On the Role of Context for Discourse Relation Classification in Scientific Writing

Stephen Wan[♣] **Wei Liu**[♢] **Michael Strube**[♢]
[♣] CSIRO Data61, Australia
[♢] Heidelberg Institute for Theoretical Studies, Germany
stephen.wan@data61.csiro.au
{wei.liu, michael.strube}@h-its.org

Abstract

With the increasing use of generative Artificial Intelligence (AI) methods to support science workflows, we are interested in the use of discourse-level information to find supporting evidence for AI generated scientific claims. A first step towards this objective is to examine the task of inferring discourse structure in scientific writing. In this work, we present a preliminary investigation of pretrained language model (PLM) and Large Language Model (LLM) approaches for *Discourse Relation Classification* (DRC), focusing on scientific publications, an under-studied genre for this task. We examine how context can help with the DRC task, with our experiments showing that context, as defined by discourse structure, is generally helpful. We also present an analysis of which scientific discourse relation types might benefit most from context.

1 Introduction

Recent Artificial Intelligence (AI) advances coupled with the agentic AI approach have seen a burst of activity in the area of "AI for Science", the application of AI techniques to help accelerate scientific discovery. Examples include usage of *Google's Co-scientist* (Penadés et al., 2025), OpenAI's *Deep Research*[1], NVidia's foundation models for life sciences[2] and the agentic AI platform *Future House*[3]. Many of these tools offer an AI research assistant that helps complex research information needs, such as question answering and research planning.

Within these AI for Science applications, generative AI approaches based on Large Language Models (e.g., Brown et al., 2020) are used to generate answers (novel text) to complex questions, introducing the problem of addressing hallucinations and lack of faithfulness (to source references) (Fang et al., 2024).

A popular approach to these problems is to show passages from the source material that supports the generated answer. This approach, sometimes referred to as "contextualising scientific claims", was the focus of the *Context24* shared task (Chan et al., 2024). Interestingly, the leading contribution in the Context24 shared task demonstrated the utility of scientific discourse cues for detecting such justification material (Bölücü et al., 2024).[4] This raises an interesting question: *can discourse information be further employed to help in providing supporting evidence for generative AI answers to scientific questions?* A necessary precursor to such an approach would be the ability to infer the discourse structure of a given paper. As a first step towards a study of this topic, in this paper, we focus here on studying the technical challenge of inferring the discourse relations between passages of scientific writing.

We focus on data from two discourse datasets for scientific text, SciDTB (Yang and Li, 2018) and CovDTB (Nishida and Matsumoto, 2021). Both datasets follow the approach that annotates discourse structure as dependency trees, introduced in the SciDTB approach (Yang and Li, 2018). To our knowledge, these are the largest discourse datasets currently available. An example of such a discourse tree is presented in Figure 1.

We present an example from the SciDTB dataset (Yang and Li, 2018) in Figure 2.[5] The figure shows segmented elementary discourse units (EDUs) for the arguments of the relation. The ground truth relation between the two arguments was annotated as *condition*.[6] This classification task is highly ambiguous. We note that, in this dataset, the word

[1] https://openai.com/index/introducing-deep-research
[2] https://www.nvidia.com/en-au/use-cases/biomolecular-foundation-models-for-discovery-in-life-science
[3] https://www.futurehouse.org/

[4] Discourse cues are described as *common expressions* in the original paper.
[5] SciDTB dataset, document ID:P14-1131.
[6] That is, a conditionality for a given situation.

Proceedings of the 6th Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences (CODI 2025), pages 96–106

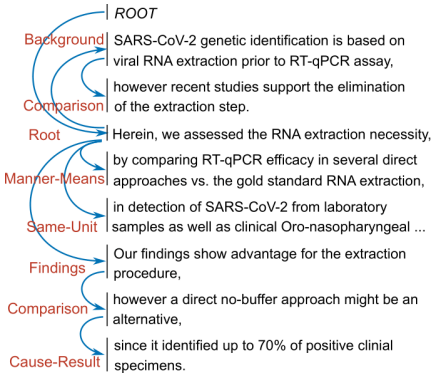

Figure 1: An example of a dependency discourse structure for an abstract for COVID19-related science article. Figure from (Nishida and Matsumoto, 2021)

EDU 1 *EDU 2*	because it can compute a single node similarity (rel:condition) *without* having to compute the similarities of the entire graph .
context	(rel:elaboration) that is efficient . . .

Figure 2: An example of how contextual information can help disambiguate which discourse relation holds between two text fragments.

