Information-Theoretic and Prompt-Based Evaluation of Discourse Connective Edits in Instructional Text Revisions

Berfin Aktaş
Natural Language Understanding Lab
University of Technology Nuremberg
berfin.aktas@utn.de

Michael Roth
Natural Language Understanding Lab
University of Technology Nuremberg
michael.roth@utn.de

Abstract

We present a dataset of text revisions involving the deletion or replacement of discourse connectives. Manual annotation of a replacement subset reveals that only 19% of edits were judged either necessary or should be left unchanged, with the rest appearing optional. Surprisal metrics from `GPT-2` token probabilities and prompt-based predictions from `GPT-4.1` correlate with these judgments, particularly in such clear cases.

1 Introduction

Discourse relations are essential for maintaining coherence and logical flow in text. This is especially critical in instructional texts such as how-to guides, where a lack of clarity can lead to misinterpretation and misunderstanding (Roth et al., 2022; Aktas and Roth, 2025). Discourse relations are often signaled explicitly through discourse connectives like "because" or "however." In the case of implicit discourse relations, where coherence is inferred from context rather without an overt connective, interpretation becomes more ambiguous. Even with explicit connectives, alternative connectives may sometimes express a relation more clearly or appropriately. Identifying when a connective is truly necessary, when it may be redundant or misleading, and which connective best fits the context remains a challenge in discourse processing.

We investigate revisions in which discourse connectives are inserted, replaced or deleted, and explore whether LLMs (`GPT-2`, `GPT-4.1`, `Mistral-7B`), can offer useful cues. Specifically, we explore whether an information-theoretic measure such as Shannon (1948)'s *surprisal* can be used to assess the necessity or appropriateness of a discourse connective in context. Our hypothesis is that connectives inserted during revision (i.e., judged necessary) and those deleted (i.e., judged unnecessary) affect the predictability of the text in different ways and and these effects can be captured through surprisal-based measures such as *average surprisal*, *variance*, and *smoothness* of surprisal change.

An initial analysis of information-theoretic measures over *all* revisions regarding connectives revealed inconsistent patterns: the presence or absence of a discourse connective did not consistently affect surprisal-based metrics (see Appendix B). To further investigate this, we conduct a qualitative study on a subset of connective replacements, which we present in Section 4. Manual annotation of this subset reveals that not all replacements serve the same function: some are optional, others are essential for coherence, and some may even be inappropriate. We apply two complementary approaches to uncover underlying patterns computationally: information-theoretic analysis using `GPT-2` (Radford et al., 2019) and prompting-based evaluation with more recent models, namely `GPT-4.1` (OpenAI et al., 2024) and `Mistral-7B` (Jiang et al., 2023). We discuss the patterns uncovered by these methods in Section 5, arguing that comparing these approaches provides valuable insights into the interpretability of model behavior.

Although our exploration of information-theoretic measures with language models yields largely inconclusive results, we believe this line of investigation remains promising. As language models become more transparent in exposing token-level probabilities and better at modeling discourse, information-theoretic metrics may offer an interpretable framework for assessing connective necessity and discourse coherence. We view this as an interesting direction for future research.

We investigate the following research questions:

- RQ1: Can information-theoretic metrics reveal when replacing a connective improves coherence?
- RQ2: To what extent do information-theoretic

Proceedings of the 6th Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences (CODI 2025), pages 228–236

and prompt-based methods align in their interpretation of discourse relations?

2 Related Work

Information *surprisal* has been extensively used in computational linguistics to model processing difficulty and expectations in language comprehension (e.g., Levy, 2008; Clark et al., 2023; Oh and Schuler, 2023). From a discourse perspective, Torabi Asr and Demberg (2015) examine the role of discourse connectives through the lens of the Uniform Information Density (UID) hypothesis. They show that connectives can help distribute information more evenly.

Recently, Aktas and Roth (2025) examine discourse connective insertions in revision data and find that multiple relations are often plausible, while language models perform inconsistently in detecting ambiguity. This underscores the challenge of identifying the necessity of connectives and motivates the need for complementary metrics.

3 Overview

In this section, we provide a brief overview of our study design. We introduce the dataset of connective edits extracted from WikiHow revisions (Section 4), including deletions, replacements, and manually annotated subsets of these edits used to assess their function. Building on this, we conduct experiments on the annotated replacement instances, comparing information-theoretic measures with prompt-based judgments from large language models (Section 5). Together, these components allow us to investigate how connective edits affect discourse coherence and how well computational models capture these preferences.

