Detecting Potentially Under-annotated Explicit Discourse Connectives in the Penn Discourse Treebank (PDTB-3) with LLMs

Yueh Ting Chuang, Xixian Liao, Bonnie Webber
School of Philosophy, Psychology and Language Science, University of Edinburgh
Barcelona Supercomputing Center
School of Informatics, University of Edinburgh
chelseachuang513@gmail.com, xixianliao@gmail.com, bonnie.webber@ed.ac.uk

Abstract

Accurate identification of explicit discourse connectives is crucial for analysing discourse relations, which supports NLP tasks such as summarisation and question answering. However, annotation inconsistencies remain a challenge, particularly for ambiguous prepositions with both discourse and non-discourse usages. This paper presents a pipeline that leverages large language model (LLM) prompting, cross-model agreement, and syntactic pattern analysis to detect likely under-annotated connectives. Evaluated on four prepositions (*by*, *with*, *without*, and *for*), the approach effectively identifies likely under-annotations for some, but not all prepositions. Results show that while the method is promising, its generalisability depends on improved prompt design, model choice, and syntactic analysis tools. The findings highlight both the potential and limitations of LLM-based approaches for corpus error detection and demonstrate how improved discourse annotation can contribute to more reliable data for downstream NLP tasks.

Keywords: Discourse Connectives, Annotation Error Detection, Prompting

1. Introduction

The Penn Discourse Treebank (PDTB) (Prasad et al., 2008, 2019) is the largest English corpus annotated with discourse relations, in which relations are marked between text spans linked by discourse connectives (Webber et al., 2019; Pitler and Nenkova, 2009). Identifying these relations is important for downstream NLP tasks such as summarisation, question answering, and text generation (Scholman et al., 2021; Atwell et al., 2021). A critical step in discourse annotation and parsing is correctly detecting explicit discourse connectives—lexical items that directly signal relations between clauses or sentences. For example, in *"He was fined for he had violated the safety regulations,"* the preposition *for* functions as a discourse connective signaling a causal relation.

Most previous discourse parsing research has relied on PDTB-2 for model training and evaluation (Pitler and Nenkova, 2009; Lin et al., 2014). Despite PDTB-3's expanded connective coverage and refined annotation guidelines (Webber et al., 2019), few studies have re-examined its explicit connective annotations. Recent work has instead focused on implicit discourse relations, which are more challenging to detect. As a result, the quality and consistency of explicit discourse connective annotations in PDTB-3 remain under-explored, particularly for prepositions that had not been annotated in the PDTB-2 as connectives.

Meanwhile, large language models (LLMs) have grown rapidly in scale and capability, thanks to advances in training data, algorithms, and hardware. Prompting has emerged as a popular way to guide LLMs in performing NLP tasks without requiring any weight updates, by providing carefully designed instructions and examples directly in the prompt (Dong et al., 2024). This capability makes them promising tools for assisting in annotation error detection in corpus.

In this study, we propose an LLM-based pipeline to identify potentially under-annotated explicit discourse connectives in PDTB-3, focusing on four prepositions—*by*, *with*, *without*, and *for*—which are especially prone to ambiguity between discourse and non-discourse usage. The pipeline combines prompt-based LLM classification, cross-model agreement analysis, and syntactic pattern analysis using Penn Treebank (PTB) features to narrow down high-confidence candidate instances for annotation review.

This work aims to evaluate whether LLMs can effectively detect under-annotated explicit discourse connectives and assist in improving corpus consistency. By providing a semi-automatic approach to highlight likely annotation gaps, the proposed method offers a practical step toward improving annotation reliability and supporting future parser development.

2. Background and Related Work

2.1. Discourse Connectives

Discourse connectives signal relationships between textual units such as clauses or sentences, contributing to discourse coherence and interpretation (Rouchota, 1996). These expressions can be explicit, as in *because*, *however*, or *for example*, or implicit, where the relation must be inferred from

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 2012–2023
11–16 May 2026.

context (Pitler et al., 2009). This study focuses on explicit connectives, particularly prepositions such as *by*, *with*, *without*, and *for*, whose syntactic and semantic flexibility makes them prone to annotation inconsistency.

A key challenge lies in distinguishing discourse from non-discourse usage. As noted by Pitler and Nenkova (2009), a common ambiguity occurs when a connective word can be used in both discourse and non-discourse contexts. For example, in the discourse usage shown in Table 1, *by* functions as a connective indicating a causal relationship between two clauses. However, in the non-discourse usage shown in Table 1, *by* does not link two discourse segments but instead operates within a clause to convey temporal information.

Pitler and Nenkova (2009) demonstrated that such ambiguity can be partially resolved using syntactic cues from the Penn Treebank (PTB) (Marcus et al., 1993). In particular, the parent and right sibling categories of the connective node are highly informative: a connective is more likely to be used in a discourse function when its right sibling in the parse tree is a clausal structure.

Usage	**Example**
Discourse usage	*By* pressuring taxi and bus drivers who needed licenses, he gained a ready cache of information.
Non-discourse usage	*By* 1997, almost all remaining uses of cancer-causing asbestos will be outlawed.

Table 1: Examples of discourse and non-discourse usage of the explicit discourse connective *by*.