"without" that begins *arg 2* is also associated with two other relations such as *contrast* or *manner*. This ambiguity can be alleviated with additional context. In the example, *arg 1* is annotated as being preceded by the fragment that has an *elaboration* relation and is about *efficiency*. Of the possible relations, this context potentially provides additional information that supports *conditionality* as a more favourable interpretation compared to *contrast* or *manner*.[7]

Inspired by examples such as this and also by recent work in discourse analytics that examines the role of context for DRC in other text genres (e.g., Zhang et al., 2021; see Section 2 for a comprehensive survey), we are further interested in examining the role discourse context for the DRC task. As the majority of these studies have focused on the Penn Discourse Treebanks (PDTB) (Webber et al., 2019) and given that scientific writing (and the discourse relations therein) as notably different from other genres (e.g., see Shi and Demberg, 2019), our goal is to study how context affects DRC for scientific writing. In particular, we are interested in examining whether context selection informed by discourse structure, which we refer to here as *discourse structured context*, has advantages over other methods, such as adjacent text spans.

This paper makes several unique contributions. We show that structured context helps to improve DRC for scientific writing, as represented by the two datasets, finding that this approach benefits both pre-trained language model and large language model approaches. Furthermore, we present an error analysis to explore the situations context in which context is helpful, revealing some interesting correspondences between scientific discourse relations within the two datasets.

2 Related Work

2.1 Discourse Relation Classification

Discourse Relation Classification (DRC), the task of inferring the relation type that holds between two EDUs, has a rich history and is an actively researched area (Pitler et al., 2009; Gessler et al., 2021; Long and Webber, 2022; Zhou et al., 2022; Liu and Strube, 2023a). The DRC task is typically divided into two types, *explicit* and *implicit* (Pitler et al., 2009). In the former, the two text spans are connected by a discourse relation signalled by an observable discourse connective. The latter is termed implicit as there is an absence of such a connective. The implicit variant is regarded as much harder than the explicit variant (Pitler et al., 2009).

In recent years, to facilitate DRC research, shared tasks have been organised within the DISRPT series (Zeldes et al., 2021; Braud et al., 2023). These shared tasks had the benefit of broadening the datasets used to evaluate DRC approaches, which has tended to focus on the PDTB (Prasad et al., 2008). Notably, the two datasets used in this paper were introduced in the 2023 shared task (Braud et al., 2023), however, the submitted approaches at the time did not focus on analysing the use of context for these datasets.

2.2 Context in DRC

The position survey article of Atwell et al. (2021) notes that "most shallow discourse parsers use only the argument pairs to determine the discourse sense without considering ... context." There have, however, been some exceptions. The effect of using discursive context on DRC has been studied in the context of annotation quality and annotator confidence in (Atwell et al., 2022). This work examines the role of context in the PDTB 2 (Prasad et al.,

[7] That is, in the example, *computation* is described as *efficient*, not because of an explicit comparison (contrast) nor an indication of how to perform a task (manner) but by virtue of conditions, in this case, *without* the described *expenses*.

2008) and 3 (Webber et al., 2019) datasets and shows that annotation quality improves with context (in this case, preceding text), particularly for certain relationships. In that same work, Atwell et al. translate the insights from human annotations and context into modelling insights, exploring the use of the XLNet model for classification with context, that incorporates some modelling of when context is needed. Atwell et al. note the prior work of Scholman and Demberg (2017) in examining the linkage between pronoun use and the need for context within PDTB, again from the perspective of acquiring human annotations.

The sequential modelling of adjacent text spans (and their relationships) has been studied by Dai and Huang (2018), Shi and Demberg (2019) and Zhang et al. (2021). These works generally evaluate on PDTB data. The exception is Shi and Demberg (2019), who also evaluate on biomedical data, albeit by mapping biomedical relations to PDTB discourse relations. Shi and Demberg also argue that the Next Sentence Prediction (NSP) capability of the the BERT model is particularly useful for modelling discourse relations. This same approach is used by Gessler et al. (2021), who also use BERT specifically for its NSP capability but add features relating to the surrounding context. This included direction features and embeddings of the surrounding sentences (to the text spans being considered). Zhang et al. (2021) model discourse structure as a graph and use graph representations in a neural network to capture discourse context, showing benefits for discourse relation classification. These investigations provide alternative representations of context to our study, which does not use specific features or graph representations of context. Our study differs in that we use various methods to select context, including a consideration of the discourse dependency tree, and we prepend contextual text to the EDUs being judged.

While prior treatments of context have shown that it is useful for DRC, these studies generally focus on non-scientific writing, like the PDTB. We note that the approach of Gessler et al. (2021), and the following work of Metheniti et al. (2024), is evaluated on the DISRPT 2021 dataset. However, while this dataset includes data a range of genres, it does not include scientific articles. Indeed, Shi and Demberg (2019) note that the discourse relations are notably different in science literature compared to relations from the PDTB. We argue that this difference thus merits a dedicated study of the effects of context for science literature.

Dataset	Train:Dev:Test size	Genres
CovDTB	(2400):2400:2587	Science (Biomed)
SciDTB	6061:1935:1912	Science (NLP)

Table 1: Overview of the scientific discourse datasets studied in this work.