4 Data

As a framework for discourse relations, we adopt the Penn Discourse Treebank (PDTB) (Prasad et al., 2018) and use its inventory of discourse connectives. Building on prior work by Aktas and Roth (2025), we use the wikiHowToImprove dataset (Anthonio et al., 2020), which contains sentence-level revisions. While that prior study primarily focused on connective insertions, we target two complementary operations: the deletion of existing connectives and the replacement of one connective with another. We introduce a subset of instructional text revisions where explicit discourse connectives are either deleted from the beginning of a sentence (§4.1) or replaced with another (§4.2).

4.1 Connective Deletions

We identified 13,597 instances where discourse connectives were deleted from the beginning of a sentence, as illustrated in the top row of Table 1. In a quantitative analysis, we find the most commonly deleted connectives to be *then*, *also*, and *and*.[1] To better understand these deletions, we qualitatively analyzed 100 examples (four for each of the 25 most frequently deleted discourse connectives). In 64% of cases, the connective was redundant or its removal improved sentence flow. In 21%, deletion was necessary to resolve syntactic or interpretability issues. In the remaining 15%, deletion should not happen as it is altering the intended meaning.[2]

We examined the connective deletion statistics in comparison to the connective insertion statistics provided by Aktas and Roth (2025). Across the dataset, there is a clear asymmetry between these two types of revisions: the number of connective deletions (13,597) significantly exceeds the number of insertions (4,274).

As shown in Appendix E, connectives such as *then*, *finally*, *and*, and *so* are deleted far more often than inserted, suggesting they are often perceived as redundant. This supports the findings of Torabi Asr and Demberg (2012), who argue that causal, temporal, and additive relations (when not marking an event shift) are frequently left implicit. In contrast, connectives like *for example* and *in addition* are more often inserted, pointing to their role in improving clarity. Interestingly, although *for example* signals an "instantiation" relation in PDTB, which Torabi Asr and Demberg (2012) describe as typically implicit, our data shows frequent explicit realization. This discrepancy suggests the need for further investigation, possibly by analyzing the full dataset in terms of discourse relations, rather than focusing only on those made explicit or implicit through sentence-initial edits.

4.2 Connective Replacements

We identified 1,841 revisions in which discourse connectives occurring at the beginning of a sentence were replaced by another connective, as

[1] We list the 15 most frequently deleted connectives in Table 7 in Appendix E.

[2] For an example, see Figure 3 in Appendix A where deleting the contrastive connective 'Otherwise" causes the instruction to lose its conditional framing.

Deletion	Cut the four shirt pieces out of the sheet material. ~~Then~~ place those pieces on the lining material, face side up, and pin them into place before cutting them out.
Replacement	Pick a music choice that agrees with your style choice. ~~So~~ For example, if your dance is fast, pick a fast song.

Table 1: Examples of connective edits in our data. In the first example, the connective is deleted (*Then*), while in the second example the connective (*So*) is replaced by another connective (*For example*).

Revision necessary	The revised version is the only option that conveys the necessary meaning or fits the syntactic context.	11 cases
Revision better	Both options are plausible, but the revision is preferred in terms of clarity, formality, or fluency.	35 cases
Either way	Both the original and revised versions are similarly acceptable.	36 cases
Original better	Both options are plausible, but the original is preferred in terms of clarity, formality, or fluency.	10 cases
Original should stay	The original version is the only option that conveys the necessary meaning or fits the syntactic context.	8 cases

Table 2: Description of labels for annotation and absolute counts after aggregation by majority vote.

illustrated in the bottom row of Table 1. The most common replacement is *but* → *however*.[3] While both connectives share the same dominant PDTB sense, COMPARISON.CONCESSION.ARG2-AS-DENIER, *however* is slightly less ambiguous, linked to only 8 different PDTB senses compared to 13 for *but*, and arguably better aligned with the formal tone of instructional texts.

However, reduced ambiguity or stylistic considerations do not fully explain all replacement choices. For instance, *since* is more ambiguous than *because*, yet *because* → *since* appears among the top five replacement pairs. Apart from replacements like *if* ↔ *even if*, which reflect a clear shift in discourse relation, most replacements preserve the original PDTB sense. This suggests many changes are guided more by tone, readability, or syntactic fit rather than discourse semantics. To gain deeper insight into the nature of these replacements, we analyze four randomly selected instances for each of the top 25 replacement pairs.[4] As this data is used for experiments (§5), we collect three independent annotations for each instance to ensure data quality and aggregate them by majority vote.[5]

The labels and aggregated counts are listed in Table 2. The average agreement with the majority label is 74%. The high frequency of the *Either_way* label (36%) is expected as the original and revised connectives convey the same discourse relation in 48% of cases.[6] By contrast, only 11% of the revisions are annotated as clearly necessary, and 8% are seen as inappropriate. The remaining cases appear to reflect stylistic or contextual preferences rather than discourse-level necessity.