2.2. The Penn Discourse Treebank

The Penn Discourse Treebank (PDTB) (Prasad et al., 2008, 2019) provides large-scale manual annotation of discourse relations over Wall Street Journal articles. The 2.0 version (PDTB-2) (Prasad et al., 2008), released in 2008, contains 18,459 instances of explicitly annotated discourse connectives (Pitler and Nenkova, 2009; Prasad et al., 2017). Building upon PDTB-2, the 2019 release of PDTB-3 (Prasad et al., 2019) expanded coverage to 24,240 explicit connectives and introduced new connective types, finer-grained sense labels, and improved annotation consistency (Webber et al., 2019).

A key addition in PDTB-3 is the inclusion of subordinators to improve the annotation of intra-sentential discourse relations—specifically, when the connective is fully contained within the span of a top-level S-node in the Penn Treebank (Webber et al., 2019). As a result, prepositions such as *despite*, *for*, *by* or *in* are only considered subordinators when they introduce a clause. This introduces challenges for annotation accuracy due to the discourse and non-discourse ambiguity of such prepositions, as illustrated in the example in Table 1.

2.3. Detecting Annotation Errors in NLP Corpora

Human-annotated corpora serve as training and evaluation benchmarks across NLP, yet they inevitably contain label noise and omissions due to annotator subjectivity, linguistic ambiguity, and task complexity (Plank, 2022).

A growing body of research has proposed automated methods to detect such errors. Swayamdipta et al. (2020) proposed a model-based diagnostic tool, *Data Maps*, which helps identify mislabeled or ambiguous examples that are difficult for models to learn effectively. Northcutt et al. (2021) introduced *Confidence Learning*, a framework for identifying label errors in classification datasets by estimating the joint distribution of noisy observed labels and latent true labels. A broader comparison by Klie et al. (2023) across various NLP tasks showed that these methods remain sensitive to dataset properties and are less reliable for tasks requiring deeper syntactic or semantic reasoning.

These studies target sentence-level classification or sequence labeling. In contrast, discourse annotation involves more complex interactions between syntax and semantics, making manual review costly and time-consuming. This motivates exploring methods that can assist annotators in systematically detecting possible under-annotation.

2.4. Prompting with Large Language Models (LLMs)

Recent advances in large language models (LLMs) have opened new possibilities for semi-automatic annotation error detection. Prompting allows models to perform classification or reasoning tasks using natural language instructions and examples, without fine-tuning (Brown et al., 2020; Dong et al., 2024). Different prompting strategies have been proposed to improve reliability and interpretability. Wei et al. (2022) introduce *Chain-of-Thought (CoT)* prompting which encourages models to generate step-by-step reasoning, while Rubin et al. (2022) trained a dense retriever to identify and rank candidate demonstrations based on their likelihood of improving output performance.

LLMs' capacity to reason about syntax and semantics makes them promising tools for identifying annotation inconsistencies. However, they are also sensitive to prompt phrasing and example choice

(Mizrahi et al., 2024). Furthermore, they may produce hallucinated or inconsistent outputs, including "faithfulness hallucination" where the answer contradicts the input and "factuality hallucination" where the answer is not able to be verified from reliable sources. Despite these limitations, LLMs offer an efficient and interpretable way to highlight potentially inconsistent annotations that can later be verified by human experts.

2.5. Research Motivation

Building on these strands of work, this study investigates whether LLM prompting can support error detection in discourse-annotated corpora, focusing on a subset of ambiguous prepositions in PDTB-3. We propose a pipeline that combines prompt-based classification, model agreement analysis, and syntactic pattern extraction from PTB parses to identify potential under-annotated explicit discourse connectives. By automating part of the error discovery process, this approach aims to improve annotation consistency and, consequently, enhance the reliability of discourse resources used in downstream NLP tasks.

3. Methodology

3.1. Experimental Pipeline

This study examines four prepositions—*by*, *with*, *without*, and *for*—which were newly annotated as discourse connectives in PDTB-3. In the corpus, these prepositions were annotated only when aligned with PropBank argument structures (Webber et al., 2019), resulting in limited coverage. Using GPT-4o and LLaMA-3-70B, we apply a prompting-based pipeline to identify potential under-annotated instances, combining cross-model agreement and syntactic pattern analysis to enhance reliability.

Each preposition is evaluated separately because their frequency, syntactic roles, and likelihood of signalling discourse relations differ substantially (Taboada, 2006). In addition, prompt effectiveness can vary with minor phrasing differences (Mizrahi et al., 2024), making per-preposition optimisation necessary.

The experimental pipeline (Figure 1) proceeds in multiple stages. First, a prompt evaluation is conducted for each preposition using 40 manually selected PDTB-3 examples (20 annotated, 20 unannotated) to compare prompt templates and identify the best configuration for each model.

Next, the better-performing model–prompt combination is applied to all corpus instances of the target preposition to generate predictions. Cases where model output disagrees with the gold annotation are then re-evaluated by the second model using its own optimal prompt to assess model agreement. High-agreement instances are treated as stronger candidates for potential under-annotation.