2.3 Other Discourse-level Analytics for Scientific Writing

Related to the task of inferring discourse relations for scientific writing are the tasks of argument zoning (Teufel et al., 2009), citation function classification and classification (Teufel et al., 2006; Wan et al., 2009), scientific argument mining (e.g., (Lawrence and Reed, 2019; Accuosto and Saggion, 2020; Binder et al., 2022; Fergadis et al., 2021) and science communication sentence classification (Louis and Nenkova, 2013; August et al., 2020); all of which infer a discourse-level features relating to argumentation or scientific writing structure. For this body of work, we note that the use of context has been studied has been studied within the topic of sentence-level categorisation for scientific function (Kiepura et al., 2024). While these associated fields can inform our work, in this paper, we focus on discourse relations rather than argument zones, single sentence classification, or discourse-level relationships across citations.

3 Data

Here, we focus on data from prior work in discourse analysis that provides ground truth annnotations for scientific discourse structure, namely the the Covid (Discourse) Dependency Treebank (CovDTB) (Nishida and Matsumoto, 2021), and the Science (Discourse) Dependency Treebank (SciDTB) (Yang and Li, 2018).

3.1 Discourse Dependency Representations

The SciDTB and CovDTB datasets use the dependency discourse structure (DDS) introduced in the SciDTB work (Yang and Li, 2018). Here, structures are directed acyclic graphs, specifically trees. An example of a DDS is shown in Figure 1. In this structure, nodes are EDUs and edges represent the labelled relations between EDUs. The DDS will be used to help select relevant contexts for relation classification. In the DDS, each directed edge (arrows) is a dependency. The figure shows name of

the relation as the text in red. The direction of the relation indicates the importance of the information, with the arrowhead indicating the less important information. By following the links back through the tree towards the root, one can select the relevant context for the classification task, which often is associated with information of greater importance.

3.2 Context Selection Schemes

A context selection scheme relies on two sub-steps: segmentation and filtering preceding text. For this study, for a given EDU in the datasets, there are several options to select context. The simplest filtering method is a *null* method, where no context is used. Alternatively, one can include a preceding text window of some size s. For example, we can also always select the previous sentence for a given argument. However, just because some text precedes an argument does not necessarily mean it is relevant context that impacts text understanding. Thus, we also explore a discourse-oriented method where a discourse structure is employed to identify other text (EDUs) that are linked via discourse relations (referred to here as *structured context*). This gives rise to the following schemes for context selection:

Default This baseline is a *null* context (that is, no context is used).

(ADn) Add-n This approach will add n sentence that precede the first argument.

(ORn) Oracle-n This scheme relies on the ground truth annotations to select the preceding context.[8] We use the following algorithm. For a pair of arguments considered for DRC, we find the parent node of the *first argument* as defined by the dependency discourse tree. By chaining together the preceding context, in principle, we can vary the amount of context to include. The intuition is that the discourse context, generally represented by a chain of EDUs following a path to the root, indicates which context is important enough to extract.

4 Models

We focus on two general neural network language models approaches, both based on the Transformer architecture (Vaswani et al., 2023). The first approach employs the RoBERTa model (Liu et al., 2019), an example of a non-auto-regressive (pretrained) language model which one can finetune for classification. The second approach uses large language models to perform prompt-based inference, typically used for generative AI.

4.1 PLM-finetuning with RoBERTa

Our RoBERTa-based approach is based on an approach that jointly models discourse connective generation and DRC (Liu and Strube, 2023a). This RoBERTa-based model performs training that combines two tasks: (1) the generation of a discourse connective that would link two arguments; and (2) the classification of the relation given the three pieces of information (argument 1, argument 2, and a connective). As such, this model is extremely flexible and can be applied to both implicit and explicit DRC.

In this study, we generate variants of the data sets, subject to the preprocessing outlined in Section 3, that differ by the amount and type of contextual information that is inserted *before* the first argument (of the relation classification task). That is, context added as per the selection schemes above. The datasets are split into training and testing subsets to train and evaluate the RoBERTa model.

We experimented with the full joint model described here and a simpler version that focuses just on classification.[9] The latter was found to perform the best and so we report only these results. We used the default training setup and parameters, following the documentation in Liu and Strube (2023b).[10]

4.2 LLM-based Inference

Prompt-based generative AI approaches using Large language models (LLMs) have been revolutionary in providing new baseline solutions for many tasks that apply across domains. A key feature of LLMs are the comprehensive training regimes that potentially captures different kinds of knowledge, including common sense knowledge and linguistic capability (e.g., Brown et al., 2020). In this work, we examine two classes of LLMs for this inference approach: *open* and *closed* weight approaches. For the former, we use Meta's LLaMA

[8] For our motivating example of *contextualising claims*, in practice we would need an initial method to compute a discourse graph connecting EDUs. We leave this to further work but note that prior work explores this task (e.g., Jeon and Strube, 2020).

[9] This is achieved by setting the connective to a constant value for all data.

