Enhancing Coreference Resolution with LLM-driven Data Augmentation and Adversarial Filtering

Anonymous ACL submission

Abstract

Coreference resolution is a fundamental task in natural language processing that involves linking different references to the same entity within a text. However, existing models often struggle to reliably identify referential relationships in contexts with extensive length or complex modifiers. This study proposes a data augmentation technique adding adjective phrases and employing a prompt-based adversarial filtering pipeline to address these challenges. Specifically, we generated and inserted contextually appropriate adjective phrases through the interaction between GPT-4o-mini based Few-shot Prompting and a Discriminative Language Model. The grammatical and semantic consistency of these phrases was validated via human evaluation and inter-annotator agreement (IAA) procedures. The generated synthetic dataset was integrated with existing data, leading to enhanced model performance. On the LitBank dataset, the CoNLL-F1 score increased by up to 2.4%, while the synthetic dataset improved linguistic diversity and the complexity of referential structures. The proposed pipeline represents a significant step towards developing coreference resolution models capable of better capturing linguistic variety and demonstrating robustness under challenging conditions.

1 Introduction

Coreference resolution (Karttunen, 1969) is a fundamental challenge in natural language processing, requiring the accurate identification and linking of multiple mentions referring to the same entity within a document. It plays a crucial role in applications such as pronoun resolution, information retrieval, document summarization, question answering, and dialogue systems. While recent advances in pre-trained Large Language Models based on the Transformer architecture (Vaswani et al., 2017) significantly improve performance, challenges remain, particularly in scenarios requiring long-range contextual reasoning or the interpretation of complex lexical structures. Existing coreference resolution datasets (Pradhan et al., 2013; Bamman et al., 2020) often consist of relatively simple sentence structures, limiting models' ability to learn linguistically diverse patterns—particularly those involving adjectival and adverbial modifiers. These more intricate expressions are especially prevalent in literary texts, and the inability to learn them effectively can substantially impair the generalization performance of a model. This limitation is further exacerbated in real-world applications, where models frequently encounter highly modified and contextually complex language, making robust coreference resolution even more challenging.

To address these issues, recent studies explore data augmentation (Feng et al., 2021) and adversarial filtering (Bras et al., 2020). Data augmentation is a well-established technique that exposes models to a variety of linguistic patterns, reducing their reliance on specific expressions or biased features. Adversarial filtering, in contrast, generates and curates sophisticated example variants, encouraging models to learn linguistic cues and complex relationships that might otherwise be overlooked. There is also growing interest in combining adversarial filtering with data augmentation to systematically adjust dataset difficulty and mitigate model weaknesses (Bhargava and Ng, 2022).

However, existing research often focuses on techniques such as synonym substitution, sentence reordering, and noise injection to generate challenging examples, even within adversarial filtering frameworks. While these techniques are effective for generating difficult-to-distinguish examples, they fall short in tasks like coreference resolution, where context preservation, referential integrity, and entity recognition are crucial. For instance, a model cannot inherently recognize that "the city" and "the breathtakingly vibrant city" refer to the same entity. This highlights the need for methods that deliberately incorporate syntactic modifiers, such as adjectives and adverbials, to enrich coreferential expressions and better reflect natural linguistic variation. This approach enables the model to perform coreference resolution based on contextual understanding and referential reasoning rather than relying on simple keyword matching.

Based on this perspective, this work presents a dataset that is augmented with complex adjective phrases, and proposes a prompt-based adversarial filtering pipeline to generate complex adjectival variants for coreferent mentions. The main contributions of this study are as follows: (1) To complement the monotonous representation of existing coreference resolution datasets, we introduce examples with modifier phrases to expand learning opportunities for complex coreference relations. (2) We design a Prompting-based Adversarial Filtering pipeline that utilizes GPT-4o-mini (Brown et al., 2020) as a Generator Language Model, proposing a data selection method that considers both contextual relevance and difficulty. The augmented dataset is validated through Inter-Annotator Agreement following human evaluation. (3) We construct a synthetic dataset by integrating the augmented dataset with the original data and fine-tune a pretrained language model, which significantly improves the F1 score of coreference resolution models. By integrating data augmentation techniques into coreference resolution research, this study introduces a novel approach that simultaneously enhances model performance and data quality.

2 Related Works

2.1 Coreference Resolution

Coreference resolution refers to identifying and linking multiple expressions that denote the same entity within a text (Karttunen, 1969). It is typically categorized into entity and event coreference. In this study, we focus on entity coreference resolution, which involves identifying groups of expressions that refer to the same real-world entity (Haghighi and Klein, 2010). The process generally comprises two stages: mention detection and mention linking (Pradhan et al., 2012). The former detects expressions that can serve as mentions of entities, while mention linking groups them into coreference clusters (Lee et al., 2017).