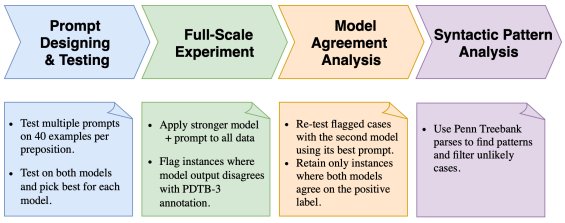

Figure 1: Pipeline for identifying potential under-annotations in PDTB-3.

Preposition	Annotated Instances	Total Instances	Annotated (%)
by	435	5,626	7.73
with	297	5,327	5.58
without	94	348	27.01
for	60	10,097	0.59

Table 2: Annotated instances of four prepositions in the PDTB-3 corpus.

Finally, syntactic pattern analysis is applied to the full set of extracted instances. By examining features such as the parent and right sibling categories from the PTB parse trees, we investigate whether certain syntactic configurations are systematically associated with discourse use. These structural cues serve to validate model predictions and filter out unlikely candidates.

3.2. Data Preparation

The selection of the four prepositions (*by*, *with*, *without*, and *for*) is based on the number of annotated instances in PDTB-3, as shown in Table 2. These four are prepositions with a relatively higher number of annotations, as they are more likely to function as discourse connectives in the PTB texts. Additionally, prompting benefits from having multiple positive examples, making prepositions with more annotated cases more suitable for prompt-based experimentation.

The Penn Treebank (PTB) comprises 25 sections (Section 0 to Section 24), each containing multiple text files. For each target preposition, we extract all sentences from these files that contain either annotated or unannotated instances of that preposition. In the extracted sentences, the target preposition is marked with square brackets (e.g., *[with]*).

Sentences with multiple occurrences of the target preposition are split into separate instances, each containing only one bracketed occurrence. Prepositions that appear as part of multi-word discourse connectives (e.g., *for example*) are excluded, as

these are annotated differently in PDTB-3 and are not the focus of this study. Figure 2 shows examples of how the preprocessing is performed.

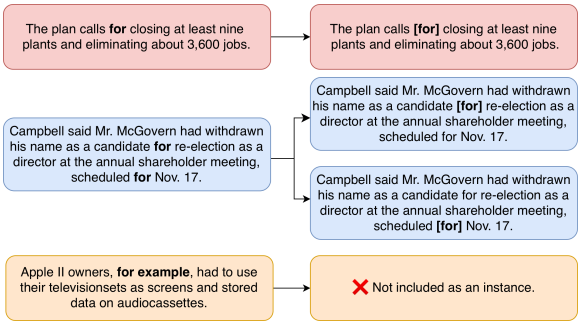

Figure 2: The figure illustrates the text preprocessing workflow, including marking, splitting and excluding of preposition instances.

3.3. Models

This study employs OpenAI's GPT-4o (OpenAI, 2023) and Meta's LLaMA-3-70B (Touvron et al., 2024) for prompt-based discourse connective classification. GPT-4o was accessed through the University of Edinburgh's Edinburgh Language Models (ELM) platform, which provides API access to OpenAI models under a fixed usage quota.[1] LLaMA-3-70B was accessed via the Groq API (Inc., 2024).

Both models were selected for their strong performance in few-shot and general NLP tasks without fine-tuning (Grattafiori et al., 2024; Shahriar et al., 2024). The 70B variant of LLaMA-3 was chosen for its balance between computational efficiency and representational capacity.

Prompts were issued as plain text with a temperature of 1.0, balancing determinism and response diversity. A consistent system prompt was used across all experiments: *"You are a linguist who always provides the correct answer as succinctly as possible. Answer precisely and avoid unnecessary elaboration."* This instruction encouraged concise and linguistically grounded responses while controlling for verbosity.

3.4. Evaluation Metrics

When detecting errors in a human-annotated corpus—especially for identifying discourse usages of prepositions, which more often serve non-discourse functions—the data is typically highly imbalanced, with far more correct than incorrect labels (Klie et al., 2023). To address this, we use precision, recall and F1 score as evaluation metrics to better assess model and prompt performance under such imbalance. A high recall indicates the model's ability to recover existing annotations, which we assume are mostly correct. On the other hand, low precision may suggest either that the model successfully detects under-annotated discourse connectives, or that it tends to over-predict positive cases. The F1 score provides a unified measure that balances precision and recall, making it particularly useful for evaluating performance on imbalanced data.

3.5. Prompt Design and Testing

We acknowledge that LLM performance is often sensitive to prompt formulation. Rather than conducting an extensive prompt search, we intentionally limited our design to three structured templates that vary in the amount and type of linguistic information provided. This controlled variation allows us to examine how additional explanatory and contextual cues influence model behaviour, while reducing the risk of overfitting results to highly specialised prompts.

We designed three prompt templates—*Basic Prompt*, *Explanation Prompt*, and *Contextualised Explanation Prompt*—shown in Figure 3. Each template includes a task description followed by k example sentences presented in random order. Due to computational constraints,[2] k was set to 4, balancing two positive and two negative examples to minimise token use. The final input to the model combines the selected template and the query sentence.

The *Basic Prompt* uses a standard few-shot format, pairing an instruction with a small set of examples. Few-shot prompting allows LLMs to infer task patterns from limited demonstrations (Brown et al., 2020; Dong et al., 2024).