[10] For computing environment, see Appendix A.

```
Replace the MASK token (a discourse
relation) by selecting only one of the
following labels: [ label1, label2,
... , labeln] Examples: Passage
1: <arg1ex1>, Passage 2: <arg2ex1>,
connective: <connectiveex1> | label1 EOL
<further examples for remaining labels>
EOL
Passage 1: <arg1>, Passage 2: <arg2>,
connective: <connective> | [MASK]
```

Figure 3: GPT4 Prompt template for discourse relation classification

model (et al., 2024). For the latter, we use OpenAI's GPT4 (OpenAI, 2024).

LLaMA 3.1 In this work, we use a locally hosted version of the LLaMa-3.1-70b-Instruct model, hosted on a server with 16 CPU cores, 128GB memory, and three A6000(48GB) which host the Llama 3.1 70B model.

GPT4 For the GPT4 model, we use OpenAI's API with the *chatcompletion* endpoint, using the gpt-4-0613 model.

For both models, for classification, we use In-Context Learning (ICL), widely acclaimed as having the ability to surpass supervised machine learning on many NLP tasks as a so-called "few-shot learner" (Brown et al., 2020). This model helps us estimate how context impacts the current methods for LLM-based inference. For this work, temperature was set to zero.

An example of prompt template for GPT4 is presented in Figure 3. This frames the relation classification task as a MASK replacement generation task. The prompt issues an instruction to replace the [MASK] token with one of the class labels provided in list format. ICL "training" examples are provided, where we randomly sample one training data point from the data set for each class label. The two arguments, the connective, and the [MASK] token are then added. The Llama 3.1 prompt is similar.

5 Experiment Results

We conduct an empirical study using ground truth linguistic data to examine the role of discourse structure in inferring discourse relations. We deploy the described models in two conditions: a control condition without discourse context and an experiment condition that includes context.

In the case of PLM-finetuning, for each pair of conditions, we run 10 different trials (that is, neural network training and testing with 10 different random seeds). We report macro-F1 classification results averaged over the 10 trials. For significance testing, we used the Wilcoxon Signed Ranked Test (WSRT) (Wilcoxon, 1945) and corrected for multiple comparisons using Bonferroni correction.

We first examine if *any* use of context leads to an improvement in the discourse classification task performance. For this investigation, we use n=1 for the context selection schemes.

Approach	CovDTB	SciDTB
default	75.54 (1.11)	57.42 (0.65)
AD1	73.45 (2.26)	**57.75 (1.08)**
OR1	**75.78 (1.52)**	**58.33 (0.77)**†

Table 2: Classification with a fine-tuned RoBERTa model. Macro-F1 scores (averaged over 10 runs) with standard deviations in parentheses. Bolded values indicate improvements above the default. Daggers indicate statistical significance improvement using the Wilcoxon Signed Rank Test († : $\alpha = 0.05$).

5.1 Context and PLM Fine-Tuning

Table 2 shows the results of including context for the PLM-finetuning approach using the RoBERTa model. The table presents macro-F1 scores to give some indication of performance across the unbalanced dataset.[11] We find that context generally helps for DRC when using a fine-tuned PLM, particularly when context is defined using discourse structure (OR1). This improvement is statistically significant for the SciDTB dataset (WSRT $p < 0.05$). The AD1 context does not lead to strong performance improvements in comparison. However, we note that the AD1 text window variant of context also helps mildly for the SciDTB but not for CovDTB.

Across the two datasets, we observe that the performance results are higher in the CovDTB dataset compared to the SciDTB dataset. This could be due to conventions in scientific writing for biomedical literature which may be more homogenous than the data from NLP domain found in SciDTB.

[11] Here we report results for $n = 1$ as our experiments showed that for larger values, adding more context confused the models. Similarly, while we explored variants of the context representations that additionally utilised the relation class (for the linked context), the results were comparable to reported the OR1 variant, from which we conclude that the extra information did not help.

Approach	CovDTB	SciDTB
default	32.07	22.06
AD1	26.19	19.28
OR1	**49.07**	**52.61**

Table 3: GPT4 model: Classification macro-F1 scores.

Approach	CovDTB	SciDTB
default	**11.20**	07.71
AD1	10.04	05.15
OR1	10.36	**11.15**

Table 4: Llama 3.1 model: Classification macro-F1 scores.

5.2 Context and LLM Prompt-based Inference

Table 3 presents the corresponding results for the GPT4 model.[12] The results indicate a poor performance by the LLM for DRC under the default setting (with no context), even when using in-context learning. Performance improves when discourse structure is used to provide context, as opposed to the adjacent text. Indeed, performance drops when the adjacent sentence is used as context. In the case of SciDTB, this brings performance closer to the PLM-finetuning result, within a margin of 5 F1 points. However, while the DRC performance on CovDTB increases with discourse structure context, the macro-F1 scores still remain far behind the fine-tuned PLM, by a margin of over 25 F1 points.

