Clarifying Underspecified Discourse Relations in Instructional Texts

Anonymous ACL submission

Abstract

Discourse relations contribute to the structure of a text and can optionally be realized through explicit connectives such as 'but' and 'while'. But when are these connectives necessary to avoid possible misunderstandings? We investigate this question by first building a corpus of 4,274 text revisions in each of which a connective was explicitly inserted. For a subset of 250 cases, we collect plausibility annotations on other connectives to check whether they would represent suitable alternative relations. The results of this annotation show that several relations are often perceived as plausible in our data. Furthermore, we analyze the extent to which large language models can identify instances with multiple plausible relations as a possible source of misunderstandings. We find that the models predict plausibility of individual connectives with up to 66% accuracy, but they are not reliable in estimating when multiple relations are plausible.

1 Introduction

Discourse relations play a crucial role in establishing coherence and logical flow between discourse segments in natural language. This role is of particular importance in instructional texts, such as how-to guides, where a lack of coherence and clarity can cause instructions to be misinterpreted (Roth et al., 2022). That is, while discourse relations in general can often be inferred by readers implicitly based on context and prior knowledge, instructions for a new task or unknown domain may require the connection between steps to be explicit to avoid confusion (e.g. when steps can be carried out in any order versus only in a specific sequence). However, it has not been fully explored when the introduction of a connective (e.g. 'meanwhile', 'afterwards') alters the perceived plausibility of discourse relations.

As a starting point for studying when an implicit or explicit relation affects plausibility, we focus on revisions of instructional texts involving the explicit insertion of discourse connectives (i.e., *explicitation* of discourse relations), which we consider a form of clarification. Specifically, we utilize wikiHow guidelines,[1] which have similarly been used in past studies to investigate other types of clarifications, such as the strengthening of arguments, resolution of vagueness and specifications of references (Afrin and Litman, 2018; Debnath and Roth, 2021; Anthonio and Roth, 2021).

How to Become a Registered Nurse			
(...) Obtain a bachelor's degree in nursing. * Programs typically take four years to complete, and vary in cost depending on which institution you choose. __$\emptyset$__ Bachelor's programs usually include more training in social sciences than other nursing programs.			
✓	However	✗	At the same time
✓	For example	✗	Thus
✓	In addition		

Table 1: Simplified example from our dataset. The top shows the title of a wikiHow guide, followed by a step name and description. In the revised version, the connective 'However' was inserted in place of __$\emptyset$__. Other connectives shown at the bottom are automatically generated, annotated as plausible (✓) or implausible (✗).

In this work, we present a novel corpus of instructional texts where explicit discourse connectives are inserted at the beginning of sentences in the revisions. An example of a pair of original and revised sentence, including one sentence from the preceding context, is shown in Example 1 below:

(1) Attending. . . meetings may not always be fun.

Original: It can improve your relationship and status with everyone at the workplace.

[1] Available under a CC BY-NC-SA 3.0 license.

Revised: **But** it can improve your relationship and status with everyone at the workplace.

In total, our dataset contains 4,274 instances, making it a substantial resource of discourse relation explicitation. Our primary focus is on cases where the absence of explicit connectives could be a potential source of misunderstanding, which we investigate by examining situations where multiple discourse relations are perceived as plausible for the same arguments. To explore this, we conduct a crowdsourced study to collect plausibility ratings for various discourse connectives, each representing a different discourse relation, within the same context. The dataset includes both naturally occurring connectives from the revisions and synthetically generated alternatives, enabling us to investigate the plausibilities of multiple interpretations for the same arguments (for an example, see Table 1). By gathering independent human ratings for each option, our corpus supports linguistic analysis of underspecified discourse relations and provides a valuable resource for evaluating machine learning models that extend beyond single-class prediction.

We conduct further analysis on the data, addressing the following research questions:

RQ1 How frequently are multiple relations perceived as plausible for the same context in our data? On the other hand, how commonly is insertion of a connective redundant or unnecessary?

RQ2 What are examples of different plausible discourse relations that may conflict or co-exist?

RQ3 Can large language models predict when multiple relations are plausible, indicating potential sources of misunderstanding?

2 Background

As a general framework for (shallow) discourse relations, we rely on the Penn Discourse Tree Bank (PDTB),[2] which provides a large annotated corpus (Prasad et al., 2018), composed of news texts, annotated for discourse relations between text segments.

In its latest version, PDTB-3, the framework employs a three-level hierarchy for the semantic categorization of relations (i.e., sense labels). At the top level of the hierarchy is the 'class' label, which distinguishes between Expansion, Comparison, Contingency, and Temporal relations. Levels 2 and 3 further refine the class semantics, with level 3 encoding directionality (e.g., *Temporal.Synchronous.Precedence*[3]) and appearing only with asymmetric level-2 relations. In total, the Penn Discourse Treebank hierarchy contains 36 fine-grained categories (full set listed in Appendix A).

PDTB-style annotations are performed on the level of semantically related arguments (Arg1 and Arg2), which are typically adjacent text segments. A discourse relation can be constructed through explicitly expressed discourse connectives (explicit relation) or inferred implicitly (implicit relation).

Annotating discourse relations is a challenging task, particularly because relations tend to be underspecified (Rohde et al., 2016; Webber et al., 2018; Scholman and Demberg, 2017; Scholman et al., 2022). While even explicit relations can have multiple readings (Pitler and Nenkova, 2009), implicit relations in particular are often interpreted in different ways (Rohde et al., 2016; Scholman et al., 2022), as reflected in low inter-annotator agreement compared to explicit relations (Zeyrek and Kurfalı, 2017; Hoek et al., 2021; Aktas and Özmen, 2024).

The ambiguity of implicit relations is also evident in the automatic classification of implicit discourse relations, as demonstrated by the performance gap between parsers handling explicit discourse relation recognition and implicit discourse relation recognition (Lin et al., 2014; Varia et al., 2019). Despite recent advances, classifying implicit relations remains a challenging task, especially for 2nd and 3rd level senses in the PDTB (Long and Webber, 2022; Chan et al., 2023). Recent studies suggest that applying modern prompting methods on large language models provides only marginal improvements in discourse parsing performance (Chan et al., 2024; Yung et al., 2024a).