5 Pilot Study on Replacements

Using the data described in Section 4.2, we evaluate whether language models can detect subtle discourse preferences between original and revised versions of connectives. We compare two different methodologies: (1) an information-theoretic approach based on token-level log-likelihoods, and (2) prompting-based judgments from more recent large language models. While the former offers an interpretable output grounded in language predictability, the latter captures higher-level discourse reasoning not available through raw probabilities. We describe information-theoretic measures (§5.1) before turning to a comparative evaluation (§5.2).

5.1 Information-Theoretic Measures

To measure a language model's uncertainty and the predictability of text, we use *surprisal*. For `GPT-2`, the input is first tokenized, and then, the model

[3] Frequent pairs are shown in Table 8 in Appendix E.

[4] Since this dataset includes only edits at the beginning of sentences, all annotated cases involve sentence-initial edits.

[5] One annotator is an author of this paper, whereas the other two are PhD students in Computational Linguistics.

[6] As determined by the most common PDTB sense annotations (e.g., *once* and *after* both signal *temporal.asynchronous.succession*).

computes the probability of each token given its preceding context. The *surprisal* of each token is then calculated as the negative base-2 logarithm of its predicted probability.

Average Token Surprisal For each token t_i, its *surprisal* $S(t_i|t_{<i})$ measures how unexpected it is given the preceding context $t_{<i}$ according to the language model:

$$S(t_i|t_{<i}) = -\log_2 P(t_i|t_{<i})$$

The *average token surprisal* for a sequence $T = (t_1, t_2, \ldots, t_N)$ is:

$$\text{Avg } S(T) = \frac{1}{N}\sum_{i=1}^{N} S(t_i|t_{<i})$$

This value reflects the overall predictability of the sequence for the language model, with lower values indicating greater predictability.

Variance of Surprisal The variance of surprisal, $\text{Var}(S)$, quantifies the spread of token-wise surprisal values around their mean:

$$\text{Var}(S) = \frac{1}{N}\sum_{i=1}^{N} \left(S(t_i \mid t_{<i}) - \text{Avg}(S)\right)^2$$

Higher variance indicates larger fluctuations in predictability across tokens, whereas lower variance suggests a more uniform distribution.

Smoothness of Surprisal Surprisal smoothness captures how abruptly the language model's predictability shifts from one token to the next. It is defined as the mean absolute difference between the surprisals of consecutive tokens:

$$\text{Smoothness}(S) = \frac{1}{N-1}\sum_{i=1}^{N-1} |S(t_{i+1}) - S(t_i)|$$

A lower value indicates more gradual changes (smoother transitions), while a higher value reflects more abrupt predictability shifts.

5.2 Comparative Evaluation

We computed model preferences using `GPT2`[7], comparing the "original" and "revision" versions based on surprisal metrics described in Section 5.1. Among these, only the differences in *average token surprisal* were statistically significant across human annotation categories ($p < 0.001$), suggesting a meaningful correlation between average surprisal changes and human preferences. In contrast, differences in *variance* and *smoothness* did not reach statistical significance.

Gold label	Equal	Orig	Revi
Either_way	**8**	14	14
Original_better	1	**6**	3
Original_should_stay	2	**6**	0
Revision_better	6	9	**20**
Revision_necessary	2	0	**9**

Table 3: Prediction distribution from `GPT2-large`. **Orig** indicates preference for the original connective; **Revi**, for the revised connective; and **Equal** indicates both versions received identical surprisal values.

For a comparison with a more recent model, we conducted prompting-based experiments using `GPT-4.1-mini`. In each prompt, the connective was replaced with "<..>" and the model was explicitly asked to suggest an appropriate discourse connective for that position (see Appendix C for examples). Each item was evaluated across five independent runs. If the original connective was predicted more frequently than the revised one, we interpreted this as a preference for the original; if the revised one was more frequent, it was considered preferred. Equal prediction rates (or no prediction) were treated as indicating no clear preference.

Results Table 3 and Table 4 summarize the predictions of `GPT-2.` and `GPT-4.1`, respectively, on the 100 manually annotated instances.[8] Despite some variation across individual categories, the overall distributional patterns between both models are largely consistent. A *chi-square test* performed across categories revealed no statistically significant differences between two models' predictions.