In Table 4, we see a similar story, although Llama 3.1 performance is much lower than GPT4. We suspect that this is primarily due to the size of the model; the Llama model used here has orders of magnitude fewer parameters than the GPT model. Again, using the adjacent sentence leads to a drop in performance. Consistent with other results, an improvement in DRC performance was observed for SciDTB when using context defined by discourse structure. Here, we failed to detect any improvement with the CovDTB dataset.

We note that, while our focus is on comparison against our default baseline and the relative difference in performance with and without context, the models described in this paper are competitive with the reported performance in the literature. We report these values for completeness in Table 5, which lists comparisons with literature, with the metrics generally reported by convention. With the RoBERTa model and the "oracle" use of the ground truth discourse annotations used here, the measured accuracy of 83.63 represents an estimate of an improvement over the prior state-of-the-art (SOTA) result for the CovDTB that could be obtained if we were to employ a fully automated version of the inference.

Approach	CovDTB	SciDTB
Performance	70.03 Acc.	75.30 Acc.
OR1 model	**83.03** Acc.	74.81 Acc.

Table 5: A comparison with performance reported in the literature. **Bolded** values indicate where our best RoBERTa model surpasses previous results. Reported performance from: covdtb and scidtb (Liu et al., 2023).

To summarise, the results show positive trends for using discourse context for the DRC task. Generally, discourse context can help with the PLM-finetuning approach and LLM-inference. When applying the text window context (AD1) results are mixed; the method does not work all the time and can decrease performance. However, when using discourse structure to determine relevant context (OR1) we generally see improved performance, with stronger gains demonstrated with the SciDTB dataset. This indicates that not all preceding text is useful for classification and that indiscriminately adding more context (without filtering) can make performance worse.

5.3 A Reflection on Datasets

We speculate that one reason why we do not see a bigger effect from the inclusion of discourse context is that our datasets may be limited to relatively short length of the text data. Indeed, Yang and Li (2018) note a related issue when studying news articles: discourse relations do not cross paragraph boundaries further making structures shallow.[13]

In this regard, the CovDTB and SciDTB datasets, as examples of short text data (i.e., abstracts) may also have simpler discourse structures than longer texts. We further investigated the nature of the discourse structures and found that, in the case of the SciDTB dataset, the structures were generally short-distance dependencies: 61% of relationships are adjacent, with 10% of relations separated by a gap of 3-5 sentences. We posit that when considering longer documents, the effect of structured context in DRC may be more pronounced.

[12] Given the cost of using the commercial GPT LLM, we report results on single trials for the datasets.

[13] This is presumably due to journalistic writing style.

6 Error Analysis: DRC for Scientific Discourse Relations

In the experiments presented above, we observed that providing context, particularly *structured* context generally helps with DRC. In this section, we perform an error analysis to better understand when context helps, analysing performance per relation type. Here, we focus on the fine-tuned PLM approach as it yielded the highest macro-F1 scores. For each of the 10 seed runs, we used predictions from the best models for the default (no context) and the OR1 (structured context) conditions. Cases where the OR1 prediction was correct and the default was not was considered a *win*. The converse case was considered a *loss*. Where both approaches agreed, this was considered a *tie*. Margins for wins and loss were averaged over the 10 runs.

Table 6 provides a list of the scientific discourse relation types in the *winning* outcome for both data sets that benefited (overall) from OR1 context and their average win margins. We can see that *elaboration*, *comparison*, *attribution*, and *temporal* relations were common to both datasets.

CovDTB	SciDTB
elaboration ($\Delta = 5.7$)	**elab**-addition ($\Delta = 1.5$)
enablement ($\Delta = 1.2$)	**elab**-aspect ($\Delta = 0.8$)
cause-result ($\Delta = 0.8$)	**temporal** ($\Delta = 0.77$)
condition ($\Delta = 0.4$)	bg-**compare** ($\Delta = 0.66$)
attribution ($\Delta = 0.4$)	joint ($\Delta = 0.44$)
comparison ($\Delta = 0.3$)	contrast ($\Delta = 0.33$)
temporal ($\Delta = 0.1$)	progression ($\Delta = 0.22$)
	exp-reason ($\Delta = 0.22$)
	elab-enum ($\Delta = 0.22$)
	comparison ($\Delta = 0.11$)
	attribution ($\Delta = 0.11$)

Table 6: Winning relations: Discourse relations who DRC performance improved with the inclusion of structured context. $\Delta =$ indicates the average win/loss margin. Bolded text indicates potential correspondences across datasets.

Table 7 presents the corresponding table for the *losing* outcome. Here we see that, both data sets have *background* relations in common for this outcome. If we assume *findings* and *results* are related relations, then we can consider this a further potential alignment.

Table 8 shows the relations that had an equal number of wins and losses. We present these for completeness. However, it may be the case that, for these datasets, there is insufficient data to assign these to either the winning or losing outcomes.