Both manual annotation processes and computational discourse parsing studies indicate that explicit relations are easier to parse than implicit ones. Liu et al. (2024) demonstrate that removing explicit relations from texts often leads to a change in the perceived sense of the relation between arguments (referred to as label shift), highlighting that connectives are generally not redundant. Building on this, we investigate connective insertions and examine how the presence or absence of a connective

[2]The term "shallow" comes from the fact that in contrast to approaches like RST (Mann and Thompson, 1988) or SDRT (Asher and Lascarides, 2003), PDTB framework does not make assumptions about the overall discourse structure. As a result, PDTB relations are not organized hierarchically.

[3]Conventionally, levels are separated by dots.

impacts plausibility.

3 Related Work

Explicitation of discourse relations has been widely studied in the context of Translation studies, particularly related to the Explicitation Hypothesis, which suggests that translations tend to be more explicit than their source texts (Blum-Kulka, 1986). Numerous studies have examined parallel texts to explore this hypothesis (e.g., Zufferey and Cartoni, 2014; Crible et al., 2019; Lapshinova-Koltunski et al., 2022), focusing largely on the insertion or omission of the connective in the translations. Yung et al. (2023) explore another form of explicitation, where a more specific connective is used in translation (e.g., translating "and" as "außerdem" in German), providing further evidence for the Explicitation Hypothesis.

In a study related to ours, Rohde et al. (2016) examine the interpretation of discourse relations and, through a crowdsourcing study framed as a connective insertion task, show that explicit markers and inferred conjunctions can coexist. This challenges the assumption that discourse relations are either explicit or inferred. In another crowdsourcing study, Yung et al. (2019) introduce a two-step method where workers first insert and then disambiguate connectives to annotate discourse relations, a method used for the DiscoGEM corpus (Scholman et al., 2022). Yung et al. (2024b) later refine this into a one-step procedure for annotating the DiscoGEM 2.0 corpus across multiple languages.

In the PDTB, the annotation of implicit relations involves first inserting a connective between the arguments, followed by labeling the relation's sense (Prasad et al., 2008). Building on this approach, several works show that generating discourse connectives between the arguments of implicit relations enhances classification (Shi and Demberg, 2019; Zhou et al., 2022; Xiang et al., 2022; Liu and Strube, 2023; Wang et al., 2023; Wu et al., 2023). Our dataset, which includes numerous instances of connective insertion, can support future research on the automatic recognition of implicit relations, in particular with regard to settings where multiple relations are plausible.

4 Data Collection

The goal of this work is to analyze to what extent implicit discourse relations may require clarification in order to avoid potential of misunderstandings. As a foundation for this analysis, we construct a data set of discourse relation explicitations through connective insertion and plausibility judgments. We proceed in three steps: first, we extract 4,274 instances of connective insertions from wikiHow by comparing different versions of the same article (§4.1); second, we generate alternative connectives indicating different relations that may also be plausible (§4.2); and third, we collect annotations from human judges on which individual connectives are perceived as plausible (§4.3). Finally, we provide statistics on the dataset (§4.4).

4.1 Extraction of Connective Insertions

As a starting point for our data, we make use of revision histories of articles in wikiHow. Using the existing wikiHowToImprove dataset (Anthonio et al., 2020), which contains revisions on the sentence level, we identify cases where a connective is inserted in the revised version of a sentence.

We define an inventory of discourse connectives by first extracting annotated instances from the PDTB corpus and compiling a list of the 100 most frequent connectives.[4] Connectives are known to be ambiguous in both their syntactic and semantic roles (Webber et al., 2019). For instance, words such as "and" can be ambiguous, functioning as a discourse connective. In Example 2, the first occurrence of 'And' is a discourse usage, whereas the second 'and' is a non-discourse usage.

(2) But don't stress make this a week project. **And** keep a dairy, to write all your feelings **and** thoughts.[5]

To avoid the complexity of connective disambiguation, we focus on instances where the inserted tokens is most likely functioning as discourse connectives. Specifically, we select cases where a connective was added at the beginning of a sentence, with no other changes made between the original and revised versions.[6] We provide statistics on the inserted connectives along with a brief discussion of what the distribution reveals in Appendix B.

[4] While genres such as social media platforms (e.g., Twitter, now X) may feature connective forms not present in the PDTB corpus, such as abbreviations or slang (e.g., 'coz', 'cos', 'cus', and 'bc' for "because" or 'b4' for "before" (Aktas and Özmen, 2024)), we assume these forms are rare in our dataset due to the edited nature of the content. Verifying this assumption could be an interesting topic for future research.

[5] All examples are taken from the wikiHowToImprove corpus, unless specified otherwise.

[6] In some instances, a comma was also added after the connective, and these cases are included as well.

4.2 Alternatives Generation

We employ two different methods to generate alternative connectives, both using the *transformers* library (Wolf et al., 2020), without additional pre-training. First, we frame the selection as a cloze task. For this method, we combine the revised sentence with the surrounding context and mask the connective at the beginning of the revision. Then a bidirectional masked language model (MLM), BERT-base-cased, predicts the masked token. From the first 50 predictions generated by the MLM, we extract those that matched the connective list from the PDTB and select the top two alternatives that differ from the original connective and signal semantically different relations. Since BERT predicts one token at a time, this method only produced single-token connectives.

