Right for the Wrong Reasons: Investigating the Unrobust Nature of High Accuracy Models in Discourse Relation Prediction

Yi Fan and Wei Liu
Heidelberg Institute for Theoretical Studies
yi.fan@h-its.org, wei.liu.llm@gmail.com

Abstract

Large language models (LLMs) can achieve strong performance on discourse relation prediction through fine-tuning. Yet it remains unclear whether such performance reflects a robust modelling of inter-sentential relations or a reliance on shallow lexical cues. This paper presents a behavioural audit of fine-tuned models across explicit and implicit settings. In explicit discourse relation prediction, we find that fine-tuned models rely heavily on connectives as categorical triggers, rather than integrating argument semantics. More surprisingly, in implicit discourse settings, we observe a severe performance degradation when masking semantically light function words from the arguments, suggesting that models may also exploit subtle surface-level cues beyond explicit connectives. Consequently, these results suggest that high accuracy alone is insufficient to demonstrate robust discourse processing. To address this limitation, we propose an intervention-based diagnostic framework that measures whether model predictions are grounded in linguistically meaningful evidence. Our metric provides a complementary perspective to standard accuracy-based evaluation and establishes a rigorous auditing toolkit for future work on discourse relation prediction.

1 Introduction

Discourse relations describe how sentences or clauses are logically and rhetorically connected in a text. In explicit discourse relations, connectives such as *because* or *however* provide clear clues about the intended relation, as shown in Example (1). However, a robust interpretation should still depend on the meaning of the two arguments, rather than on the connective alone. In implicit discourse relations, as in Example (2), no connective is present, so the relation has to be inferred mainly from the meaning of the arguments.

(1) [Friday's sell-off was a good thing,]$_{\text{Arg1}}$ **because** [it deflated a lot of crazy takeover speculation.]$_{\text{Arg2}}$

(2) [People think of the steel business as an old and mundane smokestack business,]$_{\text{Arg1}}$ (**Implicit=However**) [They're dead wrong.]$_{\text{Arg2}}$

Recent studies have shown that large language models (LLMs) can perform well on discourse relation prediction (DRP) (Fan et al., 2026; Wang et al., 2025b). However, high accuracy alone can lead to a wrong assessment of performance. In particular, it remains unclear whether fine-tuned LLMs learn to capture the semantic relation between the two arguments or rely on shallow lexical cues associated with relation labels. This vulnerability is critical: models that exploit lexical shortcuts will feed incorrect semantic information to downstream tasks such as summarisation or essay scoring (Liu et al., 2025; Liu and Strube, 2025a,b), causing unreliable results when exposed to textual variations.

In this paper, we systematically examine this question by analysing fine-tuned LLMs in both explicit and implicit discourse relation prediction. For explicit relations, discourse connectives provide strong clues, but they may also lead models to ignore the arguments. We therefore test whether model predictions remain consistent when connectives are removed or replaced. Our results show that fine-tuned models rely heavily on connectives: masking them leads to a large drop in performance, and replacing them can substantially change model predictions even when the argument content remains unchanged. These findings suggest that models often treat connectives as categorical triggers for relation labels, rather than as interpretative clues that should be processed alongside the semantic content of the arguments.

We further show that this reliance on surface cues is not limited to explicit discourse relations. In implicit relation prediction, where no connective

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 713–728
November 6-10, 2026 ©2026 Association for Computational Linguistics

is present, models still appear to rely on subtle lexical patterns. In particular, masking semantically light function words (e.g., "in", "of") from the arguments results in a clear drop in performance, suggesting that models may rely on subtle surface-level cues beyond explicit discourse connectives. This finding challenges the assumption that strong performance on implicit DRP necessarily reflects a robust performance of the relation between the two arguments.

These findings raise a broader concern about standard accuracy-based evaluation. If a model predicts the correct label for the wrong reason, high accuracy may overestimate its discourse modelling capabilities. To reveal this problem, we propose a diagnostic evaluation framework that measures whether model predictions are supported by linguistically meaningful evidence. This metric complements standard accuracy and F1 scores by evaluating not only whether the prediction is correct, but also whether it is supported by the relevant parts of the input. Our contributions are as follows:

- We introduce an intervention-based diagnostic framework that provides a fine-grained quantification of asymmetric dependencies between discourse arguments. Aligning token-level attributions with structural perturbations complements standard accuracy-based evaluations.
- We conduct a multi-level behavioural audit of fine-tuned LLMs for discourse relation prediction by using this framework, uncovering their heavy reliance on explicit connectives and function-word shortcuts.
- We provide empirical evidence across multiple architecture families, demonstrating that high fine-tuned accuracy stems from surface anchoring rather than compositional semantic processing.

2 Related Work

Ambiguity and the Accuracy Trap Discourse model performance is traditionally measured via accuracy and F1 scores (Liu et al., 2023a). However, explicit connectives are inherently ambiguous in usage, sense, and scope (Webber et al., 2019a). While early work used syntax to resolve specific types of ambiguity (Pitler and Nenkova, 2009), modern research largely focuses on achieving high classification scores (Liu and Strube, 2023; Fan et al., 2026). Since simple classifiers can exceed 90% accuracy on PDTB 2 (Prasad et al., 2008) and PDTB 3 (Webber et al., 2019b), researchers may overlook ambiguous cases. This reliance on traditional metrics has been criticised for masking model fragility (Collot et al., 2026), leading to models being "right for the wrong reasons."

Shortcut Learning and Heuristics A major concern in modern NLP is the tendency of models to rely on "shortcuts" rather than actual reasoning (Geirhos et al., 2020). These shortcuts can be categorised into occurrence, style, and concept biases (Zhou et al., 2024). In tasks such as Natural Language Inference (NLI), state-of-the-art models have been found to rely on lexical overlap and syntactic heuristics rather than on semantic processing (McCoy et al., 2019). In discourse relation prediction tasks, models often learn simple one-to-one mappings between connectives and labels, failing to capture the underlying semantic links. This issue persists in LLMs (Yuan et al., 2024), which remain highly sensitive to prompts (Habba et al., 2025) and perform poorly with simple word-level perturbations such as spelling errors or missing words (Alahmari et al., 2025; Wang et al., 2025a; Agrawal et al., 2025; Lou et al., 2024).

Behavioural Testing and Model Auditing Researchers introduce "behavioural testing", also known as black-box testing (Beizer and Wiley, 1996), which validates system capabilities by examining input-output behaviour, to systematically expose these hidden shortcuts without assuming knowledge of internal structure. Such methods reveal that high-accuracy models often remain fragile under linguistic perturbations (Ribeiro et al., 2020). Consequently, there is growing interest in auditing LLMs, a systematic diagnostic process for tracing the rationale behind predictions, to determine whether they capture critical semantic cues or follow surface-level heuristics (Chataigner et al., 2026; Amirizaniani et al., 2024; Mökander et al., 2023; Pucci and Ranaldi, 2025). Recent work shows that LLMs rely heavily on surface-level lexical patterns rather than abstract linguistic competence, showing high sensitivity to meaning-preserving lexical substitutions (Kostić et al., 2026).