There were some differences between datasets

CovDTB	SciDTB
findings ($\Delta = 0.9$)	**bg**-goal ($\Delta = 0.44$)
background ($\Delta = 0.3$)	manner-means ($\Delta = 0.44$)
	enablement ($\Delta = 0.44$)
	evaluation ($\Delta = 0.22$)
	result ($\Delta = 0.22$)
	bg-general ($\Delta = 0.22$)
	condition ($\Delta = 0.11$)
	exp-evidence ($\Delta = 0.11$)

Table 7: Losing relations: Discourse relations who DRC performance suffered with the inclusion of structured context. $\Delta =$ indicates the average win/loss margin. Bolded text indicates potential correspondences across datasets.

CovDTB	SciDTB
textual-organisation	elab-definition
manner-means	elab-process-step
	cause

Table 8: Tied relations: Discourse relations who DRC performance remained the same with the inclusion of structured context.

for a subset of relations, which were placed in different outcomes (winning, losing). These included *enablement* and *condition*. In CovDTB, a single relation is used for *cause-result* which was in the winning outcome. For the SciDTB dataset, the *result* relation was in the losing outcome. Similarly, while most background relations were in the losing outcome for both datasets, *bg-compare* was in the winning outcome for SciDTB; though this could be because the winning outcome contained more comparison-related relations. We treat these divergences as interesting outcomes to investigate further, noting that some of these may be due to annotation differences between the datasets.

In Table 9, we present some examples of data as assigned to the winning and losing outcomes. For the winning outcomes, the high-level statement of the research activity as context may contribute positively to the DRC task. For the losing outcome, we note that in the SciDTB example, the high-level context may simply be too broad. For the CovDTB, we note that both findings and background relations tended to be at the beginning of the text and so no prior context exists, explaining why these relations are in the losing outcome.

To dive deeper into what might potentially explain the difference between winning and losing outcomes, we examined the first word of the second argument, checking for a match against a list of known discourse connectives.[14] Here, we make

[14]This list was based connectives from PDTB data

Condition	Dataset	Relation	Example
winning	scidtb	comparison	**Context**: We propose a novel method of jointly embedding entities and words into the same continuous vector space . **Arg1**: that jointly embedding brings promising improvement in the accuracy of predicting facts , **Arg2**: compared to separately embedding knowledge graphs and text .
losing	scidtb	result	**Context**: We describe a search algorithm **Arg1**: Our results show **Arg2**: parsing results significantly improve
winning	covdtb	comparison	**Context**: Herein we discuss application of the Collaborative Cross (CC) panel of recombinant inbred strains **Arg1**: Although the focus of this chapter is on viral pathogenesis , **Arg2**: many of the methods are applicable to studies of other pathogens , as well as to case-control designs in genetically diverse populations .
losing	covdtb	findings	**Context**: ROOT (no context) **Arg1**: In this work , we demonstrate a design of meta - holography **Arg2**: that can achieve 2 28 different holographic frames and an extremely high frame rate in the visible range .

Table 9: Examples of discourse relations in the winning and losing outcomes for both the SciDTB and CovDTB datasets.

Relation Category	Percentage of connective matches	
	CovDTB	SciDTB
Losing Relations	7.8%	25.0%
Winning Relations	**16.6%**	**28.2%**

Table 10: Percentage of matches to a list of explicit connectives across the positive, neutral and negative relations.

the simplifying assumption that the discourse connective is found between the two arguments.

Table 10 shows the percentage of instances where, for relations in either the winning or losing outcome, the first word of the second argument was a known discourse connective. We observe that, for both datasets, winning outcomes exhibit a higher percentage of matches for connectives. We take the matches as a potential indicator of the higher proportion of explicitly marked discourse relations. This raises the potential hypothesis that perhaps context may be more beneficial for DRC of certain explicitly marked relations.

7 Future Work

Our preliminary investigation here on the role of context for DRC in scientific writing highlights two potential avenues for future research. Our error analysis suggests that structured context may potentially be more beneficial the DRC for certain explicit relations (for scientific writing). We intended to further investigate this.

We note that our investigation here is limited to dependency discourse structures and the representation of context as string concatenations. In subsequent work, we aim to explore different automatically inferred graph representations of text structure, particularly longer text documents.

Our experiments were also limited to two categories of LLM-based inference, namely In-Context Learning (ICL) for closed and open weight LLMs (or proprietary and so-called "open-source" LLMs). In the future, we intend to include LLMs that include some reasoning capability, such as the recent GPT-o1 and DeepSeek models, as well as techniques like chain of thought, to see if these inference methods help with DRC. In this work, we also used one example of a transformer network for PLM fine-tuning. In future work, we aim to experiment with the model of (Gessler et al., 2021) as an alternative competing transformer model.

Finally, returning to our motivating example, we intend to examine the role of discourse relations in identifying relevant supporting source material to validate generative AI output. We intend to conduct qualitative and quantitative user studies to better understand the potential for discourse information to help with these goals.