Several benchmark datasets are widely used for coreference resolution, including CoNLL 2012 (Pradhan et al., 2012), GAP (Webster et al., 2018), LitBank (Bamman et al., 2020), and WikiCoref (Ghaddar and Langlais, 2016). CoNLL 2012 covers multiple languages, including English, Chinese, and Arabic, and spans various text genres. GAP comprises sentence pairs containing gender-ambiguous pronouns extracted from Wikipedia articles. LitBank provides fine-grained coreference annotations for literary texts, whereas WikiCoref includes annotated with both entity types and coreference links from Wikipedia corpora.

Coreference resolution models can be broadly categorized based on their learning paradigms into mention-pair classifiers (Haghighi and Klein, 2010), entity-level models (Clark and Manning, 2016), latent-tree models (Fernandes et al., 2014), and mention-ranking models (Wiseman et al., 2016). More recently, deep learning and transformer-based large language models (Vaswani et al., 2017) are introduced to further enhance coreference resolution performance. However, challenges remain in handling complex contextual dependencies and modifier phrases.

2.2 Adversarial Filtering

Adversarial filtering is a technique designed to systematically increase the difficulty of a dataset in order to effectively evaluate the limitations of machine learning models. This method involves using a weak model to make predictions on data samples, discarding those that are easily answered correctly, and retaining only the samples for which the model produces incorrect answers or expresses uncertainty. Such a process prevents models from relying on superficial patterns or dataset biases and encourages the development of deeper reasoning capabilities and improved generalization. This method proves particularly useful in enhancing data quality in tasks and is employed in the construction of large-scale, challenging datasets such as HellaSwag (Zellers et al., 2019).

Recent advancements in adversarial filtering demonstrate its effectiveness in various domains. DISCOSENSE (Bhargava and Ng, 2022) extends the adversarial filtering framework by introducing Controlled Adversarial Filtering, leveraging discourse connectives to assess commonsense reasoning abilities and generating adversarial distractors to increase evaluation difficulty.

Specifically, we employ GPT-4o-mini (Brown et al., 2020) to generate, insert, and replace adjectival phrases in coreference expressions. The modi-

Dataset	#Train	#Dev	#Test
LitBank	80	10	10
PreCo	36,120	500	500

Table 1: Number of documents in LitBank and PreCo datasets.

Dataset	#Best	#Worst	#Weird
Augmented LitBank	184	128	23
Augmented PreCo	4,029	1,371	1,193

Table 2: Number of augmented cases in LitBank and PreCo datasets.

fied instances are then filtered via a discriminative language model, yielding a more challenging and informative dataset. Through this approach, we aim to simultaneously enhance both the performance and robustness of coreference resolution models by exposing them to more complex linguistic patterns.

3 Methodology

3.1 Task Description

Coreference resolution refers identifying and linking multiple mentions of the same entity within a given text (Karttunen, 1969). In this study, we generate a difficult dataset by augmenting correctly predicted instances with adjectival phrases. The adversarial dataset is then combined with the original data to construct the final synthetic dataset. Training on this synthetic dataset aims to enhance coreference resolution performance.

3.2 Dataset Format

OntoNotes Formatting The OntoNotes dataset (Pradhan et al., 2013) is structured as a collection of documents, each containing multiple sentences. Each sentence is represented as a word-level list, and a document is formed by aggregating these sentence lists. This hierarchical structure facilitates contextualization and enables effective modeling of document-level coreference relationships.

Cluster Structure A coreference cluster is defined as a set of mention offsets that refer to the same entity. Each offset specifies the start and end indices of a particular word or phrase within a document, uniquely identifying its occurrence. Mentions sharing the same reference are grouped into clusters based on their offsets, allowing the model to learn and distinguish different coreference relationships.

Augmented Descriptive Phrase Structure In this study, we leverage a generative language model to expand the scope and complexity of the dataset by incorporating descriptive phrases into coreferential noun phrases. For instance, if the noun phrase "the city" appears in a sentence, an adjectival phrase such as "the beautiful city" is introduced to enhance linguistic diversity while preserving the coreference relationship.

3.3 Datasets

LitBank (Bamman et al., 2020) is an annotated dataset comprising 100 works of English literature, widely utilized in NLP and computational humanities. It specializes in literary texts, containing documents with long contextual spans and complex narrative structures. These characteristics enable a more sophisticated evaluation of coreference resolution models that must process long-range dependencies. Unlike general-domain texts such as conversational transcripts or news articles, literary texts are distinguished by their stylistic diversity, frequent use of metaphors, and long-range dependencies. These features make LitBank particularly well-suited for assessing a model's long-range inference capabilities and anti-forgetting performance in long documents with intricate coreference structures. It has been widely used for tasks such as character tracking, event extraction, relationship modeling, and literary analysis.