We also evaluated `GPT-4.1-mini` as a classifier by prompting it with the same 5-way classification instructions provided to our human annotators. Each of the 100 instances was labeled through 5 runs, and the majority prediction was compared to the human majority label. The model's predictions matched the human majority in 43% of the cases, with a moderate association between the two label

[7] Specifically, the `GPT2-large` model from the `transformers` library (Wolf et al., 2020).

[8] Note that zero-shot and few-shot prompting did not yield statistically significant differences in `GPT-4.1`'s output.

Gold label	Equal	Orig	Revi
Either_way	**11**	9	16
Original_better	3	**5**	2
Original_should_stay	2	**5**	1
Revision_better	10	3	**22**
Revision_necessary	2	0	**9**

Table 4: Prediction distribution from `GPT-4.1` on the manually annotated dataset.

sets (Cramér's V = 0.306, p-value = 0.002).

We further evaluated open-source LLMs, specifically `Mistral-7B` and `LLaMA-3.1-8B`. Both models had difficulty adhering to the prompt instructions unless supplied with a short list of example connectives. In consequence, they heavily favored items from the provided list, resulting in reduced lexical diversity. To address this limitation, we also experiment with a few-shot prompting strategy rather than explicitly listing options. Under this configuration, `Mistral-7B` exhibited improved performance, producing valid connectives more consistently and with greater lexical variety. However, no statistically significant correlation with human annotations was found ($p > 0.05$; see Table 5 in Appendix D for the predictions of `Mistral-7B`).

6 Conclusion

We present a dataset of text revisions involving the deletion or replacement of discourse connectives at the beginning of sentences in the WikiHow text revisions.[9] From this dataset, we manually annotated 100 instances of connective replacements. Only 11% of these edits were judged necessary to convey the correct discourse meaning, whereas in 8% cases, the original connective was favored; the remaining edits appeared to be somewhat optional.

Using `GPT-2`, we computed information-theoretic metrics (mean surprisal, variance, and smoothness) for these annotations. Of these, only mean surprisal significantly correlated with human judgments (RQ1). A prompt-based evaluation with `GPT-4.1-mini` showed similar preferences, especially in edge cases of 5-way classification, where a revision was necessary or the original connective should be retained (RQ2). These results suggest that information-theoretic metrics and prompt-based methods capture some patterns in human decisions on discourse connectives, though their coverage remains limited.

[9]The dataset is publicly available at: `https://github.com/berfingit/connective-deletion-replacement`.

Acknowledgements

The research presented in this paper was funded by the DFG Emmy Noether program (RO 4848/2-1).

Limitations

Our human annotation analysis is limited to a small subset of the connective replacement data; expanding annotations to include more examples and other edit types (e.g., insertions and deletions) would strengthen the generalizability of our findings. Additionally, we do not explicitly ground our analysis in a theoretical framework such as the Uniform Information Density (UID) hypothesis, leaving open questions about the broader cognitive or linguistic implications of our results.

For computing surprisal-based metrics, we rely on `GPT-2`, a relatively outdated language model. This choice is motivated by the lack of token-level log-probability access in widely used API-based models like `GPT-3.5` and `GPT-4`. While more recent open-source models such as `Mistral-7B` and `LLaMA-3.1` provide access to logits (via the `transformers` library) and can be used for surprisal computation, integrating them was beyond the scope of this pilot study. We leave surprisal-based experiments with more recent models to future work.

References

Berfin Aktas and Michael Roth. 2025. Clarifying underspecified discourse relations in instructional texts. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12237–12256, Vienna, Austria. Association for Computational Linguistics.

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.

Thomas Hikaru Clark, Clara Meister, Tiago Pimentel, Michael Hahn, Ryan Cotterell, Richard Futrell, and Roger Levy. 2023. A cross-linguistic pressure for Uniform Information Density in word order. *Transactions of the Association for Computational Linguistics*, 11:1048–1065.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego

de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Byung-Doh Oh and William Schuler. 2023. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

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

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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI*. Accessed: 2024-11-15.

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.

Claude E. Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

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. 2015. Uniform surprisal at the level of discourse relations: Negation markers and discourse connective omission. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 118–128, London, UK. Association for Computational Linguistics.

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.

A Examples with Additional Context

Source: *Choreograph a Great Solo*
Section: Steps

Original:
2. Pick a music choice that agrees with your style choice. **[So]**, if your dance is fast, pick a fast song.