The *Explanation Prompt* and *Contextualised Explanation Prompt* adapt the idea of chain-of-thought prompting (Wei et al., 2022), embedding short explanations within each example to show how a preposition functions as a discourse connective. For positive examples, we highlight the linked arguments and discourse sense; for negatives, we briefly explain why the instance is not annotated as a discourse connective. The task description also asks the model to provide a short justification for its prediction.

The *Contextualised Explanation Prompt* extends this by adding linguistic context from the PDTB-3 annotation guidelines (Webber et al., 2019), including the instruction: "*A preposition should only be*

[1] Unlike LLaMA-3-70B, which was accessed via the Groq API at no cost, GPT-4o usage through ELM is subject to a fixed allocation limit, influencing design decisions in this study.

[2] GPT-4o was accessed via the University of Edinburgh's Edinburgh Language Models (ELM) platform, which provides a fixed API quota.

annotated as an explicit discourse connective if it takes a clausal complement." A concise definition of clausal complement is included to ensure clarity.

Examples were manually selected to maximise syntactic and positional diversity. For positive examples, we also ensured coverage of different discourse senses, capturing the range of relations that connectives may express.

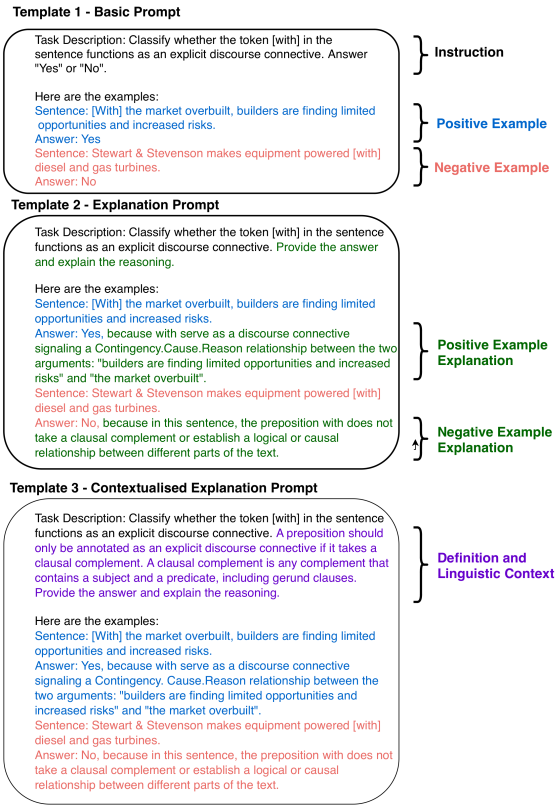

Figure 3: The three prompt templates used in this study.

All templates were tested in both zero-shot and few-shot conditions with GPT-4o and LLaMA-3-70B across four prepositions. Zero-shot tests assessed whether enhanced task descriptions alone could guide model behaviour, following the rationale of Reynolds and McDonell (2021) that well-formulated instructions may outperform few-shot examples for certain tasks.

Table 3 shows the best-performing configurations. GPT-4o generally outperformed LLaMA-3-70B, though optimal prompts varied across prepositions. All selected configurations used few-shot prompts, confirming that in-context examples consistently improve classification of discourse connectives.

While the *Contextualised Explanation Prompt* offers the richest linguistic cues, it was not always optimal—more complex prompts sometimes reduced precision. Performance for *for* was weakest across models, likely due to its low annotation rate (0.59%, see Table 2), which produces highly imbalanced training data and a less consistent contextual patterns (Reynolds and McDonell, 2021).

Prep.	Model	Best Prompt	P	R	F1
by	**GPT**	**Ctx. Expl.**	**0.87**	**1.00**	**0.93**
	LLaMA	Expl.	0.73	0.95	0.83
without	**GPT**	**Ctx. Expl.**	**0.72**	0.90	**0.80**
	LLaMA	Ctx. Expl.	0.59	**1.00**	0.74
with	**GPT**	**Expl.**	**0.76**	0.95	**0.84**
	LLaMA	Expl.	0.67	**1.00**	0.80
for	**GPT**	**Basic**	**0.67**	**0.80**	**0.73**
	LLaMA	Ctx. Expl.	0.58	0.75	0.65

Note. GPT = GPT-4o; LLaMA = LLaMA-3-70B; Expl. = Explanation; Ctx. Expl. = Contextualised Explanation; P = Precision; R = Recall.

Table 3: Best-performing prompt type and corresponding scores for each preposition and model. All prompts are few-shot.

LLaMA-3-70B exhibited more frequent overgeneralisation, generating higher false-positive rates. However, model agreement between GPT-4o and LLaMA-3-70B improved reliability: instances jointly identified as positive were typically strong candidates for true discourse usage.

3.6. Full-Scale Experiment

Based on the prompt evaluation results, we found that GPT-4o consistently achieved the best performance across all four prepositions. Consequently, we conducted full-scale experiments using GPT-4o, employing the best-performing prompt configuration for each preposition, as shown in Table 3.

We report precision, recall, and F1 score to evaluate the model's performance. Recall serves as the primary indicator of whether GPT-4o can reliably identify discourse usage, while precision and F1 score provide further insight into the quality and balance of the predictions. These metrics offer a foundational assessment of performance; we complement this quantitative evaluation with qualitative analyses of false positives, to better understand the model's predictions and the possibility of under-annotation in PDTB-3.