8 Conclusions

In this work, we showed that adding discourse context, particularly *structured context*, helps with Discourse Relation Classification for scientific writing. We demonstrated this using two dominant neural language modelling methods: finetuning using a pre-trained language model, and inference with large language models using in-context learning.

sets (2 and 3) and the collated connectives from the DiscoGEM dataset (Scholman et al., 2022). URL: https://github.com/merelscholman/DiscoGeM/tree/main/Appendix

The analysis presented here focuses on two scientific discourse datasets, CovDTB and SciDTB, representing biomedical and computer science disciplines. We found that, for the science discourse relations represented in these datasets, context might help for specific relations, such as with *elaboration*, *attribution*, *comparison* and *temporal* relations.

Acknowledgements

This work was funded by the CSIRO Julius Career Award. We are also grateful to the Heidelberg Institute for Theoretical Studies (H-ITS) for supporting this project and providing facilities for conducting this research. We would like to further acknowledge the the feedback from the CSIRO Language Technology team, the H-ITS NLP team, and the anonymous reviewers on previous versions of this paper.

Limitations

In this work, we focused on English prose language data from publicly available datasets. As such, our conclusions about discourse relations, connectives and the need for using context for discourse relation classification are limited to this language and the genres represented. We note that while we are interested in scientific writing in general, here we study data from just two science disciplines: computer science and biomedical articles about Covid. We note that we only generated a single set of results using prompt-based methods (with Llama 3.1 and GPT 4), using a temperature of zero, due to costs. It is possible that multiple trials of the approach may yield different results. Finally, we note that prompt engineering was limited. It may be possible that stronger performance gains may be obtained if further prompt engineering is employed. For further limitations, see our future work section.

Ethical Considerations

In this work, we use publicly available discourse-related datasets. Our analysis is focused discourse-related linguistic phenomena and is not focused on any individual or subgroup in the community. The work, while motivated by current trends in applied AI, is not immediately applicable in real-world usage.

References

Pablo Accuosto and Horacio Saggion. 2020. Mining arguments in scientific abstracts with discourse-level embeddings. *Data and Knowledge Engineering*, 129.

Katherine Atwell, Remi Choi, Junyi Jessy Li, and Malihe Alikhani. 2022. The role of context and uncertainty in shallow discourse parsing. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 797–811, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Katherine Atwell, Junyi Jessy Li, and Malihe Alikhani. 2021. Where Are We in Discourse Relation Recognition? Technical report.

Tal August, Lauren Kim, Katharina Reinecke, and Noah A. Smith. 2020. Writing strategies for science communication: Data and computational analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5327–5344, Online. Association for Computational Linguistics.

Arne Binder, Leonhard Hennig, and Bhuvanesh Verma. 2022. Full-text argumentation mining on scientific publications. In *Proceedings of the first Workshop on Information Extraction from Scientific Publications*, pages 54–66, Online. Association for Computational Linguistics.

Necva Bölücü, Vincent Nguyen, Roelien C Timmer, Huichen Yang, Maciej Rybinski, Stephen Wan, and Sarvnaz Karimi. 2024. CSIRO-LT at Context24: Contextualising Scientific Figures and Tables in Scientific Literature. Technical report.

Chloé Braud, Yang Janet Liu, Eleni Metheniti, Philippe Muller, Laura Rivière, Attapol Rutherford, and Amir Zeldes. 2023. The DISRPT 2023 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 1–21, Toronto, Canada. The Association for Computational Linguistics.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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. *CoRR*, abs/2005.1.

Chu Sern Joel Chan, Aakanksha Naik, Matthew Akamatsu, Hanna Bekele, Erin Bransom, Ian Campbell, and Jenna Sparks. 2024. Overview of the Context24 Shared Task on Contextualizing Scientific Claims.

In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pages 12–21, Bangkok, Thailand. Association for Computational Linguistics.

Zeyu Dai and Ruihong Huang. 2018. Improving implicit discourse relation classification by modeling inter-dependencies of discourse units in a paragraph. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 141–151, New Orleans, Louisiana. Association for Computational Linguistics.

Aaron Grattafiori et al. 2024. The llama 3 herd of models.

Biaoyan Fang, Xiang Dai, and Sarvnaz Karimi. 2024. Understanding faithfulness and reasoning of large language models on plain biomedical summaries. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9890–9911, Miami, Florida, USA. Association for Computational Linguistics.

Aris Fergadis, Dimitris Pappas, Antonia Karamolegkou, and Haris Papageorgiou. 2021. Argumentation mining in scientific literature for sustainable development. In *Proceedings of the 8th Workshop on Argument Mining*, pages 100–111, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Luke Gessler, Shabnam Behzad, Yang Janet Liu, Siyao Peng, Yilun Zhu, and Amir Zeldes. 2021. DisCoDisCo at the DISRPT2021 shared task: A system for discourse segmentation, classification, and connective detection. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 51–62, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sungho Jeon and Michael Strube. 2020. Centering-based Neural Coherence Modeling with Hierarchical Discourse Segments. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7458–7472, Online. Association for Computational Linguistics.