Specific Evidence in Discourse Analysis Specific studies in discourse relation prediction have already hinted at these hidden shortcuts. For in-

stance, an ablation study on Japanese discourse models found that removing the second argument ($Arg2$) resulted in little performance loss, whereas removing the first argument ($Arg1$) caused a significant drop (Sato et al., 2025). Since the core of the task is to identify the relationship *between* two segments, this imbalance strongly suggests that the fine-tuned model relies on a single-sided shortcut rather than processing the interaction between both arguments. Moreover, Kim et al. (2020) found that while simple fine-tuning models can achieve good performance, they overrelied on lexical cues, and a simple lexical perturbation with no semantic effect can cause a performance drop. This finding raises a concern about whether models predict the relation by capturing the semantic relation or by matching shallow patterns.

Motivation for This Study Inspired by these findings, our work bridges the gap between high accuracy and robust performance. While previous research has identified shortcuts in NLI and other tasks (Geirhos et al., 2020; McCoy et al., 2019), a systematic audit of the internal consistency and argument balance in discourse relation prediction remains lacking. Building on the principles of behavioural testing (Ribeiro et al., 2020) and evidence of argument imbalance in Japanese discourse studies (Sato et al., 2025), we propose a multi-level auditing framework. Unlike traditional evaluation that relies solely on classification scores, our approach quantifies the degree of reliance on surface-level markers versus deep semantic content. While previous work has identified shortcuts, our framework provides the first systematic measurement of asymmetric dependency. This allows us to move beyond simple observation and explore whether performance gains stem from robust reasoning or shallow heuristics.

3 Experiment Setting

We evaluate on the PDTB 2 (Prasad et al., 2008), PDTB 3 (Webber et al., 2019b), and eng.pdtb.gum (Liu et al., 2024b) datasets. The eng.pdtb.gum dataset from the 2025 DISRPT shared task (Braud et al., 2025) is annotated with 15 text genres based on the GUM corpus (Zeldes, 2017), allowing us to validate our framework beyond standard news corpora. Note that we only use the implicit discourse data from eng.pdtb.gum dataset. The models we use are Llama-3-8b, Qwen3-4b, and Qwen3-8b (Grattafiori et al., 2024; Yang et al., 2025) from Huggingface[1]. All experiments were run using simple instructions and inputs, including two arguments and the connective. The motivation is to minimise the effect of other words in the prompt. The prompt we use is shown in Appendix D. The prediction is obtained by selecting the label with the highest probability among all label candidates. We use LoRA to fine-tune the model (Hu et al., 2022), with alpha set to 128 and rank to 64. The learning rate is set to 3e-5 or 5e-5, and the number of training epochs is set to 5, with early stopping to prevent overfitting. All results are averaged over three runs. For simplicity, since the results for PDTB 2.0, PDTB 3.0, and eng.pdtb.gum are similar, we report the results on PDTB 3.0 and eng.pdtb.gum in the paper. Detailed results on PDTB 2.0 are provided in Appendix A.

4 Evaluation of Explicit Discourse Prediction

In explicit discourse relations, the presence of a connective (e.g., "*but*", "*because*") provides a strong lexical cue. But whether the LLM considers the semantic information conveyed by the entire input, or simply predicts the discourse relation based on the connective, needs to be verified. In this section, we audit whether the model's prediction is genuinely grounded in the semantic interaction between arguments or merely anchored in surface-level heuristics (Chataigner et al., 2026; Amirizaniani et al., 2024; Mökander et al., 2023; Pucci and Ranaldi, 2025). Besides, we explore during fine-tuning where the performance gains come from and how the model improves.

4.1 Prediction Transition Analysis

We first evaluate the pre-trained LLMs in a zero-shot setting to establish baseline connective reliance. Since connectives strongly hint at the labels, we can assess, based on performance, whether the models can perform well when only pre-trained. Following fine-tuning, we examine which of these zero-shot errors are most corrected, thereby improving performance. Specifically, we identify the **Top Corrected Patterns** by tracking label transitions between the zero-shot and fine-tuned settings to reveal whether different pre-trained models show a uniform bias toward a single or multiple relations.

[1] https://huggingface.co/

Top Level			
Models	WP	CP	No.
Llama-3-8b	Temporal	Expansion	661
	Temporal	Comparison	325
	Temporal	Contingency	205
Qwen3-4B	Temporal	Expansion	725
	Temporal	Comparison	404
	Temporal	Contingency	212
Qwen3-8B	Temporal	Expansion	721
	Temporal	Comparison	404
	Temporal	Contingency	207
Second Level			
Llama-3-8b	Cause	Conjunction	557
	Contrast	Concession	287
	Cause	Asynchronous	165
Qwen3-4B	Manner	Conjunction	520
	Contrast	Concession	291
	Manner	Asynchronous	119
Qwen3-8B	Cause	Conjunction	433
	Contrast	Concession	316
	Contrast	Conjunction	125

Table 1: The top three corrected pattern results on the PDTB 3 across three different models. WP: wrong prediction in zero-shot setting. CP: corrected prediction in fine-tuned setting. No.: corrected numbers.

4.2 Focus Dynamics

After assessing the model's performance in zero-shot settings, we want to explore whether the fine-tuning process guides predictions from blind bias toward a more precise distinction, or whether the model shifts to the opposite extreme by relying heavily on the connective signal.

We use a Leave-One-Out (LOO) perturbation approach to analyse the importance of individual input components. This method isolates the influence of each component $i \in \{A_1, C, A_2\}$ by defining a marginal contribution ϕ_i:

$$\phi_i = \max(0, P(y_g \mid x) - P(y_g \mid x_{\setminus i})) \quad (1)$$

where $x_{\setminus i}$ denotes the input with component i masked. The value ϕ_i measures the drop in prediction probability for the correct label y_g when component i is removed. This metric identifies which parts of the text drive the final classification outputs. To provide a stable and continuous measure of component reliance, we introduce the Normalised Connective Focus (NF):

$$NF = \frac{\phi_C}{\phi_C + \phi_{ctx}} \quad (2)$$

where $\phi_{ctx} = (\phi_{A_1} + \phi_{A_2})/2$ yields the mean contextual contribution. This formulation includes a division by 2 to compute the mean contextual contribution. This normalisation is a critical design choice to maintain dimensional alignment between the connective and the argument context. The mathematical intuition is provided in Appendix F. This metric bounds the relative attention within a stable range between 0 and 1. A value of 0.5 represents a perfectly balanced focus between the explicit connective and the contextual arguments. Values approaching 1.0 signify an absolute reliance on the connective trigger. This formulation removes the dependence on an arbitrary stability term and enables reliable statistical comparisons across different relations.