3.7. Model Agreement Analysis

Recent studies have proposed using LLMs as evaluation tools (Chang et al., 2024). One approach uses a separate LLM to evaluate the outputs of a primary model (Zheng et al., 2023). Another, like LLM-EVAL, proposes a unified prompt-based framework for multi-dimensional evaluation in open-domain conversations (Lin and Chen, 2023). Inspired by

this perspective, we adopt a second LLM as additional classifier to assess a subset of instances flagged as potential errors by the first model.

We compute the *consensus rate*, defined as the proportion of positive predictions from one model that are confirmed by the other. To further estimate precision, we also perform manual inspection on random samples of these consensus cases.

In preliminary analyses, we observed that both models frequently flagged overlapping instances as potential under-annotated connectives, suggesting that they capture similar high-confidence signals. At the same time, due to the inherent ambiguity of prepositions such as *by*, *with*, *without*, and *for*, the two models often diverged in their complete lists of potential errors. Such divergence appears to reflect task-level ambiguity rather than consistent bias in either model. To obtain a conservative and transparent filtering mechanism, we therefore prioritise instances on which both models agree, treating cross-model consensus as a high-confidence indicator for manual validation.

3.8. Syntactic Pattern Analysis

Previous research has shown that the parent category and right sibling category in a parse tree are informative for determining whether a word functions as a discourse connective (Pitler and Nenkova, 2009; Ibn Faiz and Mercer, 2013). In this study, we extract these two features for the prepositions from Penn Treebank (PTB) parse trees. They are later used to identify patterns associated with discourse usage and to filter out unlikely cases from the candidate set.

The syntactic pattern filtering stage is designed as a precision-enhancement layer rather than a core detection component. The primary identification of potential under-annotated connectives is performed by the LLM-based classification stage, which operates independently of syntactic parses. The use of PTB parse trees in this work is motivated by their close alignment with the PDTB annotation framework, facilitating direct structural comparison. However, the overall framework is not inherently tied to PTB constituency parses; adapting the filtering stage to alternative syntactic resources, such as dependency parses or other high-quality parsers, would be straightforward in principle.

4. Results and Discussion

4.1. Overall Performance

The results of the full-scale experiments are summarised in Table 4. Overall performance declines compared to the earlier small-scale experiment, as reflected by lower recall and F1 scores, which is expected given the larger task scale. GPT-4o, however, maintains high recall for three prepositions—*by*, *with*, and *without*—ranging from 0.88 to 0.92.

Prep.	Prec.	Rec.	F1	Count (Ann./Tot.)
by	0.70	0.92	0.80	435 / 5,626
with	0.33	0.91	0.49	297 / 5,327
without	0.60	0.88	0.71	94 / 348
for	0.02	0.58	0.04	60 / 10,097

Table 4: GPT-4o's performance in identifying discourse usage of four prepositions in PDTB-3. Low scores are highlighted in red. The last column shows the annotated and total counts in the corpus.

The results mirror earlier prompt evaluations: *for* performs weakest, while *by*, *with*, and *without* achieve reliable alignment with human annotations. F1 scores reveal that *with* (0.49) is limited by low precision despite high recall, whereas *for* (0.04) is weak in both metrics, suggesting frequent false positives for *with* and general difficulty for *for*.

Notably, the best-performing prompts differ: *Explanation Prompt* for *with* and *Basic Prompt* for *for* offer less contextual guidance than the *Contextualised Explanation Prompt*s used for *by* and *without*. This may contribute to higher false positive rates, a topic we explore further in the general discussion.

4.2. Model Agreement Analysis

To assess potentially under-annotated instances identified by GPT-4o, we perform a model agreement analysis using LLaMA-3-70B. Instances where GPT-4o predicts a preposition as a discourse connective despite being unannotated in PDTB-3 are re-evaluated with LLaMA-3-70B using its best-performing prompt (Table 3). Agreement between models is interpreted as a likely indication of under-annotation.

Figure 4 shows the proportion of agreement and disagreement for each preposition. Consensus rates range from 70.6% for *for* to 94.6% for *without*. The two prepositions with the highest consensus rates (*with* and *without*) use the same prompt format across both models. In contrast, *by* and *for*, which show lower agreement, are associated with different prompt formats in each model — suggesting that prompt alignment may influence cross-model consistency.

Manual inspection of sampled agreement cases confirms that many are indeed valid discourse connectives. For example, in "*He hurt himself further this summer [by] bringing homosexual issues into the debate*", both models correctly identify by as introducing a causal relation. However, the remaining errors are typically due to syntactic misinterpretation or limitations in prompt design. For instance,

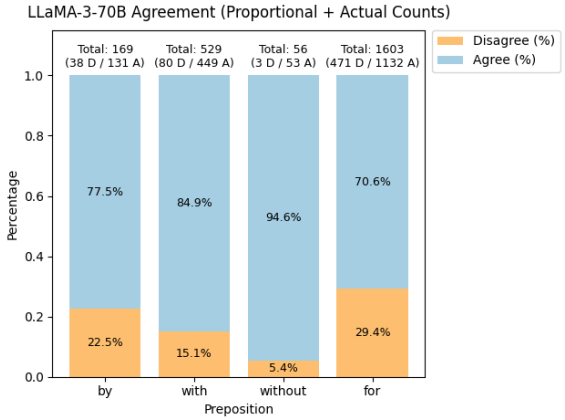

Figure 4: Proportion of agreement and disagreement between LLaMA-3-70B and GPT-4o on potentially under-annotated instances for each preposition. Each bar represents the percentage of LLaMA-3-70B predictions that agree or disagree with GPT-4o's classification. Numerical labels show the exact percentage breakdown within each bar and the total number of instances evaluated per preposition (D = Disagree, A = Agree).