We use an auto-regressive model (GPT-2) to handle multi-word connectives (e.g., "in other words," "for example"). The two best alternatives based on the perplexity scores are selected, again ensuring that they (semantically) differ from both the original connective and the MLM-based alternatives.[7]

4.3 Plausibility Annotation

The wikiHowToImprove data contains articles separated into train, development and test splits.[8] Using the extraction approach outlined before, we identify a total of 4,274 cases of discourse connective insertion, with 3,457 in the training set, 409 in the development set, and 408 in the test set. For a subset of 250 cases (125 from the development and test set each), we generate four alternatives so that there are five variants: one variant with the original connective, two with the connectives predicted by a masked language model and two based on scores from the autoregressive model. For each variant (a total of 1250 instances), we collect annotations in the form of plausibility judgments by four crowdworkers using Amazon Mechanical Turk. Following previous work on plausibility (Anthonio et al., 2022), participants are asked to rate, on a scale from 1 to 5, how well an inserted part of text fits within the context.

Context In PDTB, implicit relations are annotated only for adjacent sentences, and 91% of explicit relation arguments occur within the same or preceding sentence (Prasad et al., 2008). Bourgonje (2021) notes that only 2% of their PDTB-style annotations span more than two sentences. Therefore, we include the previous two sentences as context, either from the same or preceding paragraph (if two previous sentences are not available in the same paragraph). For paragraphs longer than three sentences, we also include the first sentence. Additionally, we include the following sentence from the same paragraph (if available), along with the article name and relevant section, to ensure sufficient context. The interface used in our Human Intelligence Tasks (HITs) is shown in the Appendix C. We pay $0.25 per HIT to ensure participants earn at least the minimum wage per hour.

Qualifications We implement several criteria in order to enhance the quality of the annotations. First, we only allow individuals located in the United States or the United Kingdom to increase the likelihood of selecting native English speakers. Second, participants must have a HIT approval rate of at least 98% and a minimum of 5,000 approved HITs. Finally, crowdworkers are required to pass a qualification test consisting of 10 questions, where they judge a set of clearly plausible and implausible cases that were selected by the authors from the wikiHow data. These 10 questions are also embedded within the actual HITs to monitor participants' attention during the task. Any submissions from participants who answer less than 75% of these attention questions correctly are filtered out.

Class labels We map continuous scores to class labels by first applying two thresholds to the individual plausibility judgments (ranging from 1 to 5) and then aggregating the annotations using majority voting. Specifically, plausibility judgments with a score of ≤ 2.5 are mapped to the label *implausible*, those with a score of ≥ 4.0 to *plausible*, and scores between these thresholds are mapped to the label *neutral*.

4.4 Data Statistics

We collected a total of 5,136 annotations for the 1250 instances in our data (625 development and

[7]Note that a single connective might correspond to multiple relations (e.g., *since*). However, discourse relation labels (e.g., *Comparison.Contrast*) are challenging for untrained crowd workers to annotate. Therefore, we follow earlier studies (e.g., Wu et al. (2023); Wang et al. (2023)) and build an inventory of relatively unambiguous connectives. Specifically, combine data from (Yung et al., 2024b) and additional connectives found in our data (e.g., "in addition").

[8]https://github.com/irshadbhat/wikiHowToImprove

[9]Due to a miscalculation in the experimental process, we ended up collecting more than four judgments for some tasks.

	Dev		Test	
Plausible	348	(55.7%)	297	(47.5%)
Implausible	249	(39.8%)	291	(46.5%)
Neutral	28	(4.5%)	37	(6.0%)
Total count	625		625	

Table 2: Distribution of plausibility judgments, based on majority aggregation of scores mapped to categories.

	Dev	Test	Σ / %
Contexts	125	125	250
1 plausible connective	15	21	14%
— Only revision plausible	7	10	7%
≥2 plausible connectives	108	95	81%
Revision not plausible	26	32	23%

Table 4: Overview of the plausibility judgments

	Dev	Test
Instances	625	625
Annotations	2501	2635[9]
Averaged agreement	53.6%	54.7%
Majority agreement	71.3%	70.7%
Averaged agr. (with MACE)	55.2%	57.9%
Majority agr. (with MACE)	72.4%	73.0%

Table 3: Inter-Annotator Agreement Statistics

		Dev	Test
Contexts with one token masked		98	101
MLM / BERT	Recall@1	34%	36%
	Recall@5	74%	67%
	Recall@50	98%	98%
GPT-2	Recall@1	8%	7%
	Recall@5	71%	74%

Table 5: Performance of top model outputs in generating the connective inserted during revision.

625 test instances). On a scale from 1 to 5, the average plausibility judgment for the original connectives is 3.98, whereas this value is 2.98 for the alternative connectives we generated. The average plausibility is higher for MLM-predicted connectives (3.18) compared to those highest ranked by GPT-2 (2.77). In Table 2, we show the distribution over the class labels *plausible*, *implausible* and *neutral* over all annotated connectives.

The class labels are based on mapping each individual annotation to a class and then aggregating the labels based on majority vote. As shown in Table 3, agreement based on majority-aggregated values is relatively high, namely 71.3% and 70.7% for the development and test set, respectively. In comparison, agreement based on average-aggregated values would be lower, namely 53.6% and 54.7%. To assess in how far disagreements reflect different perspectives or potential errors, we employ MACE (Multi-Annotator Competence Estimation) (Hovy et al., 2013) with a threshold of 0.5. In our pilot studies, 7-15% of submissions were flagged as low-quality. However, after integrating attention checks into the process (see §4.3), the number of low-quality submissions dropped significantly, with only an average of 5% being marked as such.

In our final design, MACE identified very few incompetent annotators, which had a minimal impact on overall inter-annotator agreement (see Table 3). As noted in Section 2, discourse relations can have multiple interpretations. Therefore, we believe that these agreement scores, well above chance level for the three label categories, indicate a reasonably high agreement between the crowd workers.

5 Data Analysis

A fundamental question of our work is whether the connective insertions observed in the data are necessary to avoid possible causes of misunderstandings. To answer this question, we examine how frequently different connectives in the same context are perceived as plausible. As shown in Table 4, participants in the crowdsourcing experiment annotated at least two different connectives as plausible in the vast majority of contexts (81%). There are only 36 (14%) contexts, in which a single connective was perceived as plausible. Surprisingly, the plausible connective is the actual insertion made during revision in only around half of these cases (17). In fact, there are 58 cases (23%) in which the connective inserted during revision is not perceived as plausible (i.e., it is judged as *implausible* or *neutral*). We next discuss automatically generated connectives (§5.1) and potentials for misunderstanding (§5.2).

