 2025a;b). Likewise, sentence intrusion detection has shown strong in-domain results but limited cross-domain generalization (Shen et al., 2021). In contrast to prior work that predominantly focused on improving prediction performance, our research investigates how controlled variation in linguistic realization (specifically through metaphor) affects coherence judgments on human-annotated data. This allows us to probe whether current models capture discourse-level understanding or sensitivity to local stylistic cues.

Metaphor in Discourse and Language Models. Metaphor has long been studied in discourse contexts (Asher & Lascarides, 2001; Sanford, 2008; Jang et al., 2015; 2017), and recent corpus-based work has more directly examined which discourse conditions are associated with metaphoricity and with the use of metaphorical rather than literal alternatives in context (Piccirilli & Im Walde, 2022a;b). A distinct line of research investigates what language models encode about metaphor. Probing studies show that pre-trained language models capture metaphor-relevant information, especially in their intermediate layers (Aghazadeh et al., 2022). Despite this, recent evaluation work indicates that metaphor remains challenging for current language models. MUNCH, a dedicated benchmark, was introduced to assess large language models’ metaphor understanding capabilities (Tong et al., 2024), and subsequent work reports measurable degradation on several downstream tasks, including machine translation, question answering, and natural language inference (Li et al., 2024). Our work complements these lines of research by shifting the focus from metaphor detection and understanding to its interaction with higher-level discourse evaluation. Specifically, we investigate how controlled metaphorical variation affects language models’ discourse coherence judgments and their associated probability-space behavior.

3 Methodology

Figure 1 presents an overview of our three-phase pipeline, which covers metaphor annotation, controlled rewriting, quality control, and the subsequent analysis of how metaphor features affect language models’ coherence judgments.

3.1 Task Formulation

We formulate discourse coherence evaluation as a classification task for multi-sentence texts. Given a discourse $T = \{s_1, \dots, s_n\}$, where each s_i represents a sentence, the goal is to assess the degree to which the text is logically connected and structurally well organized. Coherence is rated on a three-point ordinal scale, denoted $y \in \{\text{Low}, \text{Medium}, \text{High}\}$. In our setting, a language model produces not only a discrete prediction $\hat{y}$ but also a probability distribution $p(y | T)$ over the label space (For details of the experimental setup,

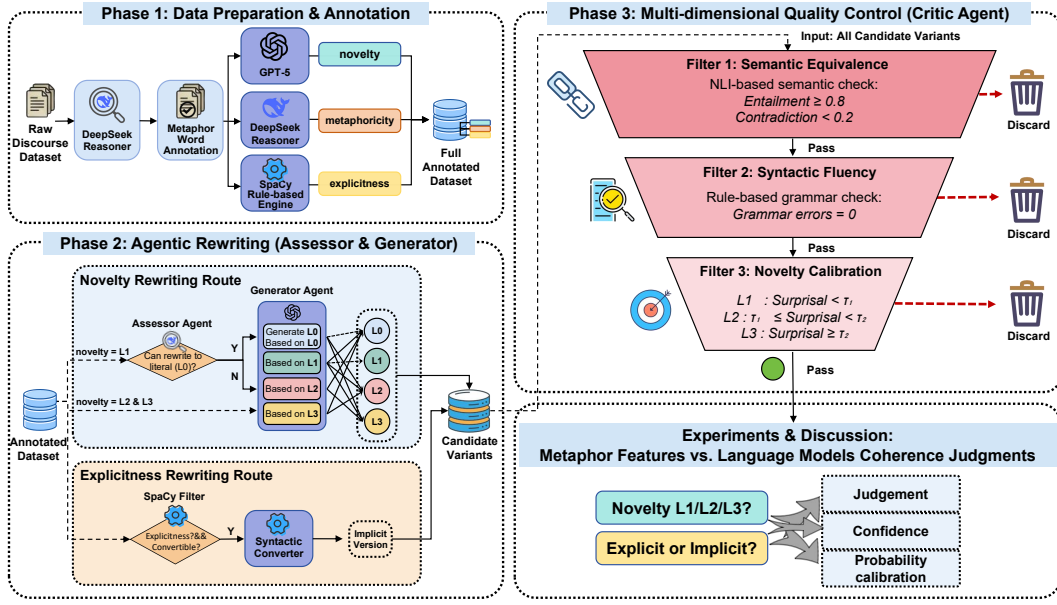


Figure 1: Overview of the pipeline for metaphor annotation, controlled rewriting, and quality control.

see the Appendix A), which allows us to examine coherence judgments beyond categorical decisions, including confidence, uncertainty, and calibration behavior.

3.2 Dataset

We use the Grammarly Corpus of Discourse Coherence (GCDC) (Lai & Tetreault, 2018), a benchmark for discourse coherence assessment covering four domains: *clinton*, *enron*, *yahoo*, and *yelp*. Each document is annotated with a three-way human coherence label. We use the expert-consensus labels in our experiments. We follow the standard 10-fold cross-validation splits provided with the dataset. Additional dataset details, including domain descriptions, document filtering, and annotation criteria, are provided in Appendix B.

3.3 Metaphor Annotation

Metaphor Identification. We adopt the MIPVU framework (Steen et al., 2019), a widely used protocol for identifying metaphors by contrasting a word’s contextual meaning with its basic sense (often coming from a dictionary). Building on this methodology, we design a token-level prompting strategy that instructs the model to evaluate each lexical unit within its sentential context. Prior to large-scale annotation, we calibrate the prompt on samples from the VU Amsterdam Metaphor Corpus (VUAMC), which was manually annotated according to MIPVU, refining the instruction wording to better align the LLM-based procedure with token-level decision criteria. Full prompt templates and calibration details are provided in Appendix C.

Metaphor Features. We annotate three metaphor features per token. *Novelty* captures the degree to which a metaphorical expression departs from conventional usage (Bowdle & Gentner, 2005; Do Dinh et al., 2018). We assign discrete novelty levels $\ell(m) \in \{\text{low, medium, high}\}$ via a prompting-based classifier benchmarked on VUAMC novelty annotations (Do Dinh et al., 2018), achieving a macro-F1 of approximately 90% (see Appendix D). *Metaphoricity* captures the degree to which a token is perceived as metaphorical rather than literal, independently of how novel that metaphor use is; it is annotated as a three-level variable which is the same as $\ell(m)$ and serves as an auxiliary control variable

rather than a directly manipulated dimension.¹ *Explicitness* captures whether the metaphorical comparison is overtly marked; we detect explicit metaphors via spaCy-based (Honnibal et al., 2020) syntactic parsing over a predefined set of comparison markers (*like*, *as*, *as if*, *similar to*, and lemma variants of *resemble*), with $e(m) = 1$ indicating an explicit construction and $e(m) = 0$ otherwise.

Discourse-level Features. From the token-level annotations, we derive four document-level metaphor features used in subsequent analyses: sentence-level metaphor density, mean novelty, novelty variability, and explicitness ratio. Sentence-level metaphor density, $\rho(T)$, captures the proportion of sentences in a discourse that contain at least one metaphor. Mean novelty, $v(T)$, reflects the average novelty of metaphorical expressions while novelty variability, $\sigma_v(T)$, captures the dispersion of those novelty scores within the discourse. Explicitness ratio, $\varepsilon(T)$, measures the proportion of metaphors that are overtly marked. Formally, for a discourse T , we define these features as follows:

$$\rho(T) = \frac{|S_{\text{met}}(T)|}{|S(T)|}, \quad v(T) = \frac{1}{|M_T|} \sum_{m \in M_T} s(m), \quad (1)$$

$$\sigma_v(T) = \sqrt{\frac{1}{|M_T|} \sum_{m \in M_T} (s(m) - v(T))^2}, \quad \varepsilon(T) = \frac{|M_{\text{exp}}(T)|}{|M_T|}. \quad (2)$$

Full definitions, notation, robustness checks (including a token-level density variant), and implementation details are provided in Appendix G. Because these dimensions may be correlated, we examine pairwise correlations before conducting subsequent regression analyses; collinearity diagnostics are reported in Appendix H.

3.4 Controlled Paraphrase Pipeline

To enable causal inference beyond observational co-variation, we construct a controlled paraphrase dataset in which a single metaphor per document is rewritten while all other content is held fixed. This intervention treats metaphor as a local lexical-semantic and stylistic manipulation, through which we vary metaphorical form, novelty, and explicitness while keeping the underlying meaning and discourse organization as stable as possible. The pipeline consists of three stages: novelty-controlled rewriting, explicitness rewriting, and multi-dimensional quality control.

Novelty-controlled Rewriting. To minimize semantic drift, we generate variants using only local transitions along the novelty scale. For each low-novelty metaphor token $m \in M_T^{\text{low}}$, an *assessor agent* first determines whether a distinct literal paraphrase can be recovered without changing the core propositional meaning. If so, a *generator agent* first produces a *Literal Baseline* variant and then generates *MEDIUM* and *HIGH* versions. Otherwise, the original low-novelty expression is treated as the baseline and only upward rewrites are generated. For *MEDIUM* items, we generate *Literal Baseline* and *HIGH*; for *HIGH* items, we generate *Literal Baseline* and *MEDIUM*. This yields controlled novelty variation while keeping each rewrite close to the original expression.

Explicitness Rewriting. We manipulate explicitness by progressively converting explicit metaphors into implicit form, removing overt comparison markers via spaCy dependency parsing. For a discourse T with $|M_T^{\text{exp}}|$ explicit metaphor instances, $T_k \in \mathcal{C}(T)$ denotes the version in which exactly k explicit metaphors are rewritten as implicit. This provides a graded intervention on explicitness while keeping discourse content as stable as possible.

Quality Control. All candidate variants pass through a three-stage critic pipeline. *Semantic equivalence* is verified using a DeBERTa-large-MNLI model, retaining only candidates with entailment probability $P(E) \geq 0.8$ and contradiction probability $P(C) < 0.2$ relative to the

¹metaphoricity and novelty are closely related and can co-vary. To avoid confounding effects between metaphoricity and novelty, we treat metaphoricity as an auxiliary control variable rather than a directly manipulated dimension.

literal baseline. The thresholds were selected through manual pilot inspection of generated rewrites to balance semantic preservation against data coverage. *Syntactic fluency* is enforced via LanguageTool, discarding any variant with grammatical errors. *Novelty calibration* is verified by computing contextual surprisal $S(w, c) = -\log P(w | c)$ using RoBERTa, where w is the generated metaphor token and c is the masked context; only variants whose surprisal falls within the target novelty band B_{target} are retained. The full pipeline is illustrated in Figure 1. To further assess pipeline reliability, we conducted a small-scale human validation study on metaphoricity detection, explicitness detection, and rewrite plausibility. Details of the sampling strategy, validation results, and evaluation protocol are provided in Appendix E.

Our validation targeted propositional-content preservation rather than full pragmatic equivalence: evaluators were instructed to accept a rewrite when the basic intended meaning was conveyed, without requiring that the rewritten expression be as idiomatic or contextually natural as the original. Under this lenient criterion, the reported acceptance rates should be read as an upper bound on rewrite quality; a stricter contextual-plausibility check (Appendix P) retains 77.4% of rewrites overall and 75.0% of high-novelty rewrites.

3.5 Models and Evaluation

Language Models. We evaluate nine models from three architecture families via 10-fold cross-validation. *Encoder-only* models (BERT-Large (Devlin et al., 2019), RoBERTa-Large (Liu et al., 2019), DeBERTa-v3-large (He et al., 2021)) are fine-tuned as sequence classifiers. *Encoder-Decoder* models (T5-Large (Raffel et al., 2020), Flan-T5-Large (Chung et al., 2024)) are trained with a sequence-to-sequence objective and scored by label likelihood under teacher forcing. *Decoder-only* models (LLaMA3-8B (Grattafiori et al., 2024), Mistral-7B (Jiang et al., 2023), Qwen3-1.7B (Yang et al., 2025)) are fine-tuned with 4-bit QLoRA (Dettmers et al., 2023) adapters. Qwen3-32B (Yang et al., 2025) is evaluated zero-shot using next-token logits over the digit tokens 1/2/3, serving as a training-free decoder baseline. We additionally reproduce two discourse-structured GCDC systems: a Discourse Relation-Enhanced model (DR-E) and an Entities and Discourse Relation-Enhanced model (E+DR-E) (Liu & Strube, 2025a;b), which inject discourse-relation structure and entity continuity into coherence prediction. Reproduction details are in Appendix O.