Preposition	Parent Category	Right Sibling	Count	Likelihood
by	PP-MNR	S-NOM	81	High
by	PP	NP-LGS	25	Low
with	PP	NP	349	Low
with	PP	S-NOM	9	High

Table 5: Syntactic patterns for under-annotated candidates agreed by both models.

in "*Big Board traders praised the Jacobson specialists for getting through yesterday [without] a trading halt*", both models misinterpret the noun phrase "a trading halt" as a clausal complement.

Overall, model agreement provides a practical way to detect likely under-annotations while also exposing systematic syntactic errors. These findings motivate the use of syntactic pattern analysis as an additional filtering step.

4.3. Syntactic Patterns

After model agreement filtering, we performed a syntactic analysis using PTB parse trees to examine whether certain syntactic patterns could reinforce or disconfirm model predictions. We focused on the prepositions *by* and *with*, which have comparable frequencies (5,626 and 5,327) and annotation rates (7.73% and 5.58%) in PDTB-3, making them suitable for comparison. A complete inventory of syntactic patterns observed in model-agreed under-annotated instances is provided in Appendix A.

Following Pitler and Nenkova (2009), we examined the parent and right sibling categories of each preposition. Both features were informative, but patterns differed from those reported for other connectives, since prepositions were not annotated in PDTB-2.

For most annotated instances, the parent node was a prepositional phrase (PP), often marked with a function tag such as `-MNR` (manner), `-TMP` (temporal), or `-EXT` (extent). Discourse-related usages of *by* frequently appeared under `PP-MNR`, while *with* showed greater variation, including a notable number of `PP-CLR` (closely related) cases among unannotated instances, which typically signal a non-discourse relation.

Right sibling patterns further distinguished discourse and non-discourse uses. For *by*, 99.5% of annotated instances had an `S-NOM` (gerund clause) sibling, while unannotated ones were mostly followed by an `NP-LGS` (logical subject). Similarly, *with* was more likely to be discourse-related when followed by an `S-NOM` rather than an NP. This suggests that prepositions followed by clausal complements are strong candidates for under-annotation.

Table 5 summarises the dominant syntactic patterns among candidate instances identified by both GPT-4o and LLaMA-3-70B. Patterns combining a `PP-MNR` parent and `S-NOM` sibling are particularly indicative of discourse usage, while cases with nominal right siblings (`NP` or `NP-LGS`) are less likely to signal discourse relations. These features can thus help prioritise which instances to review manually.

These findings show that parse-tree features can strengthen confidence in model-predicted under-annotations and filter out unlikely cases, though inconsistencies in PTB parses (notably for *with*) limit reliability. Future work should validate these observations through targeted annotation of high-confidence cases.

4.4. Discussion

The results indicate that PDTB-3 likely contains under-annotated explicit discourse connectives, and that the proposed LLM-based pipeline can effectively narrow large candidate sets into smaller, high-confidence subsets.

Prompt design had a substantial impact on performance. While optimal templates varied across prepositions and models, the *Contextualised Explanation Prompt*, which combined definitions and examples yielded the most stable results. GPT-4o consistently outperformed LLaMA-3-70B, and including examples improved classification accuracy across all tasks.

In the full-scale evaluation, GPT-4o achieved high recall for *by*, *with*, and *without*, but struggled with *for*. A clear trade-off emerged: prompts

with stronger constraints improved precision but missed valid cases, while less constrained prompts encouraged overgeneralisation and false positives—particularly for *with* and *for*. Model agreement between differently prompted LLMs reduced this effect, yielding smaller and more reliable candidate lists. Remaining errors were mainly due to syntactic misinterpretation or "faithfulness hallucination," where models confidently labelled non-discourse uses as positive.

Syntactic pattern analysis complemented model agreement by identifying structural cues linked to discourse usage. When a preposition's right sibling was a clausal constituent, especially an `S-NOM` clause, the instance was far more likely to express a discourse relation. Such patterns, common in annotated cases but rare in unannotated ones, provide useful heuristics for prioritising manual review. However, inconsistencies in PTB parses limit full reliability.

Several limitations should be noted. Prompt design was explored only to a limited extent, and computational constraints restricted the number of models and examples tested. Expanding the model pool may improve robustness, as agreement across a broader set of architectures could further strengthen confidence in high-likelihood candidates. In parallel, more systematic prompt engineering which potentially tailored to model-specific strengths, may improve stability and reduce overgeneralisation in the error-detection process.