5.1 Generated Connectives

As discussed in the data statistics (§4.4), automatically generated connectives generally re-

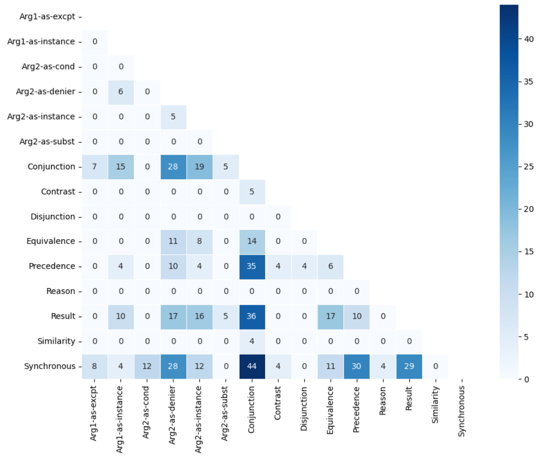

Figure 1: Relation sense co-occurrence matrix (Only the lowest level elements in the sense hierarchy are shown)

ceived lower plausibility scores than the original connective identified from the revision history (2.98 vs. 3.98). However, we restricted the model outputs to connectives that indicate semantically different relations from the original connective.

Table 5 summarizes how often the original connective would have been in the top outputs of BERT and GPT-2. We limit this analysis to one-word connectives as the MLM task always requires one token to be predicted. In the development and test set, 98 and 101 instances, respectively, contain single-token connectives. Regarding the top-1 outputs, BERT predicts a substantially higher number of these connectives than GPT-2 (about 35% vs. 8%). Among the top 5, BERT and GPT-2 achieve about the same performance (67% vs. 74% Recall@5). Finally, we find 98% of the original connectives among BERT's top-50 predictions. We conclude from this analysis that the generated connectives demonstrate a high degree of reliability.

5.2 Multiple Plausible Relations

As discussed in Section 4.2, we selected model-generated connectives such that they signal semantically different relations. In consequence, this means that when two or more connectives are judged as plausible, then multiple discourse relations seem applicable for the same arguments.[10]

The confusion matrix in Figure 1 illustrates which pairs of discourse relations are perceived as plausible in the same contexts, based on the connectives annotated in our dataset.[11] While prior studies have examined concurrent relations among the same arguments, the non-uniform distribution of connectives proposed by language models in our experiment makes direct comparisons challenging. Still, some interesting patterns emerge.

Torabi Asr and Demberg (2013) identify frequent co-occurrences like (*Conjunction&Synchronous*) and (*Synchronous&Result*) in PDTB, which also appear in our dataset. Similarly, Scholman et al. (2022) report (*Conjunction&Result*) and (*Precedence&Result*) as frequent, aligning with our findings. However, co-occurances with *Arg2-as-detail* are common in their dataset but absent from ours, likely due to the rarity of connectives indicating these relations (e.g., "in fact") in our data. A notable difference in our data is the (*Synchronous&Precedence*) co-occurrence, which neither study reports. These variations may stem from genre differences, motivating further investigation.

We note that while some relations can coexist (e.g., *Conjunction* and *Arg2-as-instance*), others are mutually exclusive. For example, synchronous and asynchronous temporal relations cannot occur simultaneously. In contexts where both are perceived as plausible, as commonly observed in Figure 1 between *Synchronous&Precedence* senses, the absence of an explicitly realized relation is likely to result in misunderstandings.

The context in Example (3) provides an actual example from our data, highlighting another potential case of misunderstanding.

(3) If you have a double sink, plug up the other side with a wet rag. **Or**, you can do the same process on both sides with two plungers, with a friend or with both your hands holding a plunger.

As shown, the relation *Expansion.Disjunction* is signaled by the connective 'or' inserted during revision. In our data collection, we found annotators to also mark the connective 'then' as plausible, which denotes a *Temporal.Asynchronous.Precedence* relation. While the *Expansion.Disjunction* relation indicates two mutually exclusive events, the temporal relation implies that the second event must also occur. Therefore, the absence of any connective in

[10]Note that while generated connectives are mostly unambiguous, the original insertions can be an underspecified connective such as "and". Therefore, two connectives may not always signal different relations.

[11]This analysis uses the most frequent relation sense for each connective, given in Appendix E.

this context may lead to confusion about the correct event sequence.

6 Computational Experiments

In Section 4, we discussed the creation of our data and showed that there exist multiple plausible relations for many of the selected contexts (82%). Our analysis in Section 5 has shown that some of the relations perceived as plausible are mutually exclusive, indicating that a lack of discourse relation explicitation could potentially lead to misunderstanding. In the following, we investigate whether more recent large language models (LLMs) can predict which connectives are plausible in a given context[12] and, by extension, whether multiple relations are plausible, signalling a potential source of misunderstanding.

6.1 Setup

The setup for the computational experiments largely follows the setup of our annotation study (§4.3), with some variation in the required output. That is, we prompt models with the article name, section and context as input, and ask for a plausibility judgment on the highlighted connective (marked by underscores) as output. We experiment with three configurations for the output: a **Binary** (*Plausible* or *Implausible*), a **3-way** categorization (*Plausible*, *Implausible*, or *Neutral*), and a five-point **Scale**-based classification (1–5).

We conduct our experiments on the test set of the annotated data to establish a baseline. We evaluate five LLMs in a zero-shot setting: **GPT4o-mini**, **GPT4o** (OpenAI et al., 2024) (via the OpenAI API), **Gemini-1.5-flash**, **Gemini-1.5-Pro** (GeminiTeam et al., 2024) (via the Google API), and **Claude3-haiku**[13] (Anthropic, 2024) (via the Amazon Bedrock API). [14] The temperature was set to 1 for all experiments. We used the development set of our data for selecting prompts, and we used the same prompts for all models on the test set (for examples, see Appendix E).