Model Outputs. For Experiment 1, we map *Low*, *Medium*, and *High* to 1, 2, and 3, respectively. We then compute three document-level outputs: absolute expected error, $AEE = \left| \sum_{c=1}^3 c p(c | T) - y_{\text{gold}} \right|$; predictive entropy; and the probability assigned to the gold label, $P(y_{\text{gold}} | T)$. For Experiment 2, we report $P(\text{High} | T)$, the margin between the two largest class probabilities, and predictive entropy. Mean adjacent-sentence cosine similarity is used as an auxiliary representation measure. In Experiment 2, we contrast conventional ($v(T) \leq 1$) and novel ($v(T) > 1$), because $s(m) \geq 1$, the conventional group contains only documents whose metaphors are all low-novelty.

Statistical Methods. In the observational analysis, we report domain-wise Spearman correlations and cross-domain OLS coefficients controlling for text length and syntactic complexity. In the controlled intervention analysis, we report Cohen’s d with bootstrap 95% confidence intervals (1,000 resamples) and Mann-Whitney U tests for group comparisons. Calibration is assessed via Expected Calibration Error (ECE) with 8 equal-width bins on $P(\text{High})$, and $\Delta\text{ECE} = \text{ECE}_{v(T)>1} - \text{ECE}_{v(T)\leq 1}$. Multiple comparisons in feature modulation analyses are corrected via Benjamini-Hochberg FDR globally across all model-domain-feature-target cells; significance markers in the grouped forest plot are Bonferroni-corrected.

4 Experiments, Results, and Discussion

We structure our empirical investigation in two complementary stages. **Experiment 1** conducts an observational analysis on naturally occurring discourse to establish whether document-level metaphor features co-vary with model behavior. While observational analysis can reveal associations, it cannot disentangle metaphor effects from potential confounds such as topic, register, or writing quality that may correlate with both metaphor usage and coherence. Therefore, **Experiment 2** employs a controlled intervention design

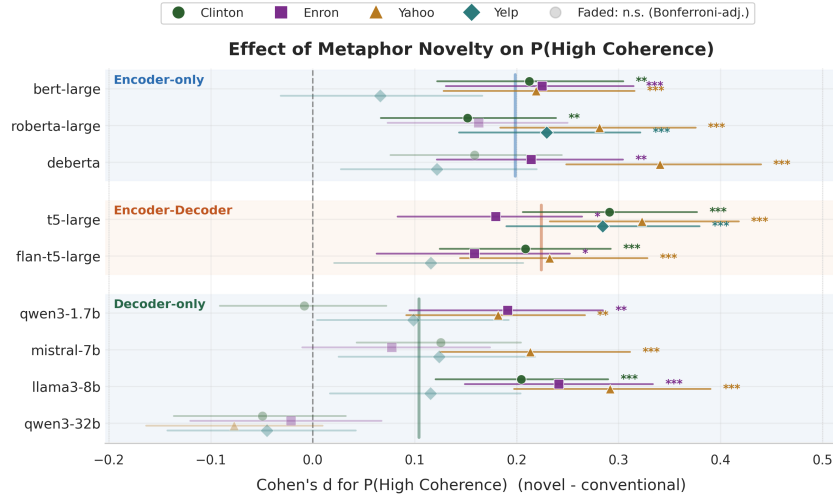


Figure 2: Novel metaphor consistently increases $P(\text{High})$ across most baseline LM conditions.

that manipulates a single metaphor per document while holding all other content fixed. Together, these two stages allow us to characterize both the ecological prevalence and the causal structure of metaphor’s influence on Language Model-based coherence evaluation.

4.1 Experiment 1: Observational Analysis of Metaphor Effects

We first analyze the original GCDC documents to determine whether naturally occurring metaphor features systematically co-vary with model behavior. This observational stage addresses two goals: (1) establishing baseline associations between metaphor features and model outputs within each domain, and (2) testing whether any such associations generalize across domains after controlling for text-level confounds.

Domain-wise associations. Experiment 1 shows that the observational relationship between metaphor and model behavior is strongly domain-dependent rather than uniform. Because metaphor co-varies with genre-specific signals in naturally occurring text, we interpret these results as associative rather than causal, which motivates the controlled intervention in Experiment 2. Detailed domain-wise Spearman correlations are reported in Appendix I. A cross-domain summary of the OLS analyses controlling for text length and syntactic complexity is provided in Appendix J; most feature effects attenuate after control, while document-level metaphoricity remains the clearest recurring association.

4.2 Experiment 2: Controlled Intervention on Coherence Judgments

In the controlled intervention setting, each accepted example differs from its source text by a single metaphor substitution that has passed multi-stage quality control for semantic equivalence and discourse preservation. This design attributes the observed differences to the controlled metaphorical-rewrite condition rather than to differences between source documents. The rewrite condition jointly alters metaphorical novelty and correlated local cues including vividness, lexical surprisal, and stylistic sophistication. As discussed in Section 3.4, our design does not isolate these properties from one another, and the results below should be read as the effect of the rewrite condition as a whole.

As an additional robustness check, we reran Experiment 2 after applying a stricter GPT-5.6 Sol-based contextual-plausibility filter to all rewrites; the principal effect directions, effect sizes, and domain-level patterns remained consistent with those obtained from the unfiltered data (Appendix P).

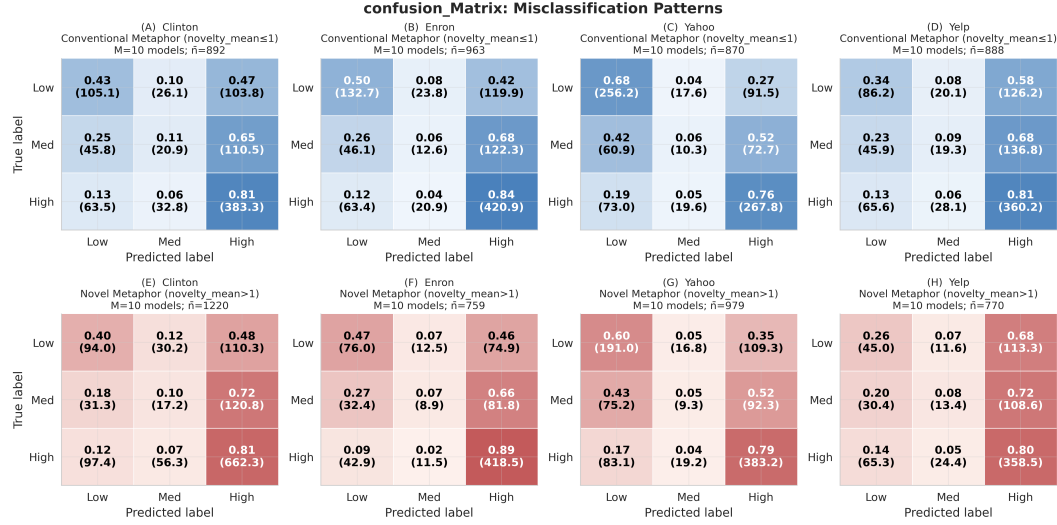


Figure 3: Under novel metaphor, models show stronger confusion toward higher coherence labels across domains. Matrices are pooled over ten models (the nine LM baselines and E+DR-E).

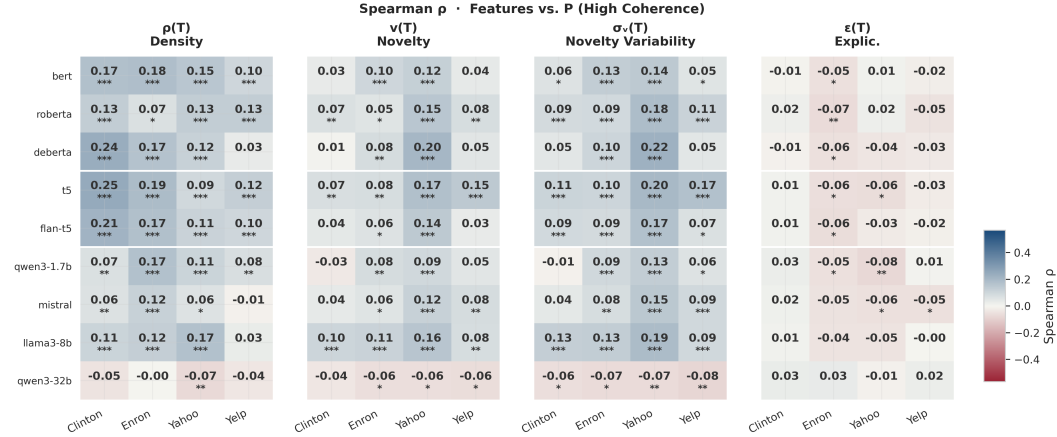


Figure 4: Metaphor-related features modulate P (High) in a domain- and model-dependent way.

4.2.1 RQ1 & RQ2: Coherence Judgments and Uncertainty

Our main finding is simple: metaphor tends to make coherence look better than it is. In the LM baselines, this appears most directly as an upward shift in P (High): novel metaphor makes models more likely to assign a *High* coherence judgment. The one exception is Qwen3-32B, the only zero-shot model, which shifts in the opposite direction (Figure 2). The aggregated confusion patterns tell the same story in a more intuitive way: under novel metaphor, low- and medium-coherence texts are more often absorbed into the *High* category (Figure 3). In other words, metaphorical novelty makes the discourse appear more coherent to the model than it really is. Appendix K provides detailed results for other metrics.

The reproduced SOTA systems remain sensitive, but their responses are strongly domain-dependent. Macro-F1 increases in *clinton* and decreases in *enron*, *yahoo*, and *yelp* for both systems (Figure 5). The pooled confusion matrices in Appendix O show that predictions remain strongly concentrated on the *High* label. Thus, explicit discourse-structural modeling reshapes, but does not eliminate, metaphor sensitivity.

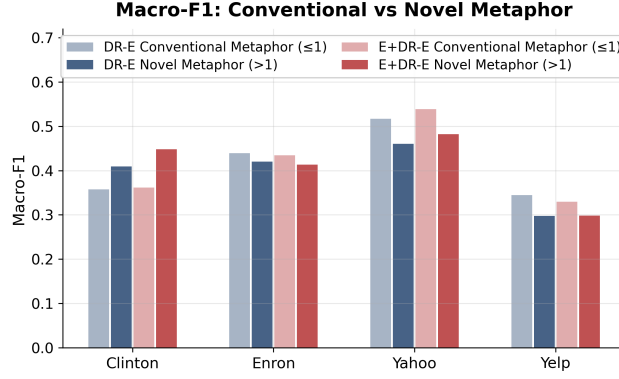


Figure 5: Novel metaphor affects class-wise performance in the reproduced SOTA coherence models.

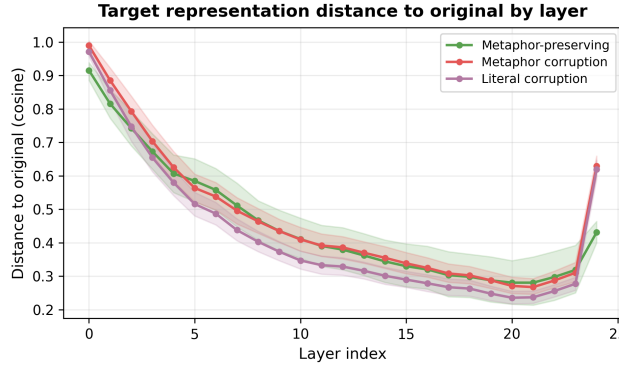


Figure 6: Layer-wise cosine distance between the original target representation and three paired rewrite conditions: *preserve_metaphor*, *corrupt_metaphor*, and *corrupt_literal*.