Revised:
2. Pick a music choice that agrees with your style choice. **[For example]**, if your dance is fast, pick a fast song.

Figure 1: Example of connective replacement

Article: *Make a Summer Dress out of a Bedsheet*
Section: Steps

Original:
9. Cut the four shirt pieces out of the sheet material. **[Then]** place those pieces on the lining material, face side up, and pin them into place before cutting them out.

Revised:
9. Cut the four shirt pieces out of the sheet material. Place those pieces on the lining material, face side up, and pin them into place before cutting them out.

Figure 2: Example of connective deletion in revision

Source: *Chat Using Facebook Messenger App on iOS*
Section: Logging into Messenger

Original:
1. Sign in. If you're already using the Facebook app and have it installed and running on your mobile device, just tap on the blue "Continue as…" button upon launch. **[Otherwise]**, input your email and password for your Facebook account.

Revised:
1. Sign in. If you're already using the Facebook app and have it installed and running on your mobile device, just tap on the blue "Continue as…" button upon launch. Input your email and password for your Facebook account.

Figure 3: Example of misleading connective deletion.

Source: *Use Every Nikon Digital SLR*

Original:
[While] there are enough similarities between all Nikon digital SLRs.
These categorisations are used here for convenience's sake and have nothing to do with image quality.

Revised:
There are enough similarities between all Nikon digital SLRs.
These categorisations are used here for convenience's sake and have nothing to do with image quality.

Figure 4: Example of connective deletion causing ungrammatical output.

B Information-Theoretic Metrics on the Full Dataset

We analyze the whole dataset using the metrics described in §5.1 and present the results in this section.

Average token surprisal: Connective insertions and replacements lead to a statistically significant increase in average surprisal, while deletions cause a significant decrease (Figure 5). An increase in average surprisal suggests that the resulting text became less predictable overall, whereas a decrease reflects a shift toward greater predictability.

Interestingly, the reverse of insertions (i.e., removing connectives that had been inserted by humans, labeled as `insertion_rev`) also leads to a decrease in average surprisal, mirroring the effect of deletions. The difference between `deletion` and `insertion_rev` is not statistically significant, suggesting that these two operations have symmetrical effects on average surprisal despite differing in context.

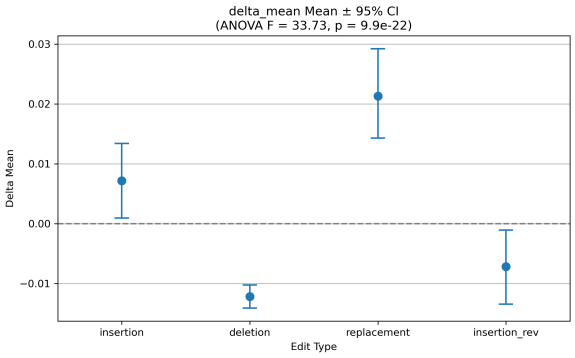

Figure 5: Mean change in surprisal (Δ Mean) across edit types. Error bars represent 95% confidence intervals. ANOVA indicates a significant effect of edit type on mean surprisal (F = 59.76, p = 1.4e-26).

Variance: As shown in Figure 6, both deletion and replacement significantly reduce surprisal variance, making the text more uniformly predictable. In contrast, insertions and therefore their reverse functions `insertion_rev` do not significantly alter variance.

This asymmetry between deletions and `insertion_rev`, despite their functional similarity, suggests that while the mean surprisal remains stable, the local variance depends more on the editing context.

Smoothness: Figure 7 shows that insertions, deletions, and replacements each lead to small

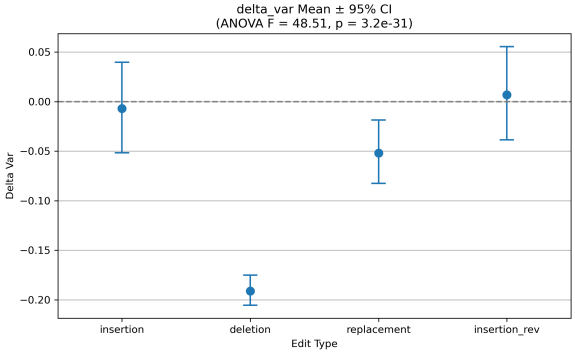

Figure 6: Mean change in variance (Δ Var) across edit types. Error bars represent 95% confidence intervals. ANOVA indicates a significant effect of edit type on surprisal variance (F = 55.11, p = 1.4e-24).

but statistically significant decreases in surprisal smoothness (i.e., negative values indicate that revisions are less smooth). However, the differences across edit types are not statistically significant, suggesting that all three introduce increases in local unpredictability.