[12]Note that we also used languge models in the creation of our data. However the applied models, BERT and GPT-2, also produced many implausible connectives (§4.4).

[13]The larger version of Claude (Sonnet) failed to follow the provided instructions.

[14]We attempted to include Llama as a more open model in our evaluation, but we neither got the small nor the large version to adhere to our instructions.

	Binary	**Scale**	**3-way**
Chance baseline	50.0	33.3	33.3
Majority class	50.5	47.5	47.5
GPT4o-mini	62.4	**48.3**	52.5
Gemini-1.5-flash	**64.6**	41.6	**59.0**
Claude3-haiku	59.9	45.6	57.8
GPT4o	61.6	47.7	57.8
Gemini-1.5-pro	**66.2**	**64.8**	**61.4**
Human agreement		70.7	

Table 6: Accuracy scores across different models in the two-way and three-way classification tasks. Bolded values indicate the highest scores within each group.

6.2 Results

Table 6 presents the performance of the LLMs across the three setups. For each, we provide results in terms of accuracy (i.e., the ratio of correct predictions). In the scale-based setup, we map predicted scores to the three classes using the same thresholds as in the annotation (§4.3). For comparison, we report individual human agreement with the aggregated label as well as expected accuracies of a random classifier and a majority class baseline.

Plausibility of connectives For individual plausibility predictions, the LLM performance always lies well above the chance level, but below the human agreement as a upper bound. The models also consistently outperform the majority baseline, except in scale-based setup, where only GPT4o-mini achieves a higher accuracy among the smaller models. All models, except Gemini-1.5-pro, perform better at predicting the class label directly, rather than predicting a score (up to +17.4 percentage points). Overall, the scale-based setup presents greater challenges for all models except Gemini-1.5-pro, which achieves outstanding performance in this setup, surpassing the second-best score by +16.5 points. The experiments reveal that larger models do not always guarantee better performance; for instance, GPT4o-mini outperforms GPT4o in the binary and scale setups. Among all models, Gemini-1.5-pro delivers the best performance across all setups, achieving results close to the human agreement upper bound.

On the binary classification task, all LLMs achieve comparable accuracy scores, between 59.9% and 66.2%. On closer inspection, it is noticeable that the models differ greatly in their errors:

Plausible connectives	≤1	≥2	Accuracy
Annotation	30	95	—
GPT4o-mini	16	109	**80.8%**
Gemini-1.5-flash	34	91	71.2%
Claude3-haiku	18	107	79.2%
GPT4o	17	108	76.8%
Gemini-1.5-pro	61	64	64.0%

Table 7: Counts of contexts with at most one (≤1) or multiple plausible (≥2) connectives according to human annotations and model predictions. The last column indicates accuracy as the overlap between human judgments and model outputs for the two classes ≤1 & ≥2.

While GPT4o models and Claude have a strong preference for the *Plausible* label, which is predicted in about 75% of cases, Gemini's predictions (and errors) are more evenly spread across all labels. In summary, we find that the models show different strengths, and there is no significant advantage of one model over the others when it comes to individual plausibility predictions.

Multiple relations As a starting point for assessing when multiple relations are plausible, we combine plausibility judgments over all connectives for each context. We then check when two or more connectives are plausible according to the human annotation as well as according to the model predictions. The results are summarized in Table 7. Similar to the individual plausibility predictions, we observe that GPT4o models and Claude have a tendency of predicting too many relations as plausible, whereas Gemini predicts too few relations as plausible. Still, the performance on the context level is relatively high, with GPT4o and Claude reaching accuracy scores between 79.2% and 80.8%.

Although these results appear promising in terms of numbers, there are at least two limitations with regard to predicting potential causes of misunderstandings. As already discussed in Section 5.2, only some discourse relations are mutually exclusive, which is why cases with two or more plausible connectives might as well be unproblematic. Still, the relative proportion of contexts in which several connectives are rated as plausible is so high that even the predictions of the best model are hardly better than a baseline that would always predict two or more connectives to be plausible (76%).

7 Discussion

We briefly summarize our findings with respect to the research questions introduced in Section 1.

RQ1 How frequently are multiple relations perceived as plausible for the same context in our data? On the other hand, how commonly is insertion of a connective redundant or unnecessary?

For the vast majority of the arguments, we found in the collected data that two or more connectives are annotated as plausible (82%, see §4.4). In contrast, there are only a few cases where the original connective is perceived as the only plausible option (6%). This indicates that the insertion of connectives as part of a revision is usually not redundant.

RQ2 What are examples of different plausible discourse relations that may conflict or co-exist?

In our analysis, we frequently found multiple discourse relations to be perceived as plausible in the same context (see §5.2). Some of these relations can co-exist (e.g. *Expansion.Conjunction* and *Expansion.Instantiation.Arg2-as-instance*) while others are conflicting (e.g. *Temporal.Precedence* and *Expansion.Disjunction*). Future work should explore in more detail when two or more relations can hold simultaneously, as context may also influence which relations are mutually exclusive.

RQ3 Can large language models predict when multiple relations are plausible, indicating potential sources of misunderstanding?

Our experiments with five LLMs revealed that predicting the plausibility of connectives is possible well above chance level, with Gemini performing best but still below the human agreement upper bound (see §6). When we compare individual plausibility judgments of LLMs with human plausibility annotations, we find that GPT and Claude tend to overpredict plausibility, while Gemini underpredicts it. In consequence, our experiments do not confirm the reliability of LLMs in predicting when multiple relations are plausible in a given context, despite seemingly high accuracy scores.

8 Conclusion

We introduced a dataset of 4,274 text revisions with explicit connective insertions and a subset with plausibility judgments on alternative connectives. While LLMs can predict plausibility with high accuracy, they still leave room for improvement when it comes to multiple or conflicting relations.