4.2.2 RQ3: Probability Calibration

If RQ1 and RQ2 show that metaphor can bias the predicted coherence label, RQ3 asks whether model confidence remains trustworthy under that bias. The answer is less uniform than the judgment shift itself. Unlike the consistent upward shift in $P(\text{High})$, calibration responses are less uniform, though the overall tendency is toward degradation: novel metaphor increases ECE in 25 of the 36 model–domain conditions. The effect concentrates in *yahoo* and *yelp*, where seven of the nine models become less well calibrated, followed by *enron* with six; *clinton*, by contrast, splits almost evenly between degradation and improvement.

Architecture accounts for only part of this pattern. Encoder-only and encoder–decoder models degrade in most conditions, while the decoder-only models divide along model lines: both Qwen variants worsen across all four domains, whereas LLaMA3-8B and Mistral-7B improve in several, notably *enron* and *yahoo*. This suggests that calibration sensitivity arises from an interaction between a model’s own confidence behavior and the discourse conventions of each domain, rather than from architecture or genre alone. Detailed ECE values and domain-level discussion appear in Appendix L.

Notably, explicit discourse-structural modeling does not remove this sensitivity: the reproduced SOTA systems show calibration shifts under metaphorical novelty as well. Metaphor, then, affects *what* models predict more consistently than *how reliably* they express those predictions.

4.2.3 RQ4: Feature Modulation and Representation-Space Analysis

The findings above raise a more fundamental question: are models reacting mainly to the novel metaphorical *form*, or have they learned to preserve deeper meaning? To address this, we first examine feature-level modulation in the baseline models. Feature-level modulation remains model- and domain-dependent (Figure 4; see Appendix M for margin and entropy). Novelty and novelty variability show the broadest associations with $P(High)$, whereas their effects on margin and entropy are less uniform; density and explicitness exhibit more localized patterns.

To distinguish sensitivity to metaphorical form from sensitivity to semantic change, we perform an additional layer-wise representation analysis on a four-condition negative-control set. Each item contains the same source document and target span under four matched conditions: *original*, *meaning-preserving metaphor*, *meaning-changing metaphor*, and *meaning-changing literal*. We retain 49 complete paired items spanning all four GCDC domains. For Flan-T5-Large, we average the encoder hidden states over the target span at each layer and compute the cosine distance from each rewrite to the original representation (see Appendix N for implementation details).

As shown in Figure 6, the three rewrite trajectories remain close through most intermediate layers. At the final encoder layer, however, both meaning-changing conditions diverge sharply from the original, while the meaning-preserving metaphor remains closer. Semantic compatibility is therefore most clearly reflected in the uppermost representation rather than uniformly across depth. This pattern suggests that metaphor-induced shifts in coherence judgments can coexist with internal representations that retain the intended semantic distinction.

5 Conclusion, Practical Implications and Future Directions

This paper shows that metaphor is a useful probe of what coherence models respond to. Across models and domains, novel metaphor often shifts judgments toward *High* coherence, indicating that automated coherence evaluation is influenced not only by discourse organization, but also by surface linguistic realization. The practical relevance of our findings lies in their systematic direction: meaning-preserving figurative variation can shift coherence judgments even when discourse organization remains stable. Such sensitivity may affect borderline decisions, rankings, or aggregated scores in applications such as automated writing feedback, essay scoring, and generated-text filtering. A natural mitigation direction is to use controlled rewrite pairs as counterfactual supervision, assigning the same coherence label to original and meaning-preserving metaphorical texts. The four-condition design could also support contrastive learning, with meaning-preserving rewrites as positive pairs and meaning-changing rewrites as hard negatives. Future work could investigate the mechanisms underlying these shifts through activation- or circuit-level analysis, and extend the framework beyond metaphor to other meaning-preserving perturbations, including formality, lexical sophistication, syntactic elaboration, and discourse-marker variation. Richer discourse analysis could further test whether rewrite effects depend on relations such as elaboration, contrast, cause, or evaluation, or influence explicit tasks such as RST relation labeling and nuclearity prediction.

Limitations

Our rewrites preserve the basic intended meaning but are not always natural: evaluators applied a lenient criterion, and some retained rewrites read as awkward or unidiomatic in context. Because our novelty levels are calibrated against surprisal, awkward phrasing and novel metaphor are hard to separate; the effect may reflect sensitivity to unusual word choice in general. The effects also persist under the stricter filter in Appendix P, though that filter does not address this confound. Our semantic-equivalence thresholds were set by pilot inspection rather than by sensitivity analysis, and our human validation was small in scale. Finally, the representation-level analysis is based on a single model and 49 paired items, and we do not test whether it generalizes to other architectures.

References

- Ehsan Aghazadeh, Mohsen Fayyaz, and Yadollah Yaghoobzadeh. Metaphors in pre-trained language models: Probing and generalization across datasets and languages. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2037–2050, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.144. URL <https://aclanthology.org/2022.acl-long.144/>.
- Nicholas Asher and Alex Lascarides. *Metaphor in Discourse*, pp. 262–289. Studies in Natural Language Processing. Cambridge University Press, 2001.
- Regina Barzilay and Mirella Lapata. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34, 2008.
- Brian F Bowdle and Dedre Gentner. The career of metaphor. *Psychological review*, 112(1):193, 2005.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of machine learning research*, 25(70):1–53, 2024.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.
- Erik-Lân Do Dinh, Hannah Wieland, and Iryna Gurevych. Weeding out conventionalized metaphors: A corpus of novel metaphor annotations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1424, 2018.
- Peter W Foltz, Walter Kintsch, and Thomas K Landauer. The measurement of textual coherence with latent semantic analysis. *Discourse processes*, 25(2-3):285–307, 1998.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- Camille Guinaudeau and Michael Strube. Graph-based local coherence modeling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 93–103, 2013.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*, 2021.
- Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spacy: Industrial-strength natural language processing in python. 2020. doi: 10.5281/zenodo.1212303.
- Dan Iter, Jong Yoon, and Dan Jurafsky. Automatic detection of incoherent speech for diagnosing schizophrenia. In *Proceedings of the fifth workshop on computational linguistics and clinical psychology: From keyboard to clinic*, pp. 136–146, 2018.
- Hyeju Jang, Seungwhan Moon, Yohan Jo, and Carolyn Rose. Metaphor detection in discourse. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 384–392, 2015.

- Hyeju Jang, Keith Maki, Eduard Hovy, and Carolyn Rose. Finding structure in figurative language: Metaphor detection with topic-based frames. In *Proceedings of the 18th Annual SIGDIAL meeting on Discourse and Dialogue*, pp. 320–330, 2017.
- Katarzyna Jankowiak. Normative data for novel nominal metaphors, novel similes, literal, and anomalous utterances in polish and english. *Journal of Psycholinguistic Research*, 49(4): 541–569, 2020.
- 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. Mistral 7b, 2023. URL <https://arxiv.org/abs/2310.06825>.
- Shafiq Joty, Muhammad Tasnim Mohiuddin, and Dat Tien Nguyen. Coherence modeling of asynchronous conversations: A neural entity grid approach. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 558–568, 2018.
- Sandra Just, Erik Haegert, Nora Ko  ranov  , Anna-Lena Br  cker, Ivan Nenchev, Jakob Funcke, Christiane Montag, and Manfred Stede. Coherence models in schizophrenia. In *Proceedings of the sixth workshop on computational linguistics and clinical psychology*, pp. 126–136, 2019.
- Philippe Laban, Luke Dai, Lucas Bandarkar, and Marti A Hearst. Can transformer models measure coherence in text: Re-thinking the shuffle test. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 1058–1064, 2021.
- Alice Lai and Joel Tetreault. Discourse coherence in the wild: A dataset, evaluation and methods. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 214–223, 2018.
- Yucheng Li, Frank Guerin, and Chenghua Lin. Finding challenging metaphors that confuse pretrained language models. *arXiv preprint arXiv:2401.16012*, 2024.
- Wei Liu and Michael Strube. Discourse relation-enhanced neural coherence modeling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4748–4762, 2025a.
- Wei Liu and Michael Strube. Joint modeling of entities and discourse relations for coherence assessment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21910–21926, 2025b.
- Wei Liu, Xiyang Fu, and Michael Strube. Modeling structural similarities between documents for coherence assessment with graph convolutional networks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7792–7808, 2023.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- Omar Momen, Emilie Sitter, Berenike Herrmann, and Sina Zarri  . Surprisal and metaphor novelty judgments: Moderate correlations and divergent scaling effects revealed by corpus-based and synthetic datasets. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8113–8127, 2026.
- Dat Tien Nguyen and Shafiq Joty. A neural local coherence model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1320–1330, 2017.

- Prisca Piccirilli and Sabine Schulte Im Walde. What drives the use of metaphorical language? negative insights from abstractness, affect, discourse coherence and contextualized word representations. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pp. 299–310, 2022a.
- Prisca Piccirilli and Sabine Schulte Im Walde. Features of perceived metaphoricity on the discourse level: Abstractness and emotionality. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 5261–5273, 2022b.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.
- Daniel Sanford. Discourse and metaphor: A corpus-driven inquiry. *Corpus Linguistics & Linguistic Theory*, 4(2), 2008.
- Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.
- Aili Shen, Meladel Mistica, Bahar Salehi, Hang Li, Timothy Baldwin, and Jianzhong Qi. Evaluating document coherence modeling. *Transactions of the Association for Computational Linguistics*, 9:621–640, 2021.
- Swapna Somasundaran, Jill Burstein, and Martin Chodorow. Lexical chaining for measuring discourse coherence quality in test-taker essays. In *Proceedings of COLING 2014, the 25th International conference on computational linguistics: Technical papers*, pp. 950–961, 2014.
- Gerard J Steen, Aletta G Dorst, J Berenike Herrmann, Anna Kaal, Tina Krennmayr, and Tryntje Pasma. Mipvu: A manual for identifying metaphor-related words. *Metaphor identification in multiple languages: MIPVU around the world*, pp. 23–40, 2019.
- Xiaoyu Tong, Rochelle Choenni, Martha Lewis, and Ekaterina Shutova. Metaphor understanding challenge dataset for llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3517–3536, 2024.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL <https://arxiv.org/abs/2505.09388>.

A Detailed Experimental Setup

Hardware and data. All experiments were conducted on two NVIDIA A100 40GB GPUs. We used the Grammarly Corpus of Discourse Coherence (GCDC) with its four domains (*clinton*, *enron*, *yahoo*, and *yelp*) under a 10-fold cross-validation setup. The task is three-way discourse coherence classification. Where applicable, the default random seed was set to 42.

Encoder models. For encoder-based classifiers (BERT-large, RoBERTa-large, and DeBERTa-v3-large), we used full fine-tuning with a maximum input length of 512, a maximum of 15 training epochs, early stopping with patience 3, and weight decay 0.01. The learning rate was 2×10^{-5} . The batch size was 32 for BERT-large and RoBERTa-large, and 8 for DeBERTa-v3-large.

Encoder-Decoder models. For T5-large and Flan-T5-large, we fine-tuned sequence-to-sequence models with a maximum source length of 512 and a maximum target length of 8. Training ran for up to 15 epochs with early stopping (patience 3), weight decay 0.01, batch size 8, and learning rate 3×10^{-5} .

Decoder-only models. For decoder-only open-weight models, we used 4-bit QLoRA fine-tuning whenever training was performed. Unless otherwise noted, these runs used NF4 quantization, LoRA rank $r = 16$, $\alpha = 32$, dropout 0.1, gradient accumulation of 4, weight decay 0.01, warmup ratio 0.1, and maximum sequence length 512. LLaMA-family models were trained for up to 5 epochs with learning rate 2×10^{-4} ; Mistral-family models were trained for up to 5 epochs with learning rate 1×10^{-4} ; Qwen fine-tuning runs used up to 3 epochs with learning rate 1×10^{-4} . Batch sizes varied by model size, with smaller models generally using larger effective batch sizes.