PreCo (Chen et al., 2018) is a large-scale coreference dataset based on English textbooks, featuring simpler syntax and explicit coreference chains compared to the literature-focused LitBank. This domain contrast allows us to test the generalizability of our method. Despite augmenting only a subset of PreCo due to its size, we still observed CoNLL-F1 improvements, suggesting effectiveness across diverse linguistic settings.

Details of the dataset composition and augmentation are provided in Table 1 and Table 2. Table 1 summarizes the number of training, development, and test instances from LitBank and PreCo used for model fine-tuning, while Table 2 presents the distribution of augmented subsets derived from the respective training sets. The three evaluation criteria used in our experiments are described in Appendix B.

3.4 Prompting-based Adversarial Filtering

The proposed data augmentation pipeline extends adversarial filtering to coreference resolution, emphasizing the interaction between a discriminative language model and a generator language model. This pipeline is designed to enhance the general-

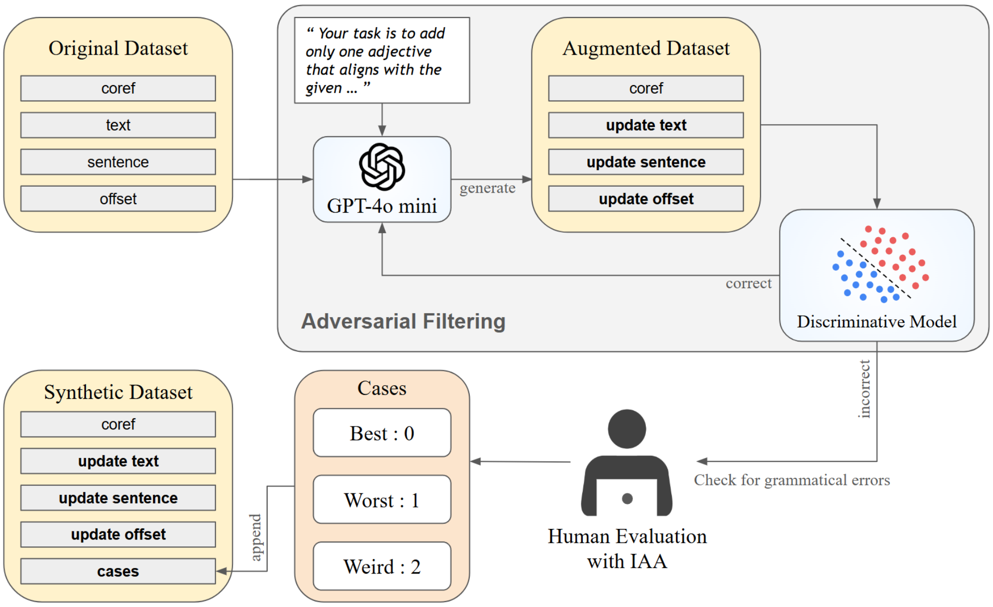

Figure 1: Overall pipeline. The gray rectangle represents the Prompting-based Adversarial Filtering process. If the discriminative model succeeds in making a prediction, the process repeats; otherwise, the data is collected and moved to the human evaluation phase.

Input (Source Sentence):
On either side of this road straggled two uneven rows of wooden buildings ; the general merchandise stores, the two banks, the drug store, the feed store, the saloon, the post-office. On the sidewalk in front of one of the stores sat a little Swede boy, crying bitterly .
Output (Augmented Sentence):
On either side of this road straggled two uneven rows of wooden buildings; the general merchandise stores, the two banks, the drug store, the feed store, the saloon, the post-office. On the sidewalk in front of one of the **various** stores sat a little Swede boy, crying bitterly.

Table 3: Example of Valid Augmented Sentence. Adjective phrases are added appropriately before nouns.

Input (Source Sentence):
In accordance with this rule, it can reasonably be assumed that Boston's forefathers built their first prison-house somewhere near Cornhill around the same time they established the earliest burial ground on Isaac Johnson's land, surrounding his honored grave. This grave later became the center of all the tombs gathered in the old churchyard of King's Chapel.
Output (Augmented Sentence):
In accordance with this rule, it can reasonably be assumed that Boston's forefathers built their first prison-house somewhere near Cornhill around the same time they established the earliest burial ground on Isaac Johnson's land, surrounding his **respected** honored grave. This grave later became the center of all the tombs gathered in the old churchyard of King's Chapel.

Table 4: Example of Invalid Augmented Sentence. An adjective phrase has been added after the pronouns unnaturally.

ization of model performance and robustness by incrementally introducing difficult examples, such as descriptive phrases, into the coreference resolution dataset through the generator model. This dataset is then filtered using the discriminative model, which filters the generated data to regulate quality and adjust difficulty levels.