Limitations

This study has several limitations. It focuses exclusively on how-to guides from wikiHow, which may not represent other discourse genres, and examines only English-language texts, leaving the behavior of discourse connectives in other languages unexplored. Additionally, the comparison with existing literature is limited, particularly regarding concurrent relations. Unlike other studies that provide a full set of connectives covering all relation senses in PDTB, our experiment did not include such a comprehensive set for plausibility judgments. Future research should conduct controlled experiments, particularly providing uniform relation sense distribution for plausibility judgments, to better understand concurrent relations within this genre. Moreover, the crowdsourced plausibility judgments, while quality-controlled, inherently involve subjectivity, which may affect the consistency of the results.

Another limitation of this study relates to the use of LLMs. While they show potential for predicting plausible relations, they struggle with identifying multiple or conflicting ones. Evaluating the reliability of LLMs in understanding discourse is essential for a comprehensive assessment of their performance on this task. Miao et al. (2024) propose a question-answering-based method to evaluate the "faithfulness" of LLMs, for instance, when discourse relations involving the same arguments shift direction (e.g., result vs. reason). Applying such methods could provide deeper insights into the limitations of LLMs, which we leave for future research.

References

Tazin Afrin and Diane Litman. 2018. Annotation and classification of sentence-level revision improvement. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 240–246, New Orleans, Louisiana. Association for Computational Linguistics.

Berfin Aktas and Burak Özmen. 2024. Shallow discourse parsing on Twitter conversations. In *Proceedings of the 20th Joint ACL - ISO Workshop on Interoperable Semantic Annotation @ LREC-COLING 2024*, pages 60–65, Torino, Italia. ELRA and ICCL.

Talita Anthonio, Irshad Bhat, and Michael Roth. 2020. wikiHowToImprove: A resource and analyses on edits in instructional texts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5721–5729, Marseille, France. European Language Resources Association.

Talita Anthonio and Michael Roth. 2021. Resolving implicit references in instructional texts. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 58–71, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Talita Anthonio, Anna Sauer, and Michael Roth. 2022. Clarifying implicit and underspecified phrases in instructional text. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3319–3330, Marseille, France. European Language Resources Association.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. Technical report, Anthropic.

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press, Cambridge.

Shoshana Blum-Kulka. 1986. Shifts of cohesion and coherence in translation. In Juliane House and Shoshana Blum-Kulka, editors, *Interlingual and Intercultural Communication: Discourse and Cognition in Translation and Second Language Acquisition Studies*, pages 17–35. Gunter Narr Verlag, Tübingen.

Peter Bourgonje. 2021. *Shallow discourse parsing for German*. doctoralthesis, Universität Potsdam.

Chunkit Chan, Cheng Jiayang, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2024. Exploring the potential of ChatGPT on sentence level relations: A focus on temporal, causal, and discourse relations. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 684–721, St. Julian's, Malta. Association for Computational Linguistics.

Chunkit Chan, Xin Liu, Jiayang Cheng, Zihan Li, Yangqiu Song, Ginny Wong, and Simon See. 2023. DiscoPrompt: Path prediction prompt tuning for implicit discourse relation recognition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 35–57, Toronto, Canada. Association for Computational Linguistics.

Ludivine Crible, Ágnes Abuczki, Nijolė Burkšaitienė, Péter Furkó, Anna Nedoluzhko, and Sigita Rackevičienė. 2019. Functions and translations of discourse markers in ted talks: A parallel corpus study of underspecification in five languages. *Journal of Pragmatics*, 142:139–155.

Alok Debnath and Michael Roth. 2021. A computational analysis of vagueness in revisions of instructional texts. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 30–35, Online. Association for Computational Linguistics.

GeminiTeam, Petko Georgiev, Ving Ian Lei, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Jet Hoek, Merel C.J. Scholman, and Ted J.M. Sanders. 2021. Is there less agreement when the discourse is underspecified? In *Proceedings of the DiscAnn Workshop*.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

Ekaterina Lapshinova-Koltunski, Christina Pollkläsener, and Heike Przybyl. 2022. Exploring explicitation and implicitation in parallel interpreting and translation corpora. *The Prague Bulletin of Mathematical Linguistics*, 119:5–22.

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.

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

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

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.

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

Yisong Miao, Hongfu Liu, Wenqiang Lei, Nancy Chen, and Min-Yen Kan. 2024. Discursive socratic questioning: Evaluating the faithfulness of language models' understanding of discourse relations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6277–6295, Bangkok, Thailand. Association for Computational Linguistics.

John D. Murray. 1997. Connectives and narrative text: The role of continuity. *Memory & Cognition*, 25(2):227–236.

OpenAI, Josh Achiam, Steven Adler, et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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

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

Rashmi Prasad, Bonnie Webber, and Alan Lee. 2018. Discourse annotation in the PDTB: The next generation. In *Proceedings of the 14th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, pages 87–97, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Hannah Rohde, Anna Dickinson, Nathan Schneider, Christopher N. L. Clark, Annie Louis, and Bonnie Webber. 2016. Filling in the blanks in understanding discourse adverbials: Consistency, conflict, and context-dependence in a crowdsourced elicitation task. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 49–58, Berlin, Germany. Association for Computational Linguistics.

Michael Roth, Talita Anthonio, and Anna Sauer. 2022. SemEval-2022 task 7: Identifying plausible clarifications of implicit and underspecified phrases in instructional texts. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1039–1049, Seattle, United States. Association for Computational Linguistics.

T.J.M. Sanders. 2005. Coherence, causality and cognitive complexity in discourse. In *Proceedings of the First International Symposium on the Exploration and Modelling of Meaning*, pages 31–46, France. Universite de Toulouse-Le Mirail.

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 Scholman, Tianai Dong, Frances Yung, and Vera Demberg. 2022. DiscoGeM: A crowdsourced corpus of genre-mixed implicit discourse relations. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3281–3290, Marseille, France. European Language Resources Association.

Erwin M. Segal and Judith Felson Duchan. 1991. The role of interclausal connectives in narrative structuring: Evidence from adults' interpretations of simple stories. *Discourse Processes*, 14:24–54.