Anna Kiepura, Yingqiang Gao, Jessica Lam, Nianlong Gu, and Richard H.r. Hahnloser. 2024. SciPara: A new dataset for investigating paragraph discourse structure in scientific papers. In *Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024)*, pages 12–26, St. Julians, Malta. Association for Computational Linguistics.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Wei Liu and Michael Strube. 2023a. 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. 2023b. Annotation-Inspired Implicit Discourse Relation Classification with Auxiliary Discourse Connective Generation.

Wei Liu, Fan Yi, and Michael Strube. 2023. HITS at DISRPT 2023: Discourse Segmentation, Connective Detection, and Relation Classification. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, (Disrpt):43–49.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: {A} Robustly Optimized {BERT} Pretraining Approach. *CoRR*, abs/1907.1.

Wanqiu Long and Bonnie Webber. 2022. Facilitating Contrastive Learning of Discourse Relational Senses by Exploiting the Hierarchy of Sense Relations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10704–10716, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Annie Louis and Ani Nenkova. 2013. What makes writing great? first experiments on article quality prediction in the science journalism domain. *Transactions of the Association for Computational Linguistics*, 1:341–352.

Eleni Metheniti, Chloé Braud, and Philippe Muller. 2024. Feature-augmented model for multilingual discourse relation classification. In *Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024)*, pages 91–104, St. Julians, Malta. Association for Computational Linguistics.

Noriki Nishida and Yuji Matsumoto. 2021. Out-of-Domain Discourse Dependency Parsing via Bootstrapping: {A}n Empirical Analysis on its Effectiveness and Limitation. *Transactions of the Association for Computational Linguistics*.

OpenAI. 2024. Gpt-4 technical report.

José R Penadés, Juraj Gottweis, Lingchen He, Jonasz B Patkowski, Alexander Shurick, Wei-Hung Weng, Tao Tu, Anil Palepu, Artiom Myaskovsky, Annalisa Pawlosky, Vivek Natarajan, Alan Karthikesalingam, and Tiago R D Costa. 2025. AI mirrors experimental science to uncover a novel mechanism of gene transfer crucial to bacterial evolution.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 683–691, 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. Technical report, Proceedings ofthe Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco.

Merel Scholman and Vera Demberg. 2017. Crowdsourcing discourse interpretations: On the influence of context and the reliability of a connective insertion task. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 24–33, Valencia, Spain. Association for Computational Linguistics.

Merel C J Scholman, Tianai Dong, Frances Yung, and Vera Demberg. 2022. DiscoGeM: A Crowdsourced Corpus of Genre-Mixed Implicit Discourse Relations. In *Proceedings of the Thirteenth International Conference on Language Resources and Evaluation ({LREC}'22)*, Marseille, France. European Language Resources Association (ELRA).

Wei Shi and Vera Demberg. 2019. Next sentence prediction helps implicit discourse relation classification within and across domains. 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 5790–5796, Hong Kong, China. Association for Computational Linguistics.

Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2009. Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1493–1502, Singapore. Association for Computational Linguistics.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 103–110, Sydney, Australia. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

Stephen Wan, Cécile Paris, and Robert Dale. 2009. Whetting the appetite of scientists: producing summaries tailored to the citation context. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '09, pages 59–68, New York, NY, USA. Association for Computing Machinery.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The Penn Discourse Treebank 3.0 Annotation Manual. Technical report.

Frank Wilcoxon. 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80–83.

An Yang and Sujian Li. 2018. SciDTB: Discourse dependency TreeBank for scientific abstracts. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 444–449, Melbourne, Australia. Association for Computational Linguistics.

Amir Zeldes, Yang Janet Liu, Mikel Iruskieta, Philippe Muller, Chloé Braud, and Sonia Badene. 2021. The DISRPT 2021 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 1–12, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yingxue Zhang, Fandong Meng, Peng Li, Ping Jian, and Jie Zhou. 2021. Context tracking network: Graph-based context modeling for implicit discourse relation recognition. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1592–1599, Online. Association for Computational Linguistics.

Hao Zhou, Man Lan, Yuanbin Wu, Yuefeng Chen, and Meirong Ma. 2022. Prompt-based Connective Prediction Method for Fine-grained Implicit Discourse Relation Recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3848–3858, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Appendix: Computing Environment

Experiments for training and evaluating the ConnRel model (RoBERTa-based, 82M parameters) were conducted on a server with 1 node (4x NVIDIA A40; 2x Intel(R) Xeon(R) Gold 6330 CPU @ 2.00GHz; 32GB RAM). Each of the 3 approaches tested was trained and evaluated with 5 datasets, over 10 trials. Each trial ranged from between 30 minutes to 1.5 hours, depending on the dataset. Estimated GPU time per approach is 36 hours. Experiments were also repeated at least twice to test for replicability. This results in approximately, 432 hours of GPU time (single jobs).