4.3 Robustness Test

Having tested whether the models rely heavily on connectives, we next measure accuracy without explicit markers to assess whether the models can predict the correct relation using only the arguments. We calculate Acc_{rv} via a connective ablation test:

$$Acc_{rv} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left(\hat{y} \backslash C, i = yg, i\right) \quad (3)$$

Acc_{rv} ranges from 0 to 100%. Since the label shift phenomenon occurs when connectives are removed in explicit cases, the robustness score should ideally be around 67% in PDTB 2 and 70.4% in PDTB 3 (Liu et al., 2024a).

4.4 Adversarial Stress Metrics

Finally, we measure the sensitivity under adversarial conditions. We use intervention-based metrics to reveal whether shortcut learning exists. **Signal Sensitivity** measures vulnerability to adversarial connective substitution using $x' = \{A_1, C_{\text{adv}}, A_2\}$:

$$\text{Sensitivity} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left(f(x'i) \neq y_{gi}\right) \quad (4)$$

Scores range from 0 to 100%, with higher sensitivity indicating that the system processes connectives as categorical triggers rather than interpretative clues. A common counter-argument is that connective substitution naturally alters rhetorical intention, justifying a label shift. While this holds true for semantically loose sentences, adversarial substitution often creates severe physical and temporal contradictions (Semantic Incompatibility) (van Leusen, 2004) that a robust model should detect. Examples are illustrated in Appendix G.

	Top Level					Second Level				
Models	Acc_{ft}	Acc_{zs}	Acc_{rv}	NF	SS	Acc_{ft}	Acc_{zs}	Acc_{rv}	NF	SS
Llama-3-8b	95.13	24.79	51.33	48.98	76.12	89.55	13.12	39.93	70.31	99.77
Qwen3-4B	94.85	19.75	50.37	73.40	75.72	90.37	14.57	35.64	84.57	99.71
Qwen3-8B	95.19	20.43	50.65	79.29	75.66	90.13	16.25	36.97	84.29	99.88

Table 2: Results on PDTB 3.0. Alongside fine-tuned (Acc_{ft}) and zero-shot (Acc_{zs}) accuracies, we report connective ablation robustness (Acc_{rv}), Signal Sensitivity (SS), and Normalised Connective Focus (NF). A robust parser should achieve Acc_{rv} around 70.4%. Conversely, high SS and NF values expose the model's high reliance on connectives rather than contextual semantics.

4.5 Results and Analysis

Our evaluation results show a disconnect between standard classification accuracy and LLMs' robustness to discourse prediction. Tables 1 and 2 provide empirical evidence on the transition patterns, structural robustness, and shortcut dependencies of the fine-tuned models.

Analysis of Prediction Transitions Table 1 presents the top corrected labels from the zero-shot to the fine-tuned setting. The pre-trained models show strong systematic biases before fine-tuning. For instance, all the models demonstrate a uniform bias toward the Temporal label at the top level. This indicates that without fine-tuning, the models default to a simple time-based sequence and fail to use the connective signal. In second-level prediction, pre-trained models often fail to predict Cause and Contrast labels. These two relations normally have strong connective signals, such as "because", "therefore" and "but". These results suggest that LLMs, even after pre-training, may still fail to capture important lexical cues. While the fine-tuning process adjusts the surface label distributions, further metrics are necessary to verify whether this adjustment reflects genuine semantic learning.

The Accuracy Trap and Structural Robustness As shown in Table 2, fine-tuned accuracy (Acc_{ft}) exceeds 94% (top) and 89% (second-level) across models, while zero-shot accuracy (Acc_{zs}) remains below 25% and 17%, respectively. This gap indicates that the underlying parameters lack a robust processing of discourse relations. The robustness test under connective ablation (Acc_{rv}) shows this vulnerability. When connectives are masked, the top-level performance drops sharply to around 50% for all models. This result falls short of the ideal linguistic distribution baseline of 70.4% (Liu et al., 2024a). At the second level, the robustness scores drop further, falling below 40%. These empirical drops demonstrate that the fine-tuned models cannot use the argument context to predict once the explicit markers are masked.

Quantifying Shortcut Mechanics The Normalised Connective Focus (NF) and Signal Sensitivity (SS) indicate the mechanics of this reliance on shortcuts. The high NF scores at the second level indicate that the internal allocation of attention is heavily toward explicit connectives rather than arguments. The SS scores provide additional evidence of shortcut learning. The SS scores reach nearly 100% across models, prompting arguments that connective substitution naturally alters the relation. However, our independent implicit experiments (Section 5) refute this defence: arguments alone achieve over 72% accuracy, demonstrating that contextual semantics are sufficient. When an adversarial connective contradicts the underlying event logic, a robust parser should detect the semantic conflict and show predictive uncertainty. Instead, the unconditional output flip provides evidence of shortcut bias (McCoy et al., 2019), demonstrating that fine-tuning forces the model to treat connectives as categorical triggers while bypassing compositional semantics. We note that Llama-3-8b shows a low NF and a high SS for top-level label prediction. This indicates the model's necessary condition: both the connective and the argument context are required for high confidence, as masking either reduces the probability. The high SS suggests the explicit connective can override the final decision; replacing it introduces a semantic contradiction.

Takeaway: Highly fine-tuned accuracy in explicit discourse relation prediction can be an artefact of surface cue anchoring. Fine-tuned systems do not process the compositional meaning of arguments. Instead, the fine-tuning process establishes a fragile, one-to-one mapping between explicit connectives and relation labels, leaving the system highly vulnerable to structural perturbations. As previous research indicates, connectives can be ambiguous (Webber et al., 2019a); blindly mapping them to

labels can yield unreliable models. As discourse relations play an important role in many downstream tasks, such as essay scoring (Liu and Strube, 2025a,b), reading comprehension (Mihaylov and Frank, 2019), and text summarisation (Liu et al., 2023b, 2025), this right-for-wrong-reasons phenomenon can raise serious concerns.

5 Evaluation of Implicit Discourse Prediction

While explicit connectives act as dominant shortcuts, implicit settings lack these salient triggers. This absence raises a question about whether the model genuinely predicts based on semantic reasoning. To make our framework and results convincing, we include our test results for the implicit discourse relation test on both PDTB 3.0 and eng.pdtb.gum in this section. Our results show that the fine-tuned models fail to transition to semantic integration in different text genres. Instead, the models treat function words (e.g., prepositions) as alternative structural shortcuts.

5.1 Semantic Attribution

We define an input as a pair of arguments $x = \{A_1, A_2\}$, where the model f predicts a relation label $\hat{y} \in \mathcal{L}$. To understand which tokens are most helpful or harmful to the model's prediction, we measure the Impact $\Phi_k = \sum_{t \in \text{Tokens}_k} \max(0, P(y_g|x) - P(y_g|x_{\setminus t}))$, which quantifies the contribution of each token to the final prediction.