Wei Shi and Vera Demberg. 2019. Learning to explicitate connectives with Seq2Seq network for implicit discourse relation classification. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 188–199, Gothenburg, Sweden. Association for Computational Linguistics.

Fatemeh Torabi Asr and Vera Demberg. 2012. Implicitness of discourse relations. In *Proceedings of COLING 2012*, pages 2669–2684, Mumbai, India. The COLING 2012 Organizing Committee.

Fatemeh Torabi Asr and Vera Demberg. 2013. On the information conveyed by discourse markers. In *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*, pages 84–93, Sofia, Bulgaria. Association for Computational Linguistics.

Siddharth Varia, Christopher Hidey, and Tuhin Chakrabarty. 2019. Discourse relation prediction: Revisiting word pairs with convolutional networks. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 442–452, Stockholm, Sweden. Association for Computational Linguistics.

Chenxu Wang, Ping Jian, and Mu Huang. 2023. Prompt-based logical semantics enhancement for implicit discourse relation recognition. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 687–699, Singapore. Association for Computational Linguistics.

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

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2018. The Penn Discourse Treebank 3.0 Annotation Manual. Report, The University of Pennsylvania.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Hongyi Wu, Hao Zhou, Man Lan, Yuanbin Wu, and Yadong Zhang. 2023. Connective prediction for implicit discourse relation recognition via knowledge distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5908–5923, Toronto, Canada. Association for Computational Linguistics.

Wei Xiang, Zhenglin Wang, Lu Dai, and Bang Wang. 2022. ConnPrompt: Connective-cloze prompt learning for implicit discourse relation recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 902–911, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Frances Yung, Mansoor Ahmad, Merel Scholman, and Vera Demberg. 2024a. Prompting implicit discourse relation annotation. In *Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 150–165, St. Julians, Malta. Association for Computational Linguistics.

Frances Yung, Vera Demberg, and Merel Scholman. 2019. Crowdsourcing discourse relation annotations by a two-step connective insertion task. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 16–25, Florence, Italy. Association for Computational Linguistics.

Frances Yung, Merel Scholman, Ekaterina Lapshinova-Koltunski, Christina Pollkläsener, and Vera Demberg. 2023. Investigating explicitation of discourse connectives in translation using automatic annotations. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 21–30, Prague, Czechia. Association for Computational Linguistics.

Frances Yung, Merel Scholman, Sarka Zikanova, and Vera Demberg. 2024b. DiscoGeM 2.0: A parallel corpus of English, German, French and Czech implicit discourse relations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4940–4956, Torino, Italia. ELRA and ICCL.

Deniz Zeyrek and Murathan Kurfalı. 2017. TDB 1.1: Extensions on Turkish discourse bank. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 76–81, Valencia, Spain. 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.

Sandrine Zufferey and Bruno Cartoni. 2014. A multifactorial analysis of explicitation in translation. *Target: International Journal of Translation Studies*, 26(3):361–384.

A Appendix A: PDTB-3 Sense Hierarchy

Level-1	Level-2	Level-3
Temporal	Synchronous	–
	Asynchronous	Precedence
		Succession
Contingency	Cause	Reason
		Result
		NegResult
	Cause+Belief	Reason+Belief
		Result+Belief
	Cause+SpeechAct	Reason+SpeechAct
		Result+SpeechAct
	Condition	Arg1-as-cond
		Arg2-as-cond
	Condition+SpeechAct	–
	Negative-condition	Arg1-as-NegCond
		Arg2-as-NegCond
	Negative-condition+SpeechAct	–
	Purpose	Arg1-as-goal
		Arg2-as-goal
Comparison	Concession	Arg1-as-denier
		Arg2-as-denier
	Concession+SpeechAct	Arg2-as-denier+SpeechAct
	Contrast	–
	Similarity	–
Expansion	Conjunction	–
	Disjunction	–
	Equivalence	–
	Exception	Arg1-as-excpt
		Arg2-as-excpt
	Instantiation	Arg1-as-instance
		Arg2-as-instance
	Level-of-detail	Arg1-as-detail
		Arg2-as-detail
	Manner	Arg1-as-manner
		Arg2-as-manner
	Substitution	Arg1-as-subst
		Arg2-as-subst

Figure 2: PDTB-3 sense hierachy (taken from Webber et al. (2018, p.17))

B Appendix B: Frequency Distribution of Inserted Connectives

Table 8 lists the 10 most frequent connectives found in the extracted connective insertion instances. The most commonly inserted discourse connectives are *Then*, *For example/For instance*[15], and *However/But*.

In their analysis, Torabi Asr and Demberg (2012) examine why some relations are made explicit while others remain implicit, suggesting that the "expectedness" of a relation determines its explicitness. Based on the Continuity hypothesis (Segal and Duchan, 1991; Murray, 1997) and the Causality-by-default hypothesis (Sanders, 2005), they argue that continuous relations (e.g., causal relations) are more likely to be implicit, whereas discontinuous relations (e.g., temporal, comparison, and instantiation relations) are often explicit. They also propose that forward-temporal relations tend to be more implicit than other temporal relations.

When we examine our data alongside the connective-sense mappings in Table 9[16], we observe that some of their findings align with our results. For example, comparison relations, signaled by connectives such as "However" and "But" are frequently explicitated, while causal relations, indicated by connectives like "So" and "Because" are rare. Temporal relations, however, reveal an intriguing exception. Although Torabi Asr and Demberg (2012) classify temporal relations as discontinuous and generally explicit, they also suggest that temporal relations following the textual order of arguments naturally tend to be implicit. Yet, in our dataset, the connective "Then" which denotes sequential events, is the most frequently inserted connective. As shown in Figure 1, the same arguments can often represent both temporal and other types of relations (e.g., *Temporal.Asynchronous.Precedence* vs. *Comparison.Concession.Arg2-as-denier*) in our experiments, suggesting that even for events in sequential order, temporal relations may not always be straightforwardly established.