Zero-shot evaluation. Qwen3-32B was evaluated in zero-shot mode rather than fine-tuned. For these models, class probabilities were obtained from next-token logits restricted to the coherence label tokens and normalized with a softmax.

Model selection. For encoder models, the best checkpoint was selected based on development accuracy. For sequence-to-sequence and QLoRA-based decoder models, the best checkpoint was selected based on development loss.

B Dataset Details

We use the Grammarly Corpus of Discourse Coherence (GCDC) (Lai & Tetreault, 2018), a corpus of real-world texts written by non-professional authors across four domains: political emails released via FOIA (*clinton*), corporate emails from the Enron corpus (*enron*), online forum responses from Yahoo Answers (*yahoo*), and business reviews from Yelp (*yelp*).

Each domain contains approximately 1,200 texts filtered to 100–300 words, with URLs and list-heavy documents excluded, each document received coherence ratings from both crowd workers and expert raters; in this work, we use the three-way consensus label derived from the expert ratings. Every document is assigned a human coherence label in $\{1, 2, 3\}$ based on understandability, organizational quality, and thematic consistency, while disregarding surface errors such as spelling or grammar. We follow the standard 10-fold cross-validation splits provided with the dataset.

C Annotation Details

C.1 Metaphor Identification Prompt

Metaphor identification is performed at the token level using a prompting-based approach grounded in the MIPVU framework (Steen et al., 2019). The prompt instructs the model to evaluate each lexical unit by comparing its contextual meaning with its most basic or conventional meaning. A token is labeled as metaphorical if it is used in a sense that contrasts with its primary literal meaning and can be understood through cross-domain comparison.

Prior to large-scale application, we calibrate the prompt on 200 randomly sampled sentences from the VU Amsterdam Metaphor Corpus (VUAMC), which was originally annotated under the MIPVU protocol. Calibration involves iterative refinement of instruction wording to reduce systematic false positives (e.g., function words misidentified as metaphorical) and false negatives (e.g., highly conventionalized content-word metaphors missed due to lack of explicit cuing).

The final prompt used for metaphor identification is provided below:

You are a linguist tasked with identifying metaphorical language using explicit step-by-step reasoning according to the procedure below. For each lexical unit (LU) in a given text, determine whether it is METAPHOR or LITERAL, and justify your reasoning process by identifying the SOURCE_DOMAIN and TARGET_DOMAIN in terms of the Conceptual Metaphor Theory.

For each LU, explicitly reason through the following steps and assign a label:

Step 1) Identify the contextual meaning of the LU and the domain of that meaning (TARGET_DOMAIN).

Step 2) Determine whether a more basic meaning (people use frequently in daily life) exists; if yes, identify it and its domain (SOURCE_DOMAIN).

Step 3) Decide whether the basic and contextual meanings are sufficiently distinct (use dictionary senses as guidance).

Step 4) Decide whether the contextual meaning can be understood via similarity or mapping from the basic meaning.

If Steps 2{4 are all satisfied, label the LU as METAPHOR.

Otherwise, label it as LITERAL.

If the context is insufficient to determine the intended contextual meaning but a metaphorical interpretation is plausible, label the LU as METAPHOR and use UNKNOWN for domains where necessary.

Examples of conceptual domains: HUMAN_BODY, HEALTH_ILLNESS, ANIMALS, MACHINES_TOOLS, BUILDINGS_CONSTRUCTION, PLANTS, GAMES_SPORT, COOKING_FOOD, ECONOMIC_TRANSACTIONS, FORCES, LIGHT_DARKNESS, HEAT_COLD, MOVEMENT_DIRECTION, WAR_CONFLICT, CONTAINER, ORIENTATION_VERTICALITY, EMOTION, DESIRE, MORALITY, THOUGHT, SOCIETY, RELIGION, POLITICS, ECONOMY, HUMAN_RELATIONSHIPS, COMMUNICATION, EVENTS_ACTIONS, TIME, LIFE_DEATH, etc.

Produce the original text with inline markup on metaphorical LUs only (do NOT output a table).

For each LU labeled METAPHOR:

Wrap the LU surface form in ... <\b>

Immediately append <MET>SOURCE_DOMAIN<TAB>TARGET_DOMAIN<\MET>

Rules:

Preserve the original text as-is (same words, order, punctuation), modifying ONLY by inserting the markup around/after metaphorical LUs.

Output markings ONLY for LUs labeled METAPHOR. If LABEL is LITERAL, leave the LU unmarked (do NOT add anything).

Use exactly the tags shown: ... <\b><MET> ... <\MET>.

SOURCE_DOMAIN and TARGET_DOMAIN must be written as domain labels; if unknown, use UNKNOWN.

Inside <MET> ... <\MET>, separate SOURCE_DOMAIN and TARGET_DOMAIN using the literal string <TAB> (do not add other fields).

Do NOT output explanations outside the marked text.

Do NOT output any additional lines such as "TEXT:" or separate lists; output only the marked text.

Worked Example (For Guidance Only)

The following example illustrates the annotation logic. Do not output these steps in your final response.

Text: Latest corporate unbundler reveals laid-back approach : Roland Franklin , who is leading a 697m pound break-up bid for DRG , talks to Frank Kane.

Output: "Latest corporate unbundler

reveals<MET>LIGHT_DARKNESS<TAB>COMMUNICATION<\MET> laid-back

approach<MET>MOVEMENT_DIRECTION<TAB>THOUGHT<\MET> : Roland

Franklin , who is

leading<MET>MOVEMENT_DIRECTION<TAB>EVENTS_ACTIONS<\MET> a 697m pound break-up bid for DRG , talks to Frank Kane",

For example, for LU "reveals":

Step 1 | Contextual meaning:

In context, "reveals" means "to disclose/make known information (to readers) about someones style or strategy."

TARGET_DOMAIN = COMMUNICATION (information disclosure in reporting)

Step 2 | More basic meaning (dictionary-attested):

A more basic, concrete sense is "to uncover something so it becomes visible; to remove what hides it and let it be seen."

SOURCE_DOMAIN = LIGHT_DARKNESS (visibility: hidden → seen)

Step 3 | Distinctness:

Making something physically visible (visibility/hiddenness) is distinct from communicating information to an audience.

Step 4 | Similarity link:

The disclosure of information is understood as bringing something from hiddenness into the open, i.e., from "not seen/unknown" to "seen/known."

Decision:

LABEL = METAPHOR

IMPORTANT:

- The worked example above is only a reference. In real annotation, do NOT output the steps.
- Your final response must ONLY output the annotation text in the required format.

Now it is your turn.

Text:

C.2 Novelty Labeling Prompt

Novelty labeling is formulated as a four-class classification task applied to each token previously identified as metaphorical. The theoretical basis follows [Bowdle & Gentner \(2005\)](#), who distinguish conventional metaphors (entrenched mappings processed as fixed expressions) from novel metaphors (creative mappings requiring active conceptual comparison). The prompt operationalizes this distinction through two cues: (1) the degree to which the cross-domain mapping is surprising or unfamiliar in everyday language use, and (2) whether the expression would be recognized as metaphorical by an average adult speaker without contextual scaffolding. Continuous novelty scores $s(m) \in [-1, 1]$ are elicited from the model and subsequently discretized as follows:

$$\ell(m) = \begin{cases} \text{none} & \text{if } s(m) \leq -0.9, \\ \text{low} & \text{if } -0.9 < s(m) < 0.1, \\ \text{medium} & \text{if } 0.1 \leq s(m) < 0.5, \\ \text{high} & \text{if } s(m) \geq 0.5. \end{cases} \quad (3)$$

The none category captures dead metaphors and fully lexicalized expressions (e.g., *table leg*) that exhibit no perceptible metaphorical activation in context. The low category captures conventional metaphors (e.g., *grasp an idea*) that are recognizably metaphorical but highly entrenched. The medium and high categories capture increasingly creative and unfamiliar mappings that require active metaphorical interpretation.

The final prompt used for metaphor novelty labeling is provided below:

You are annotating METAPHOR NOVELTY on a 3-point scale for each given metaphor (vehicle) in context.

Scale:

(1) level 1 (LOW): Highly conventional metaphor. The metaphorical use is well established in dictionaries or cultural norms; may be a dead metaphor or idiom. Functions as a category with minimal imaginative effect. Examples: "time is money", "waste your life", "shoulder responsibility".

(2) level 2 (MEDIUM): Vivid and metaphorical, but widely understood with little cognitive effort. A common trope rather than a fixed idiom. Includes emotionally intensified variants of conventional metaphors where vividness is increased but the underlying mapping is familiar.

Examples: "fight for happiness", "rollercoaster of emotions".

(3) level 3 (HIGH): Creative or novel metaphor that causes semantic surprise or defamiliarization and requires active interpretation or re-mapping between source and target domains. Often poetic or highly individual. Examples: "grief is a house with no walls", "the lie spread like ink on paper".

Decision helpers (use if uncertain):

(1) Grammatical preference (supportive only, not decisive): adding "like" weaker -> level 1; adding "like" clearer -> level 3.

(2) Category test: vehicle used as a common label for a class of people/things -> level 1/2; otherwise -> level 3.

Instructions:

Given:

(1) the full text with metaphor spans already marked in the following format: `<b>SPAN<\b>`

(2) you must assign a novelty grade to each marked metaphor span.

Output format (Inline-marked text only):

Produce ONLY the same input text, but for each marked metaphor span, append the novelty grade INSIDE the `<b> ... <\b>` tag, immediately after the span text, using this exact format: ", LOW" or ", MEDIUM" or ", HIGH".

Example:

Input: He sounds like a `<b>bird<\b>`.

Output: He sounds like a `<b>bird, MEDIUM<\b>`.

Rules:

Preserve the original text as-is (same words, order, punctuation), modifying ONLY by inserting the appended novelty grade for each already-marked metaphor span.

Do NOT add, remove, or change any existing tags (`<b>`, `<\b>`).

Mark ONLY the metaphor spans already marked in the input; do NOT mark any additional text.

The novelty grade must be exactly one of: LOW, MEDIUM, HIGH (uppercase).

Do NOT output explanations, JSON, tables, indices, or any extra lines; output only the modified text.

Text:

C.3 Metaphoricity Labeling Prompt

Metaphoricity captures the degree to which a token is perceived as metaphorical rather than literal in context, independently of how novel that metaphor use is. A highly conventional metaphor such as *grasp an idea* may exhibit low novelty but retain a measurable degree of metaphoricity, while a novel metaphor may exhibit high values along both dimensions. We model metaphoricity as a three-level categorical variable (LOW, MEDIUM, HIGH) and use it as an auxiliary control variable in regression analyses rather than as a directly manipulated dimension.

The final prompt used for metaphoricity labeling is provided below:

You are a linguist tasked with identifying metaphorical language that shows high metaphoricity. Metaphoricity reflects the degree to which a metaphorically marked words usage in context is perceived as figurative rather than literal true. Research shows that metaphoricity is not binary but varies in degree depending on factors such as semantic similarity between source and target domains, degree of conventionality, and how noticeable the figurative meaning is in context. Based on these dimensions, metaphoricity can be categorized into three levels:

LOW: weakly metaphorical with minimal semantic departure from literal meaning;

MEDIUM: clearly metaphorical but relatively familiar or moderate in metaphorical effect;

HIGH: strongly metaphorical with a large semantic departure requiring active figurative interpretation.

Instructions:

Given:

- (1) the full text with metaphor spans already marked in the following format: `<b>SPAN<\b>`
- (2) you must assign a metaphoricity grade to each marked metaphor span.