5.2 Reconstruction Accuracy

From the impact results, we found that models often rely on function words such as "in" and "on", and masking them leads to a large drop in probability. Based on our observations, we can still infer the correct discourse relation from a human perspective even when one function word is masked. Since LLMs are pre-trained on massive corpora for next-token prediction, predicting high-frequency function words should not be a hard task. Therefore, we measure **Reconstruction Accuracy** ($RAcc$). For each function word $t_i \in \mathcal{T}_f$ (e.g., prepositions and articles), the model predicts a replacement $\hat{t}_i$ based on the surrounding context: $\hat{t}_i = \arg\max_{w \in \mathcal{V}} P(w \mid x \setminus \{t_i\})$. Then the models predict the final label $\hat{y}$ using the reconstructed input $\hat{x} = \{\hat{t}_1, \ldots, \hat{t}_n\}$. $RAcc$ values range from 0 to 100%. In a robust model, $RAcc$ should be close to the fine-tuned accuracy (Acc_{ft}). Therefore, a large gap between these metrics reveals a logical disconnect: despite sufficient semantic context to infer the relation, the model fails to use it consistently in producing the final predicted label.

5.3 Fidelity and Causal Faithfulness

In the previous step, we calculated the impact for each token to understand which information drives a prediction. However, grounded in the principle of compositionality, tokens do not operate in isolation (Partee, 1995; Chomsky, 2001) but trigger complex, non-linear feature interactions (Chen et al., 2020). Consequently, a phrase's total impact often diverges from the simple algebraic sum of its individual parts. We map token-level attributions to human-interpretable linguistic structures (e.g., noun phrases (NP) and prepositional phrases (PP)) by using constituency parsing (via `Spacy`[2] and `Benepar`[3]). We also introduce a custom category, VERB_COMPLEX, that combines auxiliary and main verbs to study cues related to the state of an action. To quantify whether local token scores reliably extrapolate to phrase-level impacts, we define the Fidelity Correlation (r):

$$r = \text{Pearson}(\sum_{t \in p} \phi_t, \text{CausalImp}(p)) \quad (5)$$

where $\text{CausalImp}(p)$ is the causal impact of masking the entire syntactic phrase p. The values of r range from -1 to 1. A positive correlation score indicates that the token-level metrics align with the models' causal behaviour. Specifically, a high score near 1.0 indicates a strictly linear relationship in which individual word importance perfectly predicts the total phrase impact. In contrast, a moderate positive score indicates that the token-level metrics capture the correct causal relation, while the remaining gap reflects nonlinear interactions among tokens within the neural architecture (Lundberg and Lee, 2017; Jacovi and Goldberg, 2020).

5.4 Functional Reliance

Our previous analysis of token-level importance reveals a strong reliance on semantically light prepositions during implicit relation prediction. To test whether the classification process relies on semantic or anchors on structural placeholders, we introduce the Mean Surprisal Drop (ΔS). This metric

[2] https://spacy.io/

[3] https://spacy.io/universe/project/self-attentive-parser

	Top Level					Second Level				
Models	Acc$_{ft}$	Acc$_{zs}$	RAcc	FC*	MSD*	Acc$_{ft}$	Acc$_{zs}$	RAcc	FC*	MSD*
Llama-3.1-8B	74.42	12.42	56.51	0.36	2.09	55.53	23.06	35.54	0.34	1.60
Qwen3-4B	72.12	10.72	55.56	0.37	2.36	53.75	10.03	38.34	0.34	2.22
Qwen3-8B	74.49	12.28	60.79	0.35	2.69	55.53	23.53	40.93	0.35	3.06

Table 3: Results on PDTB 3.0 dataset. In addition to standard fine-tuned accuracy (Acc_{ft}), zero-shot accuracy (Acc_{zs}), and function-word reconstruction accuracy ($RAcc$), we report Fidelity Correlation (FC) and Mean Surprisal Drop (MSD). FC measures whether our framework is intervention-based (ideally, a positive value > 1). For MSD, values approaching 0 indicate ideal robustness, while scores > 0.5 denote exponential confidence loss. Metrics with * indicate that the values are not percentages.

Models	Acc$_{ft}$	Acc$_{zs}$	RAcc	FC*	MSD*
Llama-3.1-8B	64.28	20.53	44.30	0.26	2.45
Qwen3-4B	65.96	13.36	49.93	0.22	2.71
Qwen3-8B	65.26	13.64	49.79	0.23	3.15

Table 4: Results on eng.pdtb.gum dataset.

evaluates prediction certainty under a stress test where all function words are masked from the input arguments. Adapting the concept of surprisal from information theory (Shannon, 1948; Hale, 2001; Levy, 2008), we measure the shift in prediction confidence under structural perturbations. We calculate ΔS by quantifying the uncertainty difference before and after masking function words:

$$\Delta S = [-\log_2 P(y_g|x_{mk})] - [-\log_2 P(y_g|x_{og})] \quad (6)$$

where x_{og} represents the original complete input text and x_{mk} represents the text without functional placeholders. These function words consist of high-frequency prepositions and articles that carry minimal semantic weight. A lower ΔS value approaching 0 indicates a robust decision process. This condition indicates that the absence of structural markers does not reduce prediction confidence. Conversely, a high positive value demonstrates that classification stability is highly fragile. This increase in uncertainty demonstrates that the fine-tuned system relies heavily on surface placeholders rather than processing the core semantic meaning of the arguments.

5.5 Results and Analysis

The empirical evaluation of implicit discourse relation prediction shows that the high performance of fine-tuned models is supported by surface heuristics—a vulnerability that exists across diverse text genres. Table 3 and Table 4 provide comprehensive results across PDTB 3 and eng.pdtb.gum, respectively.

The Fine-Tuned and Reconstruction Accuracy As shown in Table 3 and Table 4, the fine-tuned accuracy (Acc_{ft}) consistently reaches high accuracy across all datasets. However, the low zero-shot accuracy (Acc_{zs}) shows that pre-trained parameters do not support robust processing of implicit discourse relations. The Reconstruction Accuracy ($RAcc$) metric further illustrates the models' dependence on function words. Given the standard pre-training process of LLMs, LLMs should recover the masked function word from context. Even if the reconstructed word differs superficially from the original, the core argument semantics and true discourse relation remain almost identical. Yet, when evaluated on these model-reconstructed inputs, classification performance degrades significantly across various domains. For instance, in the multi-genre eng.pdtb.gum dataset, $RAcc$ drops by roughly 15-20 percentage points relative to the Acc_{ft} baselines, indicating the severe degradation observed on the news-type PDTB datasets. This cross-genre result shows that function-word dependency is generic, not merely a corpus-specific artefact.

(1) [These are just baby steps toward real interactive video, which I believe will be the biggest thing yet to affect television,]$_{\text{Arg1}}$ (**Implicit=For instance**) [Although it would be costly to shoot multiple versions, TV programmers **could** let audiences vote on different endings for a movie.]$_{\text{Arg2}}$

(2) [the Clean Water Act is written **upon** the presumption–the pretense, rather– that nothing but zero risk will do]$_{\text{Arg1}}$ (**Implicit=As a result**) [it establishes a legal standard of zero discharge]$_{\text{Arg2}}$

Examples (1) and (2) from PDTB 3 illustrate this structural fragility. In (1), the model correctly predicts the gold relation (Instantiation) on the orig-

inal text, assigning a 0.942 importance score to the auxiliary "**could**". When **could** is masked, the model reconstructs it as "**can**". Although this minor substitution preserves the original Instantiation relation, the model's prediction shifts to Level-of-detail. Similarly, in (2), reconstructing the word "**upon**" as "**on**" flips the prediction from Cause to Equivalence. These failures demonstrate that the discourse prediction process is sensitive to exact lexical surface forms.