Another notable difference involves *Expansion.Instantiation* relations. Torabi Asr and Demberg (2012) classify these relations as continuous and, therefore, more likely to be implicit. In contrast, our dataset shows that connectives such as "For example" and "For instance", which strongly signal instantiation, are the second most frequent connective insertions. These differences could stem from genre-specific factors, as Torabi Asr and Demberg (2012) analyzed PDTB news data, while our dataset consists of step-by-step procedural instructions. However, it is important to emphasize that these findings are based solely on connective insertions in the revisions at the beginnings of sentences and may not represent the entire dataset. Further exploration of how these hypotheses apply to the complete dataset is reserved for future research.

C Appendix C: Interface for Crowdsourcing Experiment

[15] Grouped together as they strongly indicate the same relation according to Table 9.

[16] This analysis is based on the dominant senses indicated by connectives. A full exploration of relation senses would require parsing the dataset with a discourse parser, which we leave for future work.

Connective	%
Then	20.47
For example	15.53
However	12.95
Also	9.86
But	5.50
Or	4.29
And	4.11
So	4.04
For instance	3.23
If	2.81

Table 8: Top 10 most frequent connective insertions at the beginning of revisions

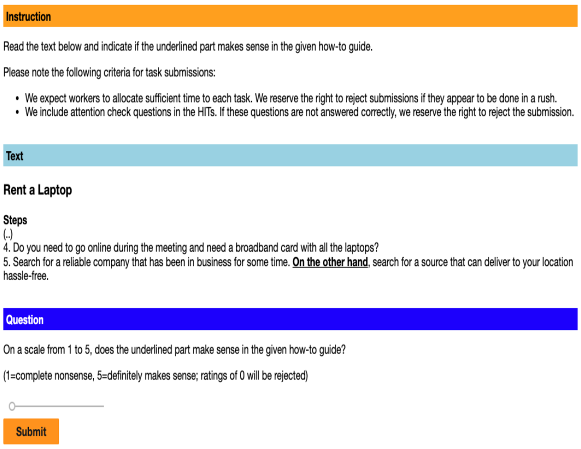

Figure 3: Interface for collecting plausibility judgments

D Appendix D: Prompt examples used in the computational experiments

Binary:

Does the text between the underscores make sense in the given how-to guide?
Please respond only with one word: Plausible or Implausible

Be a Successful Engineer
Being Successful in the Workplace
...
6. Make sure you show up at every meeting. Attending all the meetings may not always be fun. _But_ it can improve your relationship and status with everyone at the workplace. It shows you are really interested at being an engineer.

Three categories:

Does the text between the underscores make sense in the given how-to guide? Please respond only with one word: Plausible, Implausible or Unclear[a]

Be a Successful Engineer
Being Successful in the Workplace
...
6. Make sure you show up at every meeting. Attending all the meetings may not always be fun. _But_ it can improve your relationship and status with everyone at the workplace. It shows you are really interested at being an engineer.

[a] Using "Neutral" or "Undecided" did not yield responses in the third category, but "Unclear" was effective in achieving that.

Scale-based:

On a scale from 1 to 5, does the text between the underscores make sense in the given how-to guide?
(Please respond with only a single number between 1 and 5, where 1=complete nonsense, 5=definitely makes sense) Be a Successful

Engineer
Being Successful in the Workplace
...
6. Make sure you show up at every meeting. Attending all the meetings may not always be fun. _But_ it can improve your relationship and status with everyone at the workplace. It shows you are really interested at being an engineer.

E Appendix E: Connective-Relation Sense Mapping

Connective	Sense(s)
after	Temp.Asynchronous.Succession
also	Exp.Conjunction
although	Comp.Concession.Arg1-as-denier
and	Exp.Conjunction
as	Temp.Synchronous
as a result	Cont.Cause.Result
as if	Exp.Manner.Arg2-as-manner
at the same time	Temp.Synchronous
because	Cont.Cause.Reason
before	Temp.Asynchronous.Precedence
but	Comp.Concession.Arg2-as-denier
consequently	Cont.Cause.Result
even though	Comp.Concession.Arg1-as-denier
finally	Temporal.Asynchronous.Precedence
for example/instance	Exp.Instantiation.Arg2-as-instance
for that purpose	Cont.Purpose.Arg1-as-goal
furthermore	Exp.Conjunction
however	Comp.Concession.Arg2-as-denier
if	Cont.Condition.Arg2-as-cond
if not	Cont.Neg-condition.Arg1-as-negCond
in addition	Exp.Conjunction
in fact	Exp.Level-of-detail.Arg2-as-detail
in more detail	Exp.Level-of-detail.Arg2-as-detail
in other words	Exp.Equivalence
in short	Exp.Level-of-detail.Arg1-as-detail
instead	Exp.Substitution.Arg2-as-subst
meanwhile	Temp.Synchronous
moreover	Exp.Conjunction
nevertheless	Comp.Concession.Arg2-as-denier
nonetheless	Comp.Concession.Arg2-as-denier
on the other hand	Comp.Contrast
once	Temp.Asynchronous.Succession
or	Exp.Disjunction
other than that	Exp.Exception.Arg1-as-excpt
otherwise	Exp.Exception.Arg1-as-excpt
rather than	Exp.Substitution.Arg1-as-subst
similarly	Comp.Similarity
so	Cont.Cause.Result
so that	Cont.Purpose.Arg2-as-goal
specifically	Exp.Level-of-detail.Arg2-as-detail
thereby	Exp.Manner.Arg1-as-manner
therefore	Cont.Cause.Result
then	Temp.Asynchronous.Precedence
this illustrates that	Exp.Instantiation.Arg1-as-instance
though	Comp.Concession.Arg1-as-denier
thus	Cont.Cause.Result
unless	Cont.Neg-condition.Arg2-as-negCond
until	Temp.Asynchronous.Precedence
when	Temp.Synchronous
while	Temp.Synchronous

Table 9: Connectives and the main relation senses they signal (Exp: Expansion, Comp: Comparison, Cont: Contingency, Temp: Temporal)