Output format (Inline-marked text only):

Produce ONLY the same input text, but for each marked metaphor span, append the metaphoricity grade INSIDE the `<b> ... <\b>` tag, immediately after the span text, using this exact format: `" , LOW"` or `" , MEDIUM"` or `" , HIGH"`.

Example:

Input: He sounds like a `<b>bird<\b>`.

Output: He sounds like a `<b>bird, MEDIUM<\b>`.

Rules:

Preserve the original text as-is (same words, order, punctuation), modifying ONLY by inserting the appended metaphoricity grade for each already-marked metaphor span.

Do NOT add, remove, or change any existing tags (`<b>`, `<\b>`).

Mark ONLY the metaphor spans already marked in the input; do NOT mark any additional text.

The metaphoricity grade must be exactly one of: LOW, MEDIUM, HIGH (uppercase).

Do NOT output explanations, JSON, tables, indices, or any extra lines; output only the modified text.

Text:

D Novelty Annotation Evaluation

To validate the novelty-labeling procedure, we sample 500 text excerpts from the VU Amsterdam Metaphor Corpus (VUAMC), comprising 9,857 token-level instances with reference novelty annotations (Do Dinh et al., 2018). We apply the same discretization scheme defined in Equation 3 to both the model-generated scores and the reference annotations, enabling direct comparison at the four-class level. Table 2 reports low/medium/high-class precision, recall, and F1, as well as macro-averaged F1 across classes.

Level	Precision	Recall	F1	Support
Low	0.89	0.99	0.94	5924
Medium	0.98	0.81	0.89	3666
High	0.98	0.77	0.86	267
Macro avg.	0.95	0.86	0.90	9857

Table 2: Novelty labeling evaluation on VUAMC. Scores are computed after applying the discretization scheme in Equation 3 to model-generated continuous scores.

The macro-F1 of approximately 90% indicates that the prompting-based approach achieves reliable alignment with human-annotated novelty levels under the same discretization scheme. The primary source of disagreement lies at the boundary between the low and medium categories, where cross-annotator agreement is also lower in the original VUAMC annotations (Do Dinh et al., 2018).

E Human Validation of Annotation and Rewrite Outputs

In addition to the automatic quality-control pipeline, we conducted a small-scale human validation study to assess three components of the proposed pipeline: metaphoricity detection, explicitness detection, and rewrite plausibility. For each validation task, we sampled 50 cases from the four GCDC domains, balanced across relevant feature labels where applicable. Two evaluators were assigned to each validation task, with three evaluators involved in total. The same pair of evaluators assessed metaphoricity and explicitness detection, while a separate pair assessed rewrite plausibility.

For metaphoricity and explicitness detection, evaluators were asked whether they agreed with the pipeline output for each case. For rewrite plausibility, evaluators judged whether the rewritten metaphor was plausible in the surrounding discourse context. Table 3 summarizes the agreement results. Raw inter-evaluator agreement was high across all three validation tasks, ranging from 90% to 92%. After correcting for chance agreement, Cohen’s κ was 0.648 for metaphoricity detection, 0.556 for explicitness detection, and 0.752 for rewrite plausibility. These results indicate moderate to substantial agreement beyond chance, with the strongest agreement observed for rewrite plausibility. The lower κ for explicitness despite its 92% raw agreement reflects the strong concentration of both evaluators’ responses in the agreement category, which increases the level of agreement expected by chance.

F Intervention Dataset Statistics

Quality control pass rates. Table 4 reports the proportion of generated candidates that passed each quality control stage.

Validation task	E1-P	E2-P	E1-E2	Cohen’s κ
Metaphoricity detection	86%	80%	90%	0.648
Explicitness detection	90%	90%	92%	0.556
Rewrite plausibility	76%	84%	92%	0.752

Table 3: Small-scale human validation of the annotation and rewrite pipeline. Each task used 50 sampled cases from the four GDC domains. E1 and E2 denote Evaluators 1 and 2, respectively, and P denotes the automated pipeline. Thus, E1-P and E2-P indicate each evaluator’s agreement with the pipeline, while E1-E2 indicates inter-evaluator agreement. For rewrite plausibility, evaluator agreement with the pipeline is the proportion of retained rewrites judged plausible in context.

Domain	Initial	Semantic Equivalence		Syntactic Fluency		Novelty Calibration	
		Remaining	Pass %	Remaining	Pass %	Remaining	Pass %
<i>clinton</i>	31636	22952	72.6	1925	8.4	1040	54.0
<i>enron</i>	30875	24067	77.9	1200	5.0	648	54.0
<i>yahoo</i>	23276	21620	92.9	1419	6.6	771	54.3
<i>yelp</i>	20608	19089	92.6	1076	5.6	574	53.3
Overall	106395	87728	82.5	5620	6.4	3033	54.0

Table 4: Quality-control filtering statistics for intervention generation by domain.

G Discourse-level Metaphor Features

This appendix provides the full definitions and implementation details for the discourse-level metaphor features used in our analyses.

Let a discourse be denoted as $T = \{s_1, \dots, s_n\}$, where n is the number of sentences in T . Let $S(T)$ denote the set of sentences in T , and let M_T denote the set of metaphorical instances identified in T from token-level annotation. For each metaphorical instance $m \in M_T$, let $s(m) \in \{1, 2, 3\}$ denote its novelty score. Let $M_{\text{exp}}(T) \subseteq M_T$ denote the subset of explicitly marked metaphors, and let $S_{\text{met}}(T) \subseteq S(T)$ denote the set of sentences containing at least one metaphorical expression.

Metaphor Density. Our primary density measure is sentence-level metaphor density:

$$\rho(T) = \frac{|S_{\text{met}}(T)|}{|S(T)|}, \quad (4)$$

where $S_{\text{met}}(T) \subseteq S(T)$ denotes the set of sentences containing at least one metaphorical expression. This definition captures how broadly metaphor is distributed across the discourse, rather than how many metaphorical tokens appear in total. Intuitively, it distinguishes texts in which metaphorical language is concentrated in only one sentence from those in which it is spread more widely across the discourse.

For robustness, we additionally consider a token-/instance-level density variant:

$$\rho_{\text{token}}(T) = \frac{|M_T|}{|\text{tokens}(T)|}, \quad (5)$$

where $|\text{tokens}(T)|$ is the number of tokens in T . Unless otherwise noted, the main analyses use the sentence-level measure $\rho(T)$, since it more directly reflects discourse-level metaphorical dispersion.

Metaphor Novelty. Let $s(m) \in \{1, 2, 3\}$ denote the novelty score assigned to a metaphorical instance $m \in M_T$. We define discourse-level novelty as the average instance-level

novelty score:

$$\nu(T) = \frac{1}{|M_T|} \sum_{m \in M_T} s(m). \quad (6)$$

This quantity summarizes the overall degree of metaphorical novelty in a discourse.

To account for potential heterogeneity within a discourse, we additionally consider the maximum novelty value:

$$\nu_{\max}(T) = \max_{m \in M_T} s(m). \quad (7)$$

This auxiliary statistic captures whether a discourse contains at least one highly novel metaphor, even when its average novelty is only moderate.

Finally, to reflect intra-discourse variability in metaphorical deviation, we compute the standard deviation of novelty scores:

$$\sigma_\nu(T) = \sqrt{\frac{1}{|M_T|} \sum_{m \in M_T} (s(m) - \nu(T))^2}. \quad (8)$$

This measure distinguishes texts whose metaphorical instances are relatively homogeneous in novelty from those mixing more conventional and more novel metaphorical expressions.

Metaphor Explicitness. Let $M_{\text{exp}}(T) \subseteq M_T$ denote explicitly marked metaphors. We define discourse-level explicitness as

$$\varepsilon(T) = \frac{|M_{\text{exp}}(T)|}{|M_T|}. \quad (9)$$

This proportion-based measure distinguishes discourses dominated by implicit metaphorical usage from those containing more overt metaphorical cues.

Main and Auxiliary Features. In the main text, we report sentence-level density $\rho(T)$, mean novelty $\nu(T)$, novelty variability $\sigma_\nu(T)$, and explicitness ratio $\varepsilon(T)$ as the primary discourse-level features. The auxiliary quantities $\rho_{\text{token}}(T)$ and $\nu_{\max}(T)$ are used for robustness checks and interpretive support.

Implementation Details and Edge Cases. All discourse-level features are derived from token-level metaphor annotations and then aggregated to the document level. Density is computed over sentences, while novelty and explicitness are computed over annotated metaphorical instances.

Documents with no metaphorical instances ($|M_T| = 0$) are excluded from analyses involving discourse-level metaphor features. We make this design choice to ensure that all document-level variables are grounded in observed metaphor annotations and that subsequent metaphor-focused rewrites remain evidence-based and controlled. In contrast, introducing metaphor into originally non-metaphorical documents would require unconstrained selection of candidate expressions for metaphorical rewriting, which could introduce additional variability in rewrite location, semantic preservation, and discourse function. By restricting analysis to documents with at least one annotated metaphorical instance, we maintain tighter control over the intervention space and improve the interpretability of downstream comparisons.

Overall, this feature design allows us to distinguish how widely metaphor is distributed across a discourse, how novel its metaphorical expressions are on average, how much novelty varies within the discourse, and how explicitly metaphor is marked.

H Collinearity Diagnostics

H.1 Pairwise Correlations among Metaphor Features

Figure 7 visualizes pairwise correlations among the four discourse-level metaphor features: density $\rho(T)$, mean novelty $\nu(T)$, novelty variability $\sigma_\nu(T)$, and explicitness $\varepsilon(T)$. Correla-

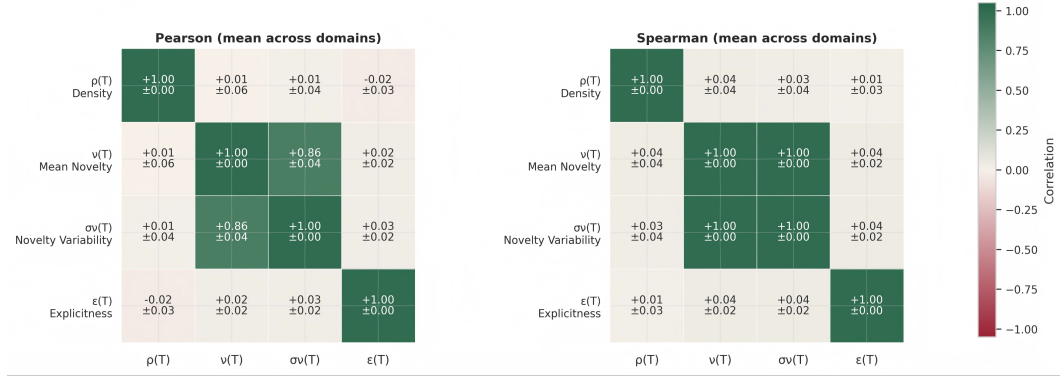


Figure 7: Pairwise Pearson (left) and Spearman (right) correlations among discourse-level metaphor features, computed separately for each GCDC domain and averaged across domains. Each cell reports the mean correlation and its standard deviation across domains. Mean novelty and novelty variability are strongly correlated, whereas density and explicitness remain nearly orthogonal to the other feature dimensions.



Domain	Dominant feature	AEE	Entropy	$P(\text{true})$
<i>clinton</i>	Metaphoricity $\mu(T)$	↑	↑	↓
<i>enron</i>	Metaphoricity $\mu(T)$	↑	↑	↓
<i>yelp</i>	Density $\rho(T)$	↓	↓	↑
<i>yahoo</i>	None stable	—	—	—

Table 5: Summary of domain-wise observational associations in Experiment 1. Arrows indicate the dominant localized direction of association across models; — means no stable pattern. Novelty and explicitness do not show stable cross-domain effects. The full model-by-model Spearman heatmap is provided in Figure 8.

with higher error and uncertainty, consistent with political emails in which metaphor accompanies rhetorically dense or semantically layered discourse. *Enron* shows a weaker version of the same pattern, plausibly because its more formulaic corporate-email register leaves less room for metaphorical variation to shape coherence. By contrast, *yelp* reverses the expected disruption story: higher density $\rho(T)$ is linked to lower error and higher confidence, suggesting that in review writing, denser metaphor use may function more as a cue for expressive fluency or writing quality than as a source of semantic instability. *Yahoo* shows the weakest pattern, consistent with a noisier and more heterogeneous Q&A genre. These results suggest that metaphor is not a clean observational variable in natural discourse, but is bundled with genre-specific signals, which motivates the controlled intervention in Experiment 2.