Causal Validation via Fidelity Correlation The Fidelity Correlation (FC) metric empirically validates that our token attribution method reliably reflects the underlying decision process. Across all models, prediction levels, and domains, the correlation scores remain stable within a moderately positive range. This cross-genre consistency demonstrates that our metrics align with the system's causal behaviour during phase perturbations rather than dataset-specific noise. Moreover, this moderate positive correlation is expected on mathematical grounds. As established in Lundberg and Lee (2017), individual first-order attributions cannot fully capture higher-order, non-linear feature interactions within deep neural architectures. Therefore, the remaining variance does not indicate algorithmic noise. Instead, it accurately preserves the expected non-linear behaviour of fine-tuned language models during compositional sentence processing (Jacovi and Goldberg, 2020). The remaining gap indicates that individual word importances cannot be simply summed to predict global phrase impacts, reflecting the complex, non-linear token interactions within the fine-tuned layers.

Quantifying Functional Reliance and Fragility The Mean Surprisal Drop (MSD) metric provides empirical evidence of shortcut learning. To interpret MSD, we must establish its theoretical baseline: for a robust parser that relies on core compositional semantics, masking function words should yield an MSD approaching 0 (i.e., near-zero confidence loss). While minor fractional values (e.g., $MSD < 0.5$) represent an acceptable tolerance zone for natural context disruption, our results exceed this threshold across all domains. Crucially, this fragility is across multiple genres. Even the lowest observed MSD of 1.60 indicates that the model's predicted probability of the correct labels drops by nearly a factor of 5 ($e^{1.6} \approx 4.95$). At the higher end, values reaching 3.06 imply an exponential plunge in certainty, roughly 23 times lower ($e^{3.15} \approx 23.3$). As established in previous work (Geirhos et al., 2020; McCoy et al., 2019), when neural models exhibit such severe probability degradation in response to minor structural perturbations, they are anchoring to superficial lexical cues rather than modelling the semantics. Consequently, the fine-tuned systems heavily exploit these function words as categorical triggers. One might assume that scaling up model parameters naturally mitigates shortcut learning. However, as demonstrated by Kostić et al. (2026), the relationship between model size and robustness is non-monotonic and task-dependent; increased scale does not universally confer stability against lexical variations.

Takeaway: High accuracy in implicit discourse prediction could be unreliable across diverse genres. Fine-tuned models fail to develop robust compositional processing of argument semantics. Instead, they use function words as prediction shortcuts, a structural vulnerability that shows in both the news type (PDTB) and the multi-domain eng.pdtb.gum datasets. Consequently, standard accuracy (Acc_{ft}) alone cannot evaluate the robustness of the models' semantic processing. We suggest establishing causal robustness metrics, specifically Reconstruction Accuracy ($RAcc$) and Mean Surprisal Drop (MSD), as complementary standards to expose superficial heuristic reliance in future discourse benchmarks.

6 Conclusions

We present a complementary diagnostic framework to evaluate fine-tuned LLMs for discourse relation prediction, showing a disconnect between standard classification accuracy and robust semantic integration. In explicit settings, connectives function as absolute triggers rather than interpretative clues. This reliance is also observed in implicit discourse settings, where function words significantly affect the model's predictions. Fine-tuned LLMs are highly sensitive to function words. Even if one function word is replaced with another, while the discourse relation remains unchanged, the model's prediction is highly liable to shift. These empirical findings suggest that researchers cannot be satisfied with high classification scores alone. Future work should shift focus toward robust and semantically grounded models, moving beyond the pursuit of raw accuracy. Meanwhile, incorporating additional linguistic metrics to assess model performance will enhance the comprehensiveness of the evaluation,

helping detect the right-for-wrongs phenomenon and thus reducing concerns about applying the models to various downstream tasks.

Limitations

While we propose our diagnostic framework as a complement to traditional metrics, it is not without limitations. First, our empirical evaluation is currently restricted to English corpora. Because the distribution and syntactic reliance on function words vary significantly across typologies, future work can extend this audit to morphologically richer or low-resource languages to verify the cross-lingual universality of these shortcuts. Second, measuring token-level probability drops inherently requires access to the model's output logits. Consequently, this diagnostic framework is currently applicable only to open source models or APIs that expose full logit distributions. It cannot be applied to closed-source commercial models. Last, our analysis is grounded in the PDTB theoretical framework. Integrating alternative formalisms, such as Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) or Segmented Discourse Representation Theory (SDRT) (Lascarides and Asher, 2007), alongside broader psycholinguistic theories, will be crucial for developing a more comprehensive evaluation of hierarchical discourse comprehension in LLMs.

Acknowledgements

The authors would like to thank the anonymous reviewers for their comments. This work has been funded by the Klaus Tschira Foundation, Heidelberg, Germany.

References

Aryan Agrawal, Lisa Alazraki, Shahin Honarvar, Thomas Mensink, and Marek Rei. 2025. Enhancing LLM robustness to perturbed instructions: An empirical study. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

Saeed S Alahmari, Lawrence Hall, Peter R Mouton, and Dmitry Goldgof. 2025. Large language models robustness against perturbation. *Scientific Reports*.

Maryam Amirizaniani, Elias Martin, Tanya Roosta, Aman Chadha, and Chirag Shah. 2024. Auditllm: A tool for auditing large language models using multiprobe approach. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 5174–5179, New York, NY, USA. Association for Computing Machinery.

B. Beizer and J. Wiley. 1996. Black box testing: Techniques for functional testing of software and systems. *IEEE Software*, 13(5):98–.

Chloé Braud, Amir Zeldes, Chuyuan Li, Yang Janet Liu, and Philippe Muller. 2025. The DISRPT 2025 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In *Proceedings of the 4th Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2025)*, pages 1–20, Suzhou, China. Association for Computational Linguistics.

Clea Chataigner, Rebecca Ma, Prakhar Ganesh, Yuhao Chen, Afaf Taik, Elliot Creager, and Golnoosh Farnadi. 2026. Say it another way: Auditing LLMs with a user-grounded automated paraphrasing framework. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1441–1467, Rabat, Morocco. Association for Computational Linguistics.

Hanjie Chen, Guangtao Zheng, and Yangfeng Ji. 2020. Generating hierarchical explanations on text classification via feature interaction detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5578–5593, Online. Association for Computational Linguistics.

Noam Chomsky. 2001. Derivation by phase.

Stephane Collot, Colin Fraser, Justin Zhao, William F. Shen, Timon Willi, and Ilias Leontiadis. 2026. Balanced accuracy: The right metric for evaluating LLM judges - explained through youden's J statistic. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 927–936, Rabat, Morocco. Association for Computational Linguistics.