Additionally, syntactic analysis relied solely on the PTB; preliminary inspection with other parsers revealed promising patterns, but these were not systematically validated. The present analysis also focused primarily on instances flagged by the models, and did not systematically examine cases that were not identified as candidates. As a result, the extent to which the pipeline may miss valid but under-annotated instances remains unclear. Although the present study focuses on prepositions that can function as explicit discourse connectives, the proposed semi-automatic pipeline is designed to support human annotators rather than replace them. With further refinement of prompts and syntactic pattern filters, the framework could be generalised to other types of ambiguous expressions.

Finally, this work provides a foundational step toward more reliable explicit discourse connective detection. By reducing annotation ambiguity in PDTB-style corpora, improved corpus consistency may in turn benefit downstream discourse parsers trained on such resources, potentially enabling higher performance in connective identification and relation classification.

5. Conclusion

This study introduced a prompting-based LLM pipeline for detecting potentially under-annotated explicit discourse connectives in PDTB-3, focusing on four newly annotated prepositions. The proposed approach effectively reduces large candidate sets to smaller, high-confidence subsets, providing a practical semi-automatic aid for improving corpus consistency.

While the pipeline achieved strong performance for several prepositions, its generalisability remains constrained. Performance depends on the availability of annotated examples, and syntactic inconsistencies in the PTB parses can affect the reliability of structural filtering. Moreover, prompt sensitivity and model-specific behaviour indicate that broader model consensus and more systematic prompt design may further enhance robustness.

Future work will extend connective coverage, incorporate additional models to strengthen agreement-based validation, and explore more systematic prompt engineering. Integrating alternative syntactic resources may also improve structural reliability. Targeted analysis of cases not identified by the models will also help characterise potential false negative behaviour of the pipeline. A crucial next step is the involvement of human annotators in reviewing both high- and lower-confidence subsets identified by the models. Evaluating the proportion of valid annotations in each subset will provide an empirical assessment of the pipeline's practical effectiveness.

More broadly, this work demonstrates how LLMs can function as annotation-support tools in discourse research. By reducing ambiguity and improving consistency in PDTB-style corpora, the proposed framework has the potential to enhance the reliability of downstream discourse parsing systems trained on these resources.

Acknowledgements

We thank Prathyusha Jwalapuram for insightful discussions, valuable suggestions, and guidance throughout the development of this study, including the early conceptualisation of the research direction.

6. Bibliographical References

Joshua Achiam, Steven Adler, Sahil Agarwal, Liane Ahmad, Ilge Akkaya, Francisco López Aleman, Bob McGrew, et al. 2023. Gpt-4 technical report. ArXiv preprint arXiv:2303.08774.

Katherine Atwell, Junyi Jessy Li, and Malihe Alikhani. 2021. Where are we in discourse relation recognition? In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 314–325.

Ann Bies, Mark Ferguson, Karen Katz, Robert MacIntyre, Victoria Tredinnick, Grace Kim, Mary Ann Marcinkiewicz, and Britta Schasberger. 1995. Bracketing guidelines for treebank ii style penn treebank project. Technical Report 97, University of Pennsylvania, Philadelphia, PA, USA.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Yulin Chang, Xin Wang, Jing Wang, Yuwei Wu, Liang Yang, Kai Zhu, and Xiaohui Xie. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3563–3578, Bangkok, Thailand. Association for Computational Linguistics.

Markus Dickinson and Detmar W. Meurers. 2003. Detecting inconsistencies in treebanks. In *Proceedings of the Third Workshop on Treebanks and Linguistic Theories (TLT 2003)*, volume 3, pages 45–56, The Ohio State University, Columbus, OH, USA.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.

Andrea Grattafiori, Aman Dubey, Ankit Jauhri, Arpit Pandey, Aditya Kadian, Ahmad Al-Dahle, and Dejan Vasic. 2024. The llama 3 herd of models. ArXiv preprint arXiv:2407.21783.

Liang Huang, Wenhao Yu, Wen Ma, Wei Zhong, Zhen Feng, Han Wang, Ting Liu, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Shamsuddeen Ibn Faiz and Robert E. Mercer. 2013. Identifying explicit discourse connectives in text. In *Canadian Conference on Artificial Intelligence*, pages 64–76, Berlin, Heidelberg. Springer.

Groq Inc. 2024. Groq api documentation. `https://groq.com`. Accessed August 2025.

Isaac J. Jacob, Shankar K. Shanmugam, Selwyn Piramuthu, and Piotr Falkowski-Gilski. 2022. *Data Intelligence and Cognitive Informatics*. Springer Singapore, Singapore.

Anders Johannsen and Anders Søgaard. 2013. Disambiguating explicit discourse connectives without oracles. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 997–1001.

Jan-Christoph Klie, Bonnie Webber, and Iryna Gurevych. 2023. Annotation error detection: Analyzing the past and present for a more coherent future. *Computational Linguistics*, 49(1):157–198.

Rico Knaebel and Manfred Stede. 2020. Contextualized embeddings for connective disambiguation in shallow discourse parsing. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 65–75.

Joshua Ong Jun Leang, Aryo Pradipta Gema, and Shay B Cohen. 2025. CoMAT: Chain of mathematically annotated thought improves mathematical reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20245–20274, Suzhou, China. Association for Computational Linguistics.

Yen-Ting Lin and Yun-Nung Chen. 2023. LLM-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 47–58, Toronto, Canada. Association for Computational Linguistics.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A pdtb-styled end-to-end discourse parser. *Natural Language Engineering*, 20(2):151–184.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, and Bill Schasberger. 1994. The penn treebank: Annotating predicate argument structure. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8–11, 1994*, Plainsboro, New Jersey.