Figure 8 provides the full model-by-model heatmap of domain-wise Spearman correlations between metaphor features and observational outputs (AEE, predictive entropy, and $P(\text{true})$), complementing the compact summary in Table 5.

J Regression-based Cross-domain Effects in Experiment 1

Figure 9 summarizes the distribution of domain-specific OLS coefficients across the four domains, with text length and syntactic complexity included as control variables. Each point denotes the mean coefficient across domains, and the error bar denotes ± 1 standard deviation. After controlling for these confounds, most feature effects attenuate, and several apparent domain-wise patterns fail to persist as stable cross-domain trends.

Metaphoricity $\mu(T)$ remains the clearest exception. Its coefficient is generally positive for AEE ($\beta > 0$, indicating larger prediction errors), positive for entropy, and negative for $P(\text{true})$, with relatively consistent direction across models. The magnitudes remain modest, but among all observational features, overall metaphorical loading is the only one that repeatedly points toward higher uncertainty and lower reliability.

Explicitness $\varepsilon(T)$ shows a largely opposite pattern, with higher explicitness generally associated with lower AEE and entropy and higher $P(\text{true})$. Overt comparison markers may reduce interpretive ambiguity, although substantial between-domain variation indicates that this tendency is not uniform across genres.

K Detailed Effect of metaphor novelty on model judgment

Figure 10 reports Cohen’s d for the novelty contrast across decision margin, entropy, and $P(\text{High})$.

Novel metaphor changes judgment direction more consistently than certainty. The clearest effect is the upward shift in $P(\text{High})$: across nearly all model–domain conditions, novel metaphor increases the probability of a *High* coherence judgment ($d \approx 0.1\text{--}0.3$). Rather than behaving like a disruption cue, metaphorical novelty is more often treated as a positive signal of coherence or discourse quality.

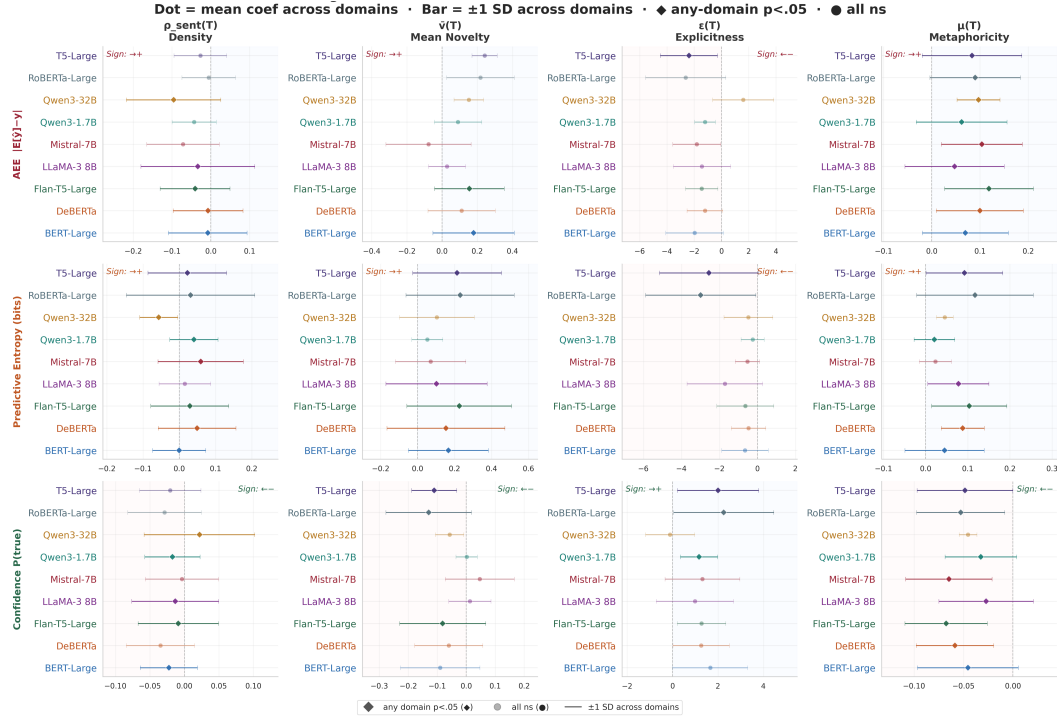


Figure 9: Cross-model OLS coefficient forest plot: each panel reports the mean coefficient across domains (dot) with between-domain variability (± 1 SD, error bar) for the four focal metaphor features on three outputs (AEE, entropy, confidence), while controlling for text length and syntactic complexity.

Certainty effects are narrower and depend on architecture and genre. Changes in margin and entropy are smaller and less uniform. Encoder–Decoder models show the clearest pattern, with higher margin and lower entropy under novel metaphor, suggesting that these models not only shift toward *High* coherence but also do so more decisively. Encoder-only models remain closer to zero, while Decoder-only models are more mixed. The domain pattern points in the same direction: in some genres, novelty raises $P(\text{High})$ and sharpens the judgment, while in others it mainly shifts the label without resolving uncertainty.

L Detailed Probability Calibration Analysis

Table 6 summarizes calibration shift as $\Delta \text{ECE} = \text{ECE}_{v(T)>1} - \text{ECE}_{v(T)\leq 1}$, where positive values indicate worse calibration under novel metaphor and negative values indicate improvement. The results reveal a clear asymmetry: calibration worsens in 25 of the 36 model–domain conditions, improves in 9, and stays roughly unchanged in 2.

Degradation also clusters by domain. Novel metaphor worsens calibration for seven of the nine models in both *yahoo* and *yelp*, and for six models in *enron*, while *clinton* is split almost evenly between degradation and improvement. This does not reduce to a simple formal–informal distinction. In heterogeneous Q&A and evaluative reviews, metaphorical novelty may compound the already greater variation in topic, stance, and expressive intensity, leaving confidence estimates less stable. In the more formulaic corporate-email register, novelty may instead stand out against a stable institutional backdrop. The mixed results for political email suggest that rhetorical framing does not impose a consistent calibration effect across models.

Architecture accounts for only part of this variation. Encoder-only and encoder–decoder models degrade in most conditions, but the decoder-only models split into two groups:

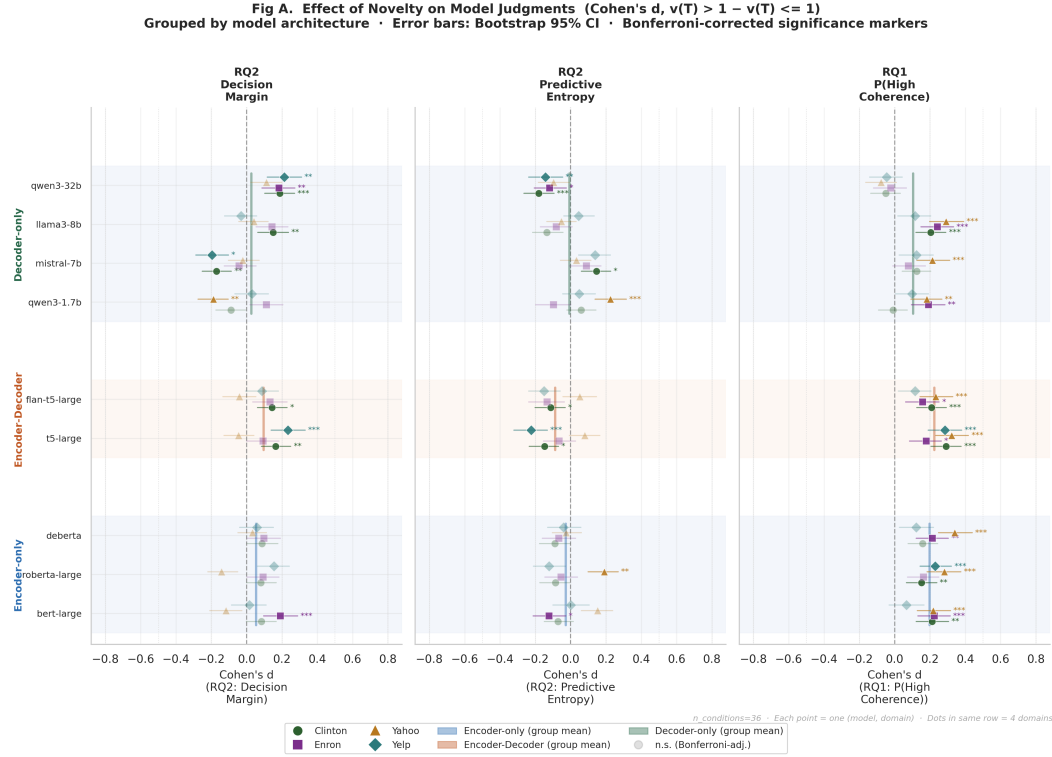


Figure 10: **Effect of metaphor novelty on model judgments (RQ1 & RQ2).** Cohen's d for novel ($v(T) > 1$) vs. conventional ($v(T) \leq 1$) metaphor across decision margin, entropy, and $P(\text{High})$. Models grouped by architecture; colored shapes indicate domains; horizontal bars show group means; error bars are bootstrap 95% CIs. Significance: Bonferroni-corrected (* $p < .05$, ** $p < .01$, *** $p < .001$); gray = non-significant.

both Qwen variants worsen across all four domains, whereas LLaMA3-8B and Mistral-7B improve in several conditions, notably *enron* and *yahoo*, while still degrading in *yelp*. That two models from the same family diverge so clearly indicates that calibration sensitivity is governed not by architecture alone, but by how a model's confidence behavior interacts with the discourse conditions of each domain. Metaphor, in short, changes *what* models predict more consistently than *how reliably* they express those predictions.

M Detailed Feature Modulation of Probability-space Outputs

Figure 11 reports feature-level correlations between discourse-level metaphor features and probability-space outputs. Unlike the global novelty contrast in RQ1–RQ3, these modulation effects depend jointly on the metaphor feature, model, and domain, while also revealing several feature-specific regularities. Density $\rho(T)$ shows the clearest cross-output pattern in *clinton*, where higher density is associated with larger decision margins, higher $P(\text{High})$, and lower entropy for most models. This alignment across the three outputs suggests that metaphor density in this domain is not only associated with a greater preference for the *High* label, but also with sharper judgments. Density is also positively related to $P(\text{High})$ in many conditions outside *clinton*, although its effects on margin and entropy are less consistent across the other domains.

Mean novelty $v(T)$ and novelty variability $\sigma_v(T)$ exhibit an even more systematic effect on judgment direction. In both *enron* and *yahoo*, mean novelty is significantly associated with $P(\text{High})$ for every evaluated model: the association is positive for eight of the nine models, with Qwen3-32B showing a consistent negative relationship. Novelty variability closely

Model	Clinton	Enron	Yahoo	Yelp	# worse
<i>Encoder-only</i>					
BERT-LARGE	▼ 0.028	▲ 0.031	▲ 0.023	▲ 0.067	3/4
ROBERTA-LARGE	▲ 0.007	▼ 0.039	▲ 0.048	▲ 0.046	3/4
DEBERTA	▼ 0.029	▲ 0.010	▲ 0.059	— 0.002	2/4
<i>Encoder-Decoder</i>					
FLAN-T5-LARGE	▲ 0.025	▲ 0.029	▲ 0.042	— 0.003	3/4
T5-LARGE	▼ 0.028	▲ 0.050	▲ 0.026	▲ 0.052	3/4
<i>Decoder-only</i>					
LLAMA3-8B	▲ 0.022	▼ 0.064	▼ 0.005	▲ 0.017	2/4
MISTRAL-7B	▼ 0.025	▼ 0.064	▼ 0.020	▲ 0.037	1/4
QWEN3-1.7B	▲ 0.098	▲ 0.077	▲ 0.036	▲ 0.007	4/4
QWEN3-32B	▲ 0.072	▲ 0.029	▲ 0.024	▲ 0.032	4/4

Table 6: Calibration shift under novel metaphor, measured as $\Delta ECE = ECE_{v(T)>1} - ECE_{v(T)\leq 1}$. ECE uses eight equal-width bins over $P(\text{High})$. Positive values (▲) indicate worse calibration under novel metaphor; negative values (▼) indicate improvement; changes with $|\Delta ECE| \leq 0.003$ are marked as —.