Yi Fan, Michael Strube, and Wei Liu. 2026. Linguistic cues for LLM-based implicit discourse relation classification. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 4585–4602, Rabat, Morocco. Association for Computational Linguistics.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Eliya Habba, Ofir Arviv, Itay Itzhak, Yotam Perlitz, Elron Bandel, Leshem Choshen, Michal Shmueli-Scheuer, and Gabriel Stanovsky. 2025. DOVE: A large-scale multi-dimensional predictions dataset towards meaningful LLM evaluation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11744–11763, Vienna, Austria. Association for Computational Linguistics.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Najoung Kim, Song Feng, Chulaka Gunasekara, and Luis Lastras. 2020. Implicit discourse relation classification: We need to talk about evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5404–5414, Online. Association for Computational Linguistics.

Bogdan Kostić, Conor Fallon, Julian Risch, and Alexander Löser. 2026. Same meaning, different scores: Lexical and syntactic sensitivity in llm evaluation. *Preprint*, arXiv:2602.17316.

Alex Lascarides and Nicholas Asher. 2007. Segmented discourse representation theory: Dynamic semantics with discourse structure. In *Computing meaning*, pages 87–124. Springer.

Roger Levy. 2008. A noisy-channel model of human sentence comprehension under uncertain input. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 234–243, Honolulu, Hawaii. Association for Computational Linguistics.

Dongqi Liu, Xi Yu, Vera Demberg, and Mirella Lapata. 2025. Explanatory summarization with discourse-driven planning. *Transactions of the Association for Computational Linguistics*, 13:1146–1170.

Wei Liu, Yi Fan, and Michael Strube. 2023a. HITS at DISRPT 2023: Discourse segmentation, connective detection, and relation classification. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 43–49, Toronto, Canada. The Association for Computational Linguistics.

Wei Liu, Xiyan Fu, and Michael Strube. 2023b. 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)*, pages 7792–7808, Toronto, Canada. Association for Computational Linguistics.

Wei Liu and Michael Strube. 2023. Annotation-inspired implicit discourse relation classification with auxiliary discourse connective generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15696–15712, Toronto, Canada. Association for Computational Linguistics.

Wei Liu and Michael Strube. 2025a. Discourse relation-enhanced neural coherence modeling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4748–4762, Vienna, Austria. Association for Computational Linguistics.

Wei Liu and Michael Strube. 2025b. Joint modeling of entities and discourse relations for coherence assessment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 21910–21926, Suzhou, China. Association for Computational Linguistics.

Wei Liu, Stephen Wan, and Michael Strube. 2024a. What causes the failure of explicit to implicit discourse relation recognition? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2738–2753, Mexico City, Mexico. Association for Computational Linguistics.

Yang Janet Liu, Tatsuya Aoyama, Wesley Scivetti, Yilun Zhu, Shabnam Behzad, Lauren Elizabeth Levine, Jessica Lin, Devika Tiwari, and Amir Zeldes. 2024b. GDTB: Genre diverse data for English shallow discourse parsing across modalities, text types, and domains. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12287–12303, Miami, Florida, USA. Association for Computational Linguistics.

Qian Lou, Xin Liang, Jiaqi Xue, Yancheng Zhang, Rui Xie, and Mengxin Zheng. 2024. CR-UTP: Certified robustness against universal text perturbations on large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9863–9875, Bangkok, Thailand. Association for Computational Linguistics.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Neural Information Processing Systems*.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Todor Mihaylov and Anette Frank. 2019. Discourse-aware semantic self-attention for narrative reading comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2541–2552, Hong Kong, China. Association for Computational Linguistics.

Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. 2023. Auditing large language models: a three-layered approach. *AI and Ethics*, 4(4):1085–1115.

Barbara Partee. 1995. Lexical semantics and compositionality. *An invitation to cognitive science: Language*, 1:311–360.

Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 13–16, Suntec, Singapore. Association for Computational Linguistics.

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

Giulia Pucci and Leonardo Ranaldi. 2025. Advancing oversight reasoning across languages for audit sycophantic behaviour via X-agent. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12949–12965, Suzhou, China. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Takuma Sato, Ai Kubota, and Koji Mineshima. 2025. Is partial linguistic information sufficient for discourse connective disambiguation? a case study of concession. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 977–990, Vienna, Austria. Association for Computational Linguistics.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423.

Noor van Leusen. 2004. Incompatibility in context: A diagnosis of correction. *Journal of Semantics*, 21(4):415–415.

Haoyu Wang, Guozheng Ma, Cong Yu, Ning Gui, Linrui Zhang, Zhiqi Huang, Suwei Ma, Yongzhe Chang, Sen Zhang, Li Shen, Xueqian Wang, Peilin Zhao, and Dacheng Tao. 2025a. Are large language models really robust to word-level perturbations? *Transactions on Machine Learning Research*.

Zhipang Wang, Yu Hong, Weihao Sun, and Guodong Zhou. 2025b. Using subtext to enhance generative IDRR. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 464–473, Vienna, Austria. Association for Computational Linguistics.

Bonnie Webber, Rashmi Prasad, and Alan Lee. 2019a. Ambiguity in explicit discourse connectives. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 134–141, Gothenburg, Sweden. Association for Computational Linguistics.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019b. The Penn Discourse Treebank 3.0 Annotation Manual. *Philadelphia, University of Pennsylvania*, 35:108.

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, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Yu Yuan, Lili Zhao, Kai Zhang, Guangting Zheng, and Qi Liu. 2024. Do LLMs overcome shortcut learning? an evaluation of shortcut challenges in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12188–12200, Miami, Florida, USA. Association for Computational Linguistics.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Yuqing Zhou, Ruixiang Tang, Ziyu Yao, and Ziwei Zhu. 2024. Navigating the shortcut maze: A comprehensive analysis of shortcut learning in text classification by language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2586–2614, Miami, Florida, USA. Association for Computational Linguistics.

A More Results

The explicit and implicit results on PDTB 2.0 are shown in Table 6 and Table 7, respectively.

B Computational Resource

In this work, all our experiments are conducted on an NVIDIA H100 GPU.

C Details About Datasets

The dataset statistics are shown in Table 5. The top-level labels in PDTB 2 and PDTB 3 are *Temporal, Contingency, Comparison,* and *Expansion*. The second-level labels in PDTB 2 are *Concession, Contrast, Cause, Pragmatic cause, Conjunction, Instantiation, Alternative, List, Restatement, Asynchronous* and *Synchrony*. The second-level labels in PDTB 3 are *Concession, Contrast, Cause, Cause+Belief, Condition, Purpose, Conjunction, Equivalence, Instantiation, Level-of-detail, Manner, Substitution, Asynchronous* and *Synchronous*. The labels in eng.pdtb.gum are *Explanation, Query, Causal, Condition, Purpose, Alternation, Temporal, Mode, Elaboration, Conjunction, Concession, Contrast* and *Reformulation*.