There is a clear contrast between deletions and `insertion_rev` in terms of smoothness. Removing human-inserted connectives appears to make the surprisal values change more gradually, yielding a smoother predictability pattern.

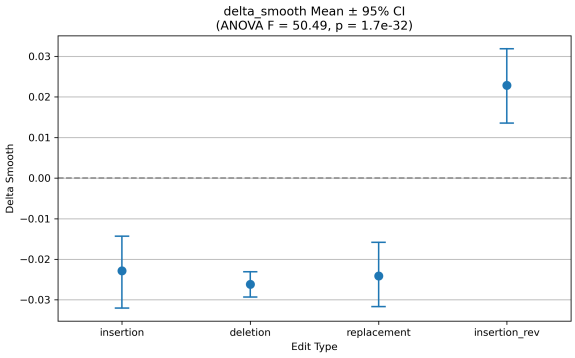

Figure 7: Mean change in smoothness (Δ Smooth) across edit types. Error bars represent 95% confidence intervals. ANOVA does a not indicate a significant effect of edit type on smoothness (F = 0.42, p = 0.66).

Discussion These results show that while `deletion` and `insertion_rev` have similar effects on average surprisal, they differ notably in their impact on surprisal variance and smoothness. This suggests that although a connective's overall informativeness may remain stable, its influence on local predictability (i.e., whether it smooths or disrupts the flow) depends on context.

Overall, the surprisal-based metrics offer a nuanced view of how connective edits affect a language model's expectations (as measured by GPT-2). Interestingly, connective insertions and replacements increase average surprisal and reduce smoothness, indicating that these edits do not improve predictability as one might expect with explicit connectives. Since some of these findings run counter to intuition and we do not yet provide a deeper analysis, we consider them inconclusive.

C Prompts

Prompt Example:

```
TASK: Insert an appropriate discourse
connective into the gap marked by <...>
in the following text.


Text:

Steps 1.  To easily interpret your
quotation and make a plausible argument
and analysis, ask yourself questions.
For example, why was this said? What
possible situation made this quotation
significant? 2. For your main points,
ask yourself questions based on your
interpretation.  Always remember why,
what, and how. 3. <...> you have your
main points, come up with solid examples.
Books, movies, politics, current events,
music, art, culture, history, and any
other category will work. Just remember
to not use all of one category! Mix them
up for a more solid analysis.


Answer:
```

D Mistral-7B's Connective Predictions on Annotated Data

Human Majority	Equal	Orig	Revi
Either_way	31	2	3
Original_better	10	0	0
Original_should_stay	5	2	1
Revision_better	27	6	2
Revision_necessary	10	0	1

Table 5: Mistral-7B's match breakdown by human majority class

E Frequency Distribution of Connective Deletions and Replacements

Connective	Count	Normalized (%)
Then	882	20.5%
For example	669	15.5%
However	558	13.0%
Also	425	9.9%
But	237	5.5%
Or	185	4.3%
And	177	4.1%
So	174	4.0%
For instance	139	3.2%
If	121	2.8%
Finally	118	2.7%
Instead	106	2.5%
In addition	70	1.6%
Otherwise	42	1.0%
In fact	35	0.8%

Table 6: Most frequent connective insertions with raw counts and normalized percentages.

Connective	Count	Percentage (%)
Then	4287	31.5%
Also	1690	12.4%
And	1600	11.8%
So	971	7.1%
But	960	7.1%
However	917	6.7%
Finally	851	6.3%
Or	414	3.0%
For example	374	2.8%
Instead	147	1.1%
Because	120	0.9%
In order	114	0.8%
Therefore	99	0.7%
If	97	0.7%
For instance	88	0.6%

Table 7: Most frequent connective deletions with raw counts and normalized percentages.

Replacement	Count	Percentage (%)
But → However	314	17.1
When → If	106	5.8
If → When	80	4.3
And → Also	65	3.5
Because → Since	39	2.1
However → But	35	1.9
If → Even if	31	1.7
Also → In addition	29	1.6
When → Once	23	1.2
And → In addition	22	1.2
Once → When	18	1.0
Even if → If	18	1.0
While → Although	18	1.0
Then → Finally	18	1.0
So → Therefore	18	1.0

Table 8: Most frequent connective replacements with raw counts and normalized percentages.