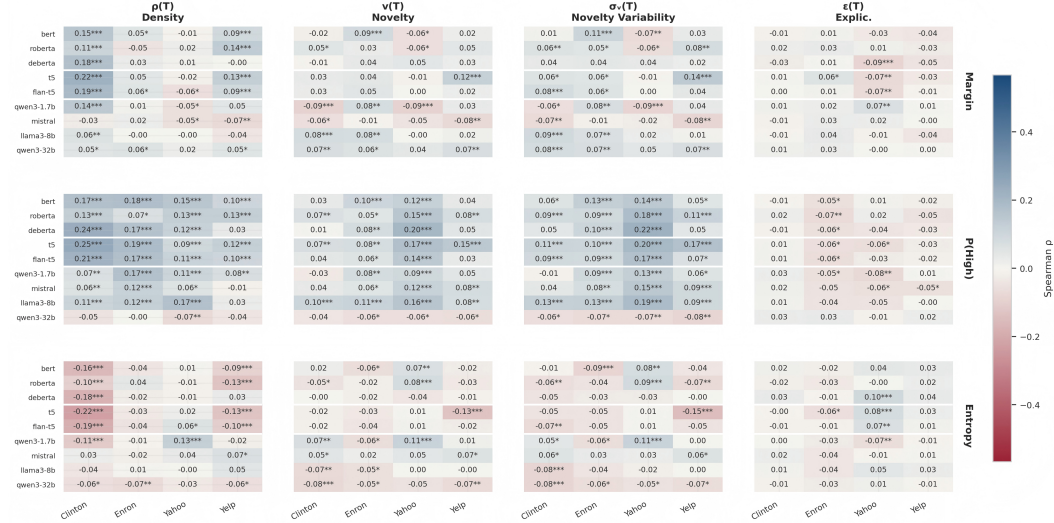


Figure 11: **Feature modulation of probability-space outputs (RQ4).** Spearman ρ between metaphor features (density, novelty, novelty variability, explicitness) and model outputs (margin, $P(\text{High})$, entropy). Rows = models; columns = domains. Significance: BH-FDR corrected (* $p < .05$, ** $p < .01$, *** $p < .001$).

mirrors this pattern and often yields correlations of comparable or greater magnitude, suggesting that the mixture of more and less conventional metaphors within a discourse may matter alongside its average novelty. Similar, though less uniform, positive associations also appear in *yelp*.

Importantly, these strong effects on $P(\text{High})$ do not translate into a uniform change in confidence. For some models, higher novelty is accompanied by larger margins and lower entropy, indicating a sharper shift toward *High*; for others, particularly in *yahoo*, the same increase in $P(\text{High})$ co-occurs with smaller margins or higher entropy. Novelty can therefore alter the preferred coherence label without necessarily making that preference more decisive.

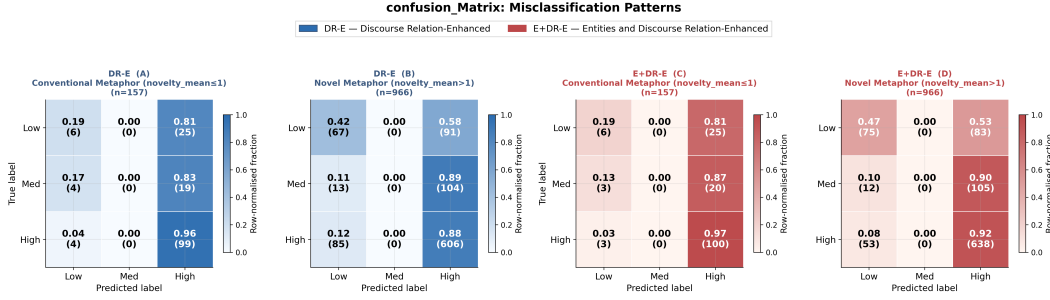


Figure 12: Row-normalized confusion matrices for the reproduced DR-E and E+DR-E coherence models under conventional ($\nu(T) \leq 1$) and novel ($\nu(T) > 1$) metaphor conditions, pooled across the four GCDC domains. Raw item counts are shown in parentheses.

Qwen3-32B provides a systematic reversal: higher novelty and novelty variability are associated with lower $P(\text{High})$ but larger margins and lower entropy across domains, indicating a more confident shift away from the *High* label rather than weak sensitivity to novelty.

By contrast, explicitness $\varepsilon(T)$ exhibits smaller and more localized effects, with no stable cross-domain pattern. Overall, density produces the clearest localized alignment between judgment direction and confidence, whereas novelty and novelty variability exert a stronger and more broadly shared influence on which coherence label models prefer. Their less consistent effects on margin and entropy further show that metaphor-related shifts in judgment direction and judgment sharpness are related but separable aspects of model behavior.

N Details of Representation-Space Analysis

Implementation details. The 49 complete four-condition items comprise 15 *clinton*, 12 *enron*, 9 *yahoo*, and 13 *yelp* items. For each condition, we locate the target span in the rewritten text using character-level anchoring and tokenizer offset mappings. When the target span is split into multiple subword tokens, we average the hidden states of all tokens in the span to obtain the layer-specific target representation. The main analysis uses target-span encoder representations from Flan-T5-Large; the corresponding results are reported in Figure 6.

Illustrative example. The following example illustrates the four matched conditions.

- **Original:** “The lawyer defended his client.”
- **Meaning-preserving metaphor:** “The lawyer shielded his client.”
Different wording, with the intended meaning preserved.
- **Meaning-changing literal:** “The lawyer attacked his client.”
Literal wording, but with a changed meaning.
- **Meaning-changing metaphor:** “The lawyer roasted his client.”
Metaphorical wording, but with a changed meaning.

O Detailed Results for Reproduced Discourse-Structured GCDC SOTA Systems

This appendix reports detailed results for two reproduced RoBERTa-based GCDC SOTA systems: a *Discourse Relation-Enhanced* model (DR-E) from Liu & Strube (2025a) and an *Entities and Discourse Relation-Enhanced* model (E+DR-E) from Liu & Strube (2025b). Both are fusion-style coherence models rather than prompt-based variants. This distinction matters for our analysis: DR-E explicitly injects discourse relation structure into coherence prediction, while E+DR-E strengthens the model further by jointly encoding both discourse relations

and entity continuity. Our goal is therefore not simply to ask whether these stronger systems are still affected by metaphor, but to test how increasingly explicit discourse structure changes what metaphor probes in coherence prediction.

Aggregate class-wise patterns. Before examining the domain-specific effect sizes, Figure 12 summarizes the pooled confusion patterns of the two reproduced discourse-structured systems. Predictions from both models remain strongly concentrated on the *High* label, while the *Medium* category is rarely predicted under either metaphor condition. Novel metaphor does not affect all gold classes uniformly: *Low*-class recall increases, whereas *High*-class recall decreases, and *Medium*-labeled items remain predominantly assigned to *High*. These pooled matrices suggest that novel metaphor reshapes the class-wise error profile rather than uniformly degrading performance. As the domain-level analyses below show, the consequences of this redistribution vary substantially across *clinton*, *enron*, *yahoo*, and *yelp*.

O.1 Judgment Shifts: Relation Enhancement Makes Genre Dependence More Visible

Figure 13 shows Cohen’s d for the novelty contrast across decision margin, predictive entropy, and $P(\text{High})$.

Both reproduced systems exhibit internally coherent but sharply domain-specific response profiles. In *yelp*, novel metaphor increases decision margin and $P(\text{High})$ while reducing entropy, indicating a sharper shift toward the *High* label. *Clinton* shows the opposite pattern: novelty lowers margin and $P(\text{High})$ and raises entropy, suggesting weaker and less certain high-coherence judgments. *Enron* follows the same directional pattern as *yelp*, although the effects are smaller and become statistically reliable mainly for E+DR-E. By contrast, *yahoo* remains close to zero across all three outputs, providing little evidence of a systematic novelty effect.

These reversals indicate that discourse-structured models do not treat metaphor as a uniform stylistic cue. A plausible interpretation is that novel metaphor interacts with the discourse and register expectations learned for each domain: it may align with evaluative language in review discourse, remain compatible with some relation patterns in corporate email, but interfere with the rhetorical or relational cues used in political email. Because these discourse functions are not directly annotated, this account should be treated as an interpretation of the observed domain profiles rather than a tested mechanism.

E+DR-E largely preserves the direction of the DR-E effects but does not uniformly amplify them. Its clearest strengthening appears in *yelp* and *enron*, whereas the two systems behave similarly in *clinton* and remain near zero in *yahoo*. Adding entity-continuity information therefore appears to reinforce metaphor sensitivity selectively, when entity and discourse-relation signals remain compatible with the rewritten expression, rather than increasing sensitivity in every domain.

O.2 Calibration: Structural Signals Reshape Reliability in a Domain-Specific Way

Figure 14 reports expected calibration error (ECE) by domain and the calibration shift

$$\Delta\text{ECE} = \text{ECE}_{v(T)>1} - \text{ECE}_{v(T)\leq 1}.$$

For DR-E, novel metaphor improves calibration in *clinton*, *yahoo*, and *yelp*, while producing almost no change in *enron*. E+DR-E preserves the improvements in *clinton* and *yelp*, with larger reductions in ECE, and remains close to neutral in *yahoo*. The main reversal occurs in *enron*, where adding entity continuity changes a nearly neutral DR-E response into clear calibration degradation.

These results show that stronger discourse-structural modeling does not have a uniform effect on confidence reliability. Instead, entity and relation signals reshape the calibration response selectively across domains. In *clinton* and *yelp*, the added structure appears to make confidence estimates more reliable under novelty, whereas in *enron* it increases the mismatch between confidence and correctness.

The calibration shifts also do not simply mirror the judgment shifts reported in Figure 13. For example, novel metaphor produces a sharper shift toward *High* coherence in *yelp* while improving calibration, whereas the corresponding shift in *enron* is accompanied by worse calibration for E+DR-E. Thus, metaphor can alter what a model predicts and how reliably it expresses that prediction through partially independent, domain-conditioned pathways.

O.3 Feature Modulation: Discourse Structure Reconfigures Domain-Specific Associations

Figure 15 reports Spearman correlations between discourse-level metaphor features and model outputs. The results reveal distinct modulation patterns for density, novelty-related features, and explicitness.

Density $\rho(T)$ provides the most consistent directional signal. It is positively associated with $P(\text{High})$ in all four domains for both systems, with the strongest effects in *clinton*. In that domain, higher density is also associated with larger margins, lower entropy, and greater adjacent-sentence similarity, indicating a coordinated shift toward sharper *High*-coherence judgments. Outside *clinton*, density continues to shift label preference upward, but its relationships with confidence and representation similarity are less uniform.