Dataset	Train	Val	Test
PDTB 2 (Prasad et al., 2008)	12632	1183	1046
PDTB 3 (Webber et al., 2019b)	17085	1653	1474
eng.pdtb.gum (Liu et al., 2024b)	10519	1682	1678

Table 5: The data distribution for the datasets we use in this work.

D Prompt Design

The prompts we use for explicit and implicit cases are shown in Figure 1 and Figure 2, respectively.

E Targeted Function Words for Structural Perturbations

To audit the models' reliance on superficial lexical shortcuts during implicit discourse relation prediction, we systematically mask specific categories of semantically light function words. These words were selected because they serve primarily grammatical roles within a sentence and lack the core propositional content required to logically infer the relations. More importantly, after the pre-training, LLMs are expected to recover those words easily.

The targeted function words belong to the following four lexical categories:

- Articles (such as: a, an, the)
- Prepositions (such as: to, for, with, on, in, at, from, by)
- Auxiliary Verbs (such as: be, have, do, can, will, shall, should, would, may, might)
- Determiners (such as: this, that, these, those, some, any, each)

By strictly constraining our structural perturbations to these predefined categories, we ensure that the masking process disrupts only the surface-level syntax. Crucially, this guarantees that the fundamental argument semantics necessary for a robust parser to infer the correct discourse relation remain functionally sufficient.

F Mathematical Motivation for Contextual Normalisation

In Section 4.2, the formulation of the Normalised Connective Focus (NF) incorporates a division by two to calculate the mean contextual contribution $\phi_{ctx} = (\phi_{A_1} + \phi_{A_2})/2$. This normalisation is a critical design choice to maintain dimensional alignment between the connective and the argument context.

To illustrate the necessity of this normalisation, consider a hypothetical "perfectly balanced" semantic parser. This ideal model relies equally on all functional units within the discourse structure, assigning an identical impact of 1/3 to each component: $\phi_C = 1/3$, $\phi_{A_1} = 1/3$, and $\phi_{A_2} = 1/3$. Intuitively, since the model relies equally on the explicit marker and the context, the focus ratio should reflect a 50% balance.

If the contextual contribution were calculated as a simple sum ($\phi_{ctx} = \phi_{A_1} + \phi_{A_2} = 2/3$), the resulting ratio would be:

$$NF_{unnormalized} = \frac{1/3}{1/3 + 2/3} \approx 33.3\%$$

This unnormalized approach mathematically penalises the context simply because it has a bipartite structure (two arguments), shifting the baseline so that a 50% score would actually indicate a strong bias towards the connective.

By applying the mean normalisation, we compress the two arguments into a single standard semantic unit:

$$NF_{normalized} = \frac{1/3}{1/3 + 1/3} = 50.0\%$$

Instruction: Analyze Arg1, Connective and Arg2, then predict the discourse relation
Input:
Arg1: {Argument 1 text}
Connective: {Connective}
Arg2: {Argument 2 text}
Relation:

Figure 1: The prompt for explicit cases.

Instruction: Analyze Arg1 and Arg2, then predict the discourse relation
Input:
Arg1: {Argument 1 text}
Arg2: {Argument 2 text}
Relation:

Figure 2: The prompt for implicit cases.

This operation aligns the denominator with the connective's single-token nature, ensuring that the 50% threshold objectively reflects equal reliance on explicit cues and underlying semantics. It is important to note that this 50% normalised threshold is not proposed as a psycholinguistic law dictating that context and connectives must be equally weighted in human cognition. Rather, it serves as a mathematically calibrated diagnostic baseline. While explicit connectives are undeniably strong discourse signals, a robust parser must still integrate contextual semantics to ensure logical consistency. By establishing a symmetric 50% reference point, our normalisation provides an objective mathematical scale for quantifying extreme deviations—specifically, the heuristic exploitation in which a model ignores context entirely.

G Semantic Incompatibility in Adversarial Substitution

Consider the following adversarial intervention based on a causal relation:

(1) [The streets are completely flooded.]$_{\text{Arg1}}$ **Explicit=Because** [It rained heavily throughout the night.]$_{\text{Arg2}}$

(2) [The streets are completely flooded.]$_{\text{Arg1}}$ **Explicit=But** [It rained heavily throughout the night.]$_{\text{Arg2}}$

In the adversarial case, the explicit connective "but" forces a *Contrast* relation. However, the underlying physical events at Arg1 and Arg2 are inherently strongly causally linked. A robust semantic parser should recognise that heavy rain naturally leads to flooded streets. When presented with the adversarial sentence, the system should encounter a structural clash between the surface cue ("but") and the physical reality of the arguments, leading to hesitation in prediction, high entropy, or a drop in confidence. Instead, our empirical results show that the fine-tuned LLMs process the adversarial sentence with near-100% sensitivity, unhesitatingly predicting *Contrast* with high confidence. This behaviour shows that the models do not attempt to relate the connective with its context; they blindly execute the connective's categorical trigger, completely discarding the semantic absurdity of the resulting sentence.

H Theoretical Justification for Causal Faithfulness

In Section 5.3, we define the Fidelity Correlation (FC) as an intervention-based metric. It is important to clarify the distinction between observational attribution and causal intervention in this context. Traditional token-level impact, whether derived from gradients or attention weights, is fundamentally observational. They highlight which input features strongly activate the network, but do not guarantee that these features causally drive the final prediction. As established by Jacovi and Goldberg (2020), an interpretation is faithful only if it accurately reflects the system's true decision process under causal perturbations.

To bridge the gap between observation and causation, our FC metric incorporates a direct causal

	Top Level					Second Level				
Models	Acc_{ft}	Acc_{zs}	Acc_{rv}	NF	SS	Acc_{ft}	Acc_{zs}	Acc_{rv}	NF	SS
Llama-3-8b	94.55	35.18	55.8	58.57	71.28	85.98	30.88	37.65	77.09	98.91
Qwen3-4B	95.41	22.88	50.19	79.28	70.66	85.48	31.72	35.81	80.89	98.00
Qwen3-8B	95.41	22.49	49.65	75.57	70.74	84.56	35.89	38.31	79.17	98.83

Table 6: Results on PDTB 2.0. Acc_{ft}: Fine-tuned Accuracy, Acc_{zs}: Zero-shot Accuracy, Acc_{rv}: fine-tuned models' Accuracy after masking the connective, NF: normalised focus, SS: Signal Sensitivity.