Mor Mizrahi, Gal Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? a call for multi-prompt llm evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949.

Curtis Northcutt, Lu Jiang, and Isaac Chuang. 2021. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411.

OpenAI. 2023. Gpt-4 technical report. https://doi.org/10.48550/arXiv.2303.08774. ArXiv:2303.08774.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 683–691. Association for Computational Linguistics.

Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 13–16. Association for Computational Linguistics.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rashmi Prasad, Bonnie Webber, and Aravind Joshi. 2017. The penn discourse treebank: An annotated corpus of discourse relations. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 1197–1217. Springer Netherlands, Dordrecht.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7, New York, NY, USA. Association for Computing Machinery.

Villy Rouchota. 1996. Discourse connectives: What do they link. *UCL Working Papers in Linguistics*, 8:199–214.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, Seattle, United States. Association for Computational Linguistics.

Merel Scholman, Tao Dong, Frances Yung, and Vera Demberg. 2021. Comparison of methods for explicit discourse connective identification across various domains. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 95–106.

Syed Shahriar, Björn Lund, Nikhil R. Mannuru, Muhammad A. Arshad, Khaled Hayawi, Raj V. K. Bevara, and Laila Batool. 2024. Putting gpt-4o to the sword: A comprehensive evaluation of language, vision, speech, and multimodal proficiency. ArXiv preprint arXiv:2407.09519.

Haotian Sun, Xinyang Li, Yikai Xu, Yusuke Homma, Qiyang Cao, Mengzhou Wu, Diamond Charles, et al. 2023. Autohint: Automatic prompt optimization with hint generation. ArXiv preprint arXiv:2307.07415.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Maite Taboada. 2006. Discourse markers as signals (or not) of rhetorical relations. *Journal of Pragmatics*, 38(4):567–592.

Ann Taylor, Mitchell Marcus, and Beatrice Santorini. 2003. The penn treebank: An overview. In Anne Abeillé, editor, *Treebanks: Building and Using Parsed Corpora*, pages 5–22. Springer, Dordrecht.

Hugo Touvron et al. 2024. Llama 3: Open foundation and fine-tuned chat models. https://ai.facebook.com/blog/large-language-model-llama-3-open-foundation-and-fine-tuned-chat-models. Accessed: August 2025.

Ji Wang and Man Lan. 2015. A refined end-to-end discourse parser. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, pages 17–24.

Bonnie Webber, Rashmi Prasad, Aravind Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual. Technical Report 35, University of Pennsylvania, Philadelphia.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Xinyu Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. In *Advances in Neural Information Processing Systems*, volume 35, pages 30378–30392. Curran Associates, Inc.

Lingzhi Zheng, Wei-Lun Chiang, Yizhe Sheng, Shuchang Zhuang, Ziyi Wu, Yong Zhuang, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

7. Language Resource References

Marcus, Mitchell P. and Marcinkiewicz, Mary Ann and Santorini, Beatrice. 1993. *The Penn Treebank*. Linguistic Data Consortium. Linguistic Data Consortium (LDC), University of Pennsylvania, Treebank Resources, 1.0, ISLRN 060-785-139-403-2.

Prasad, Rashmi and Dinesh, Nikhil and Lee, Alan and Miltsakaki, Eleni and Robaldo, Livio and Joshi, Aravind and Webber, Bonnie. 2008. *The Penn Discourse Treebank 2.0*. Linguistic Data Consortium. Linguistic Data Consortium (LDC), University of Pennsylvania, PDTB Resources, 2.0, ISLRN 488-589-036-315-2.

Prasad, Rashmi and Webber, Bonnie and Lee, Alan. 2019. *The Penn Discourse Treebank 3.0*. Linguistic Data Consortium. Linguistic Data Consortium (LDC), University of Pennsylvania, PDTB Resources, 3.0, ISLRN 977-491-842-427-0.

Appendix

A. Syntactic Pattern Inventory

A.1. Preposition *by*

Table 6: Syntactic patterns for under-annotated *by* instances agreed upon by GPT-4o and LLaMA-3-70B.

Parent Sibling	Right Sibling	Count
PP-MNR	S-NOM	81
PP	NP-LGS	25
PP	S-NOM	8
PP-MNR	NP	8
PP	NP	3
PP-PRD	S-NOM	2
WHPP-1	WHNP	2
PP	S-NOM-LGS	1
PP	SBAR-NOM-LGS	1

A.2. Preposition *with*

Table 7: Syntactic patterns for under-annotated *with* instances agreed upon by GPT-4o and LLaMA-3-70B.

Parent Sibling	Right Sibling	Count
PP	NP	349
PP-CLR	NP	51
PP-MNR	NP	18
PP	S-NOM	9
PP	FRAG	3
SBAR	S	3
SBAR-ADV	S	3
S	NP-SBJ	3
S	PP	1
PRT	NP-SBJ	1
PP-TMP	NP	1
PP-3	NP	1
NP-SBJ	NP	1
PP	PRN	1
PP	SBAR	1
PP	UCP	1
PP-CLR-TPC-1	NP	1
PP-TPC-2	NP	1