Mean novelty $\nu(T)$ and novelty variability $\sigma_\nu(T)$ exhibit closely parallel but more strongly domain-conditioned patterns. In *clinton*, both features are associated with lower margins and $P(\text{High})$, higher entropy, and lower cosine similarity. In *enron*, they instead increase margins and cosine similarity while reducing entropy, but produce little corresponding change in $P(\text{High})$. This suggests that novelty sharpens the model’s judgment in this domain without substantially changing its preferred label. In *yahoo*, novelty mainly reduces $P(\text{High})$ and representation similarity, with little systematic change in margin or entropy. By contrast, the positive novelty pattern in *yelp* becomes clear under E+DR-E, where higher novelty is associated with larger margins, higher $P(\text{High})$, lower entropy, and greater cosine similarity.

E+DR-E therefore does not uniformly amplify the DR-E associations. Instead, entity-continuity information selectively strengthens some feature–output relationships, particularly the density pattern in *clinton* and the novelty pattern in *yelp*, while weakening or reconfiguring others. Explicitness $\varepsilon(T)$ remains more localized: its effects are close to zero in *clinton*, but in *yahoo* higher explicitness is associated with lower $P(\text{High})$ and cosine similarity, and under E+DR-E also with smaller margins and higher entropy.

Overall, the added discourse structure does not simply increase or suppress metaphor sensitivity. It changes which metaphor features contribute to label preference, confidence, and representation similarity in each domain. The results also show that these outputs can dissociate: a feature may sharpen a judgment without shifting it toward *High*, or change label preference without producing a corresponding change in confidence.

P Robustness to Additional Contextual-Plausibility Filtering

To assess whether the intervention results were driven by residual contextually implausible rewrites, we applied an additional plausibility filter to all rewritten texts using the then-current flagship reasoning model, GPT-5.6 Sol in ChatGPT (accessed August 2026). Given the surrounding discourse context and the rewritten text, the model was asked only whether the rewrite made sense in context. The filter did not use the gold coherence label or any downstream model prediction. We removed rewrites that failed this check and reran Experiment 2 using the same model settings and analysis procedures as in the main experiment.

Data retention. Table 7 summarizes the retained data. Overall, 2,348 of 3,033 rewrites were retained (77.4%). Retention was similar across domains, ranging from 75.4% in *yahoo* to 78.6% in *clinton*, indicating that the filter did not disproportionately remove data from a particular domain. Retention varied more across rewrite targets: 85.7% of literal rewrites, all medium-novelty rewrites, and 75.0% of high-novelty rewrites were retained. The filtered

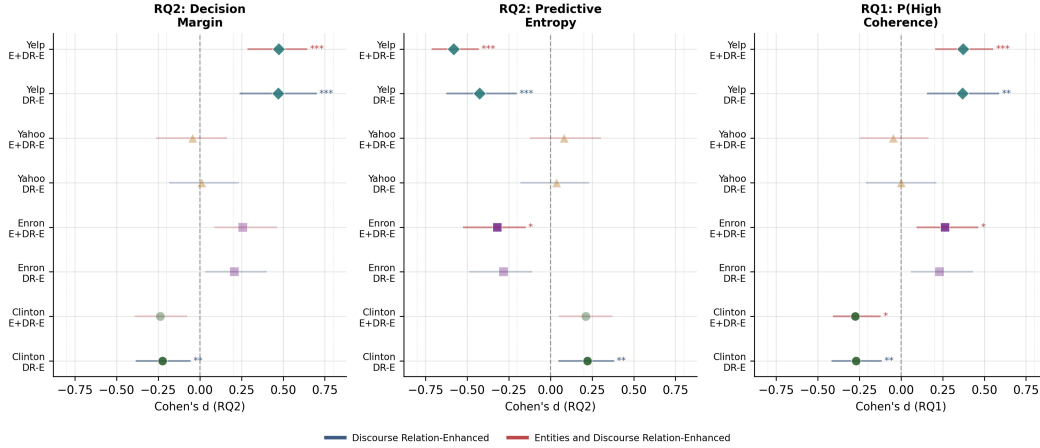


Figure 13: Novelty effects on judgment direction and uncertainty for the reproduced ROBERTa-based DR-E and E+DR-E models.

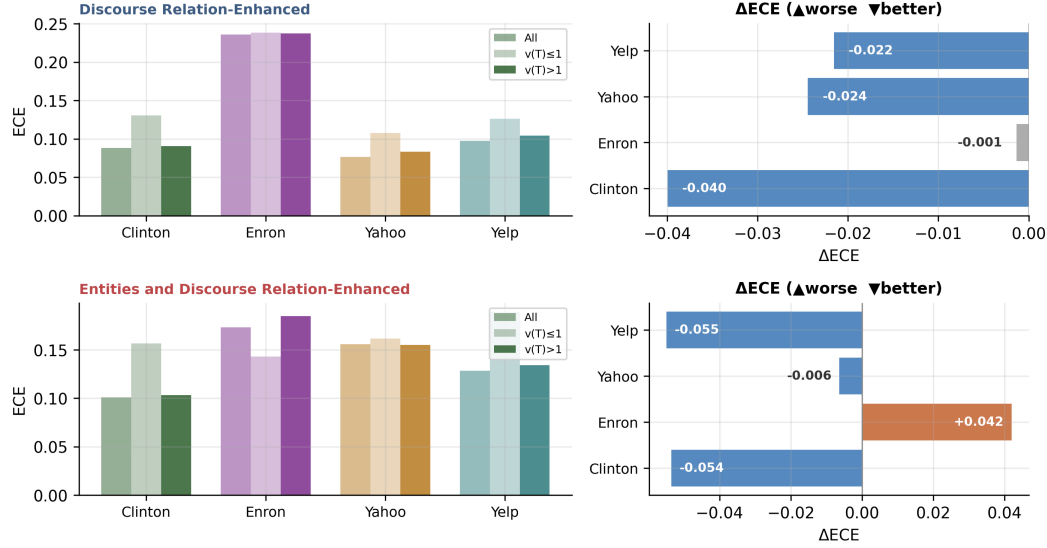


Figure 14: Calibration behavior under novelty for the reproduced ROBERTa-based DR-E and E+DR-E models.

experiment therefore tests the intervention on a stricter subset, particularly among the high-novelty rewrites.

General-model robustness. The principal judgment effects remain stable after filtering (Table 8). Across the nine general models, effect direction is preserved in 35 of 36 model-domain conditions for $P(\text{High})$ and in all 36 conditions for margin and entropy. The mean effect size for $P(\text{High})$ changes only slightly, from $d = 0.162$ to $d = 0.152$, and the condition-level estimates remain strongly correlated ($r = 0.988$). Although the number of Bonferroni-significant effects decreases from 21 to 18, the domain-level ordering is unchanged ($\rho = 1.00$): *yahoo* > *enron* > *clinton* > *yelp*. Calibration is similarly stable in aggregate, with mean ECE changing from 0.121 to 0.120, mean ΔECE from 0.019 to 0.018, and shift direction preserved in 35 of 36 conditions.

SOTA-model robustness. The reproduced SOTA systems also retain the main judgment patterns. Effect direction is preserved in 7 of 8 conditions for $P(\text{High})$ and in all 8 conditions

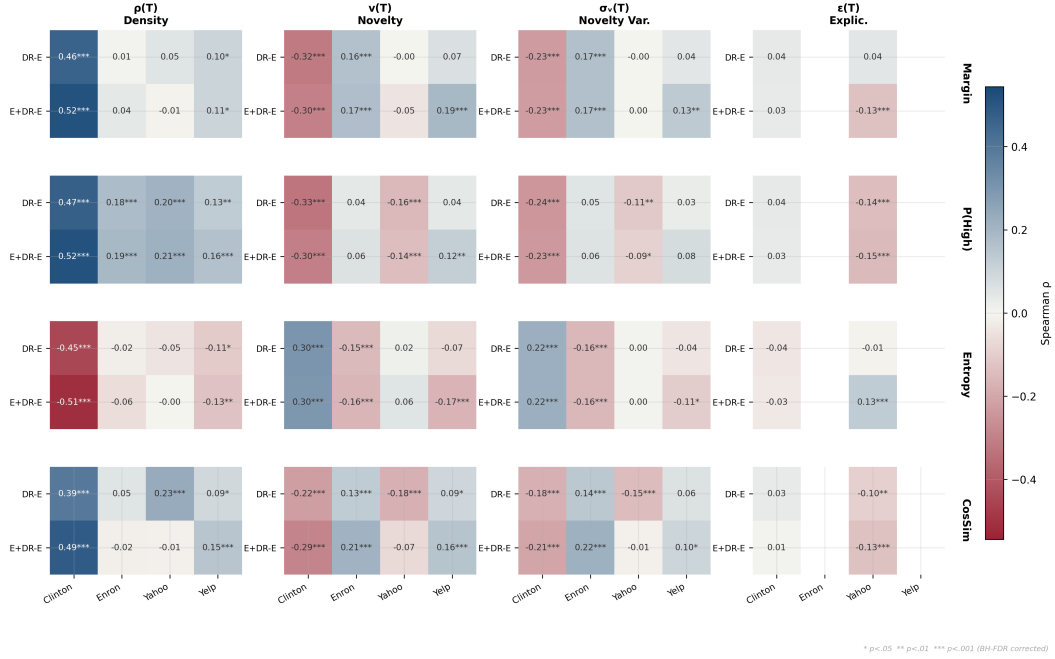


Figure 15: Feature-level modulation under novelty for the reproduced RoBERTa-based DR-E and E+DR-E models.

(a) Retention by domain			
Domain	Before	Retained	Rate (%)
Clinton	1,040	817	78.6
Enron	648	504	77.8
Yahoo	771	581	75.4
Yelp	574	446	77.7
All	3,033	2,348	77.4

(b) Retention by rewrite target			
Target	Before	Retained	Rate (%)
Literal baseline	446	382	85.7
Medium novelty	103	103	100.0
High novelty	2,484	1,863	75.0

Table 7: Data retained after additional contextual-plausibility filtering, reported by domain and rewritten novelty target.

for margin and entropy. The mean $P(\text{High})$ effect changes from $d = 0.080$ to $d = 0.097$, with near-perfect agreement between condition-level estimates ($r = 0.998$). The domain ordering also remains unchanged ($\rho = 1.00$): $yelp > enron > yahoo > clinton$. Calibration is somewhat less stable: mean ECE changes from 0.132 to 0.138, mean ΔECE from -0.020 to -0.011 , and shift direction is preserved in 6 of 8 conditions.

Summary. The filtered analysis reproduces the main qualitative conclusions obtained from the unfiltered data. Judgment directions, effect sizes, and domain-level patterns remain highly consistent, while calibration continues to show greater condition-level variation than judgment shifts. The central findings of Experiment 2 therefore do not depend on the rewrites excluded by the additional contextual-plausibility filter.

(a) Statistics before and after filtering				
Statistic	General models (9×4)		SOTA models (2×4)	
	Before	After	Before	After
Mean $d_{P(\text{High})}$	0.162	0.152	0.080	0.097
Bonferroni-significant $P(\text{High})$ effects	21/36	18/36	5/8	4/8
Mean ECE	0.121	0.120	0.132	0.138
Mean ΔECE	0.019	0.018	-0.020	-0.011

(b) Cross-condition stability after filtering		
Stability measure	General models	SOTA models
Correlation of $d_{P(\text{High})}$ estimates	$r = 0.988$	$r = 0.998$
Same-sign $P(\text{High})$ effects	35/36 (97.2%)	7/8 (87.5%)
Same-sign margin effects	36/36 (100%)	8/8 (100%)
Same-sign entropy effects	36/36 (100%)	8/8 (100%)
Domain-rank correlation	$\rho = 1.00$	$\rho = 1.00$
Same-sign ΔECE effects	35/36 (97.2%)	6/8 (75.0%)

Table 8: Comparison of Experiment 2 results before and after additional contextual-plausibility filtering. Panel (a) reports the principal aggregate statistics. Panel (b) reports agreement between the filtered and unfiltered analyses across model–domain conditions. Same-sign counts indicate conditions in which the estimated effect remained positive or negative after filtering. Pearson’s r compares condition-level $P(\text{High})$ effect sizes, and Spearman’s ρ compares domain-level effect rankings.