	Top Level					Second Level				
Models	Acc_{ft}	Acc_{zs}	RAcc	FC*	MSD*	Acc_{ft}	Acc_{zs}	RAcc	FC*	MSD*
Llama-3.1-8B	71.70	13.86	59.75	0.42	1.27	53.71	21.66	33.69	0.39	1.95
Qwen3-4B	68.74	7.46	59.37	0.36	1.61	58.83	19.15	33.69	0.38	2.74
Qwen3-8B	70.27	6.88	59.75	0.39	1.47	57.36	21.56	38.11	0.40	2.06

Table 7: Results on PDTB 2.0 dataset. Acc_{ft}: fine-tuned accuracy, Acc_{zs}: zero-shot accuracy, RAcc: reconstructed accuracy, FC: fidelity correlation, MSD: Mean Surprisal Drop. Metrics with * indicate that the values are not percentages.

intervention. By systematically masking the entire syntactic phrase from the input, we perform a black-box causal intervention similar to the behavioural testing principles outlined by Ribeiro et al. (2020). The resulting drop in prediction probability (Causal Impact) represents the absolute causal effect of that specific phrase on the output. By measuring the Pearson correlation between the sum of observational token scores and the actual causal impact, the FC metric assesses whether the token-level explanations align with the model's true causal dynamics.

I Causal Mechanics of the Mean Surprisal Drop

The Mean Surprisal Drop (MSD) introduced in Section 5.4 serves as a causal diagnostic tool rather than a standard performance metric. Readers might question why masking function words serves as empirical evidence of shortcut learning. The theoretical foundation lies in the concept of counterfactual intervention used to diagnose linguistic heuristics (McCoy et al., 2019; Geirhos et al., 2020).

In standard evaluation, high accuracy indicates that the input maps to the correct output. However, it fails to reveal which variables control the prediction. The MSD framework applies a targeted structural intervention by masking semantically light function words while preserving the core argument semantics. This intervention mathematically isolates the causal influence of surface placeholders on the system's internal probability distribution. If the fine-tuned system possessed robust compositional semantics, this minor structural intervention would exert a negligible causal effect on the output certainty. However, the severe spike in surprisal (entropy) following this specific ablation mathematically proves a strong causal dependency on these exact functional tokens. This intervention directly aligns with the shortcut learning mechanisms described in prior literature, demonstrating that the system's confidence is causally anchored to superficial cues rather than to the integrated logic of the argument.

J The Impact of Connective Ambiguity on Downstream Tasks

In Section 4.5, we highlight that fine-tuned systems often treat connectives as absolute categorical triggers. A critical danger of this shortcut learning is that explicit connectives in natural language are inherently ambiguous (Webber et al., 2019a). When a system blindly maps a specific connective to a high-frequency relation label without integrating argument semantics, it leads to cascading failures in downstream applications such as Reading Comprehension (Mihaylov and Frank, 2019) and Text Summarisation (Liu et al., 2025).

To illustrate this vulnerability, consider the connective "since", which frequently exhibits sense ambiguity between Temporal (time-based) and Contingency (causal) relations.

(3) **Case A (Temporal): Explicit=Since** [the manager left the company,]$_{\text{Arg1}}$ [the team has been using the new software.]$_{\text{Arg2}}$

(4) **Case B (Causal): Explicit=Since** [the man-

ager left the company,]$_{\text{Arg1}}$ [the team has to restructure its workflow.]$_{\text{Arg2}}$

In Case A, "since" functions purely as a temporal marker indicating when the action started. In Case B, "since" functions as a causal marker indicating the reason for the restructuring. A robust parser must process the compositional semantics of the arguments to disambiguate the relation.

However, if a fine-tuned discourse parser relies on a superficial shortcut (e.g., statistically anchoring "since" strictly to a Temporal label due to training data bias), it will misclassify Case B. This single classification error severely corrupts downstream tasks.

By treating ambiguous connectives as programmatic triggers rather than context-dependent clues, shortcut-biased models generate fundamentally unrobust discourse representations. This demonstrates why evaluating models using adversarial and implicit structural metrics is essential before deploying them in complex downstream linguistic tasks.

K The Fragility of Function Word Shortcuts in Downstream Applications

In Section 5, our diagnostic audit reveals that fine-tuned models rely heavily on semantically light function words (e.g., prepositions) to predict implicit discourse relations. This severe dependency introduces critical vulnerabilities when these discourse parsers are deployed in real-world downstream applications, particularly those that evaluate human- or machine-generated text.

An example is Automated Essay Scoring (AES). Researchers use discourse relations to measure the coherence and quality of argumentation in student essays (Liu and Strube, 2025a,b). If a coherence scoring system relies on a not robust discourse parser biased toward function words, the evaluation becomes highly unreliable. A student might construct a good, logical and coherent argument. However, a minor typographical error, such as a misspelt preposition or a synonymous preposition (e.g., "on" instead of "in"), will disrupt the parser's expected surface cues. The parser will consequently fail to recognise the implicit semantic relation, leading to an artificially low coherence score. This shortcut penalises stylistic variations and minor grammatical slips rather than assessing actual logical argumentation.

L Unified Diagnostic Framework and Metric Mapping

To systematically audit shortcut learning across both explicit and implicit discourse relation prediction, our diagnostic framework relies on a unified four-step operational pipeline. Although the specific metric instantiations differ depending on the presence (explicit) or absence (implicit) of connectives, they share the same underlying methodological purposes. Table 8 maps these unified methods to their corresponding experimental settings.

This framework operates sequentially, with clear dependencies between each methodological step to ensure causal validity:

- **Method 1 (Attribution Measurement):** This initial stage serves as the exploratory foundation. It identifies potential Shortcut Triggers (STs) — such as explicit connectives or specific semantically light function words — by evaluating which input components predominantly drive the model's predictions.
- **Method 2 (Target Ablation):** Building upon the findings of Method 1, this step conducts targeted ablations on the identified STs. It evaluates the model's structural robustness by measuring how much accuracy degrades when these critical triggers are disrupted.
- **Method 3 (Confidence Dependency):** Utilising the exact ablated inputs from Method 2, this step shifts the evaluation from hard accuracy to soft predictive probabilities. It mathematically quantifies the severity of the model's confidence loss, revealing exponential probability drops even when the categorical label remains unchanged.
- **Method 4 (Intervention & Adversarial Stress):** The final stage acts as the causal validator. By applying structural interventions and adversarial stress to the STs identified in Method 1, it confirms that the models explicitly anchor to these structural heuristics. This step guarantees that the performance degradations observed in Methods 2 and 3 stem from flawed reasoning mechanics rather than simple observational noise.

Method	Explicit Setting	Implicit Setting	Purpose
1. Attribution Measurement	Leave-One-Out (LOO) perturbation	Semantic Attribution	Using masking tests to assess which part of the input the model heavily relies on to infer the relation.
2. Target Ablation	Robustness Test	Reconstruction Accuracy	Measuring performance degradation when the identified shortcut trigger is masked or replaced.
3. Confidence Dependency	Normalised Connective Focus	Mean Surprisal Drop (MSD)	Quantifying the loss of predictive confidence (via probability or entropy shifts) mathematically when the shortcut trigger is removed.
4. Intervention & Adversarial Stress	Signal Sensitivity	Fidelity Correlation	Executing structural interventions or adversarial substitutions to prove that the reliance on structural heuristics is causal rather than observational noise.

Table 8: The mapping of unified diagnostic operations across explicit and implicit discourse evaluation settings.