Discriminative models predict coreference relationships from input data and compare them to gold-standard annotations to identify instances where the model already makes correct inferences. In this study, we employ Maverick-mes (Martinelli et al., 2024) as the discriminative model. The generator model increases the complexity of the dataset by adding or replacing descriptive phrases before coreference expressions. The newly generated examples are then validated by the discriminative model. For this purpose, GPT-4o-mini is utilized as the generator model.

To ensure that the generator model accurately determines the appropriate placement and integration of descriptive phrases, we provide explicit examples within the prompts to facilitate the generation of more natural and contextually appropriate adjectival phrases. Furthermore, we develop an automated pipeline to generate modified data based on the prompts, which is subsequently validated and filtered using the discriminative model. Figure 1 illustrates the complete process of Prompting-based Adversarial Filtering. Starting with the original dataset, the generator model inserts appropriate descriptive phrases before coreference expressions.

Table 3 shows the valid cases of the augmented LitBank dataset, and Table 4 shows the invalid cases from our augmented LitBank dataset. Although the underlying format follows OntoNotes, we present the examples in standard sentence format for readability. The underlined words indicate coreference mentions, while bold-faced words represent augmented descriptive phrases. An example prompt template for adversarial filtering is provided in Appendix A.

3.5 Human Evaluation with Inter-Annotator Agreement(IAA)

We conduct a human evaluation to assess the quality of the data generated by the Few-shot Prompt-based Adversarial Filtering process. This evaluation aims to directly assess the grammatical correctness, semantic appropriateness, and coreference relevance of the augmented data. Three researchers perform the evaluation based on predefined criteria, systematically reviewing all augmented datasets produced through the adversarial filtering process. The evaluation criteria are shown in Appendix B.

To ensure the reliability of annotations and assess the consistency of the data, we measure IAA. IAA quantitatively indicates the degree to which multiple annotators consistently make judgments on the same items. We employ two widely used IAA metrics: Fleiss' Kappa (Krippendorff, 2011) and Krippendorff's Alpha (Fleiss, 1971). Fleiss' Kappa measures the agreement level among multiple raters labeling categorical data; in our case, it yields a score of 0.5911, which can be interpreted as moderate agreement. Krippendorff's Alpha, a more generalized metric applicable to various data types and tolerant of missing values, records a score of 0.5915, indicating a level of agreement that reflects acceptable reliability. Given the moderate agreement, final labels are determined via majority voting. In cases where all labels received an equal number of votes, the data is considered uncertain and classified as the worst case. Only the best cases, those with clear annotator consensus, are included in the augmented dataset for further training and evaluation.

4 Experiments

4.1 Model

Maverick-incr is a coreference resolution model based on the Shift-Reduce Paradigm (Clark and Manning, 2016) that incrementally updates the clusters formed in the previous step. The model processes text sequentially and determines whether newly emerged mentions can be linked to existing clusters. If a mention can be included in an existing cluster, it is merged. Otherwise, a new cluster is created to maintain the coreference relationship. Unlike traditional sentence-by-sentence approaches, Maverick-incr favors real-time and sequential processing, making it particularly well-suited for coreference resolution in streaming data or interactive environments where incremental inference is required.

Maverick-s2e is a coreference resolution model based on the Coarse-to-Fine method (Lee et al., 2017). This approach consists of two steps: mention extraction and mention-antecedent classification. In the mention extraction step, the model identifies potential mentions in the text that can be part of a coreference chain. In the next step, the hidden state corresponding to the start and end tokens of an antecedent candidate mention is compared to classify whether it refers to the same entity. Mentions identified as coreferential are grouped into clusters. This two-step approach improves inference efficiency by first narrowing down candidate mentions before applying a more refined classification, avoiding the need for computationally expensive contextual processing.

Maverick-mes follows the same Coarse-to-Fine-based structure as Maverick-s2e but introduces a Multi-Expert Scorer instead of a Mention-Pair Scorer to refine linguistic pattern recognition. Specifically, it defines six linguistic synchronization categories—PRON-PRON-C, PRON-PRON-NC, ENT-PRON, MATCH, CONTAINS, and OTHER—, determines which category a mention belongs to, and computes a score for each category to form clusters. This approach enhances coreference resolution by pre-typing linguistic features such as pronoun-pronoun agreement, noun phrase-pronoun relations, and partial inclusion relationships.

4.2 Comparing Other Discriminative

To comprehensively evaluate the proposed augmented dataset, this study assesses a total of three discriminative coreference models, including the Maverick and two additional architectures, thereby demonstrating the generalizability and practical applicability of the approach across model variants.

LingMess (Otmazgin et al., 2023) is an encoder-only model based on Longformer, designed to efficiently process extended contexts and capture long-range dependencies through a sparse attention mechanism optimized for long documents. It has exhibited strong performance on literary text-based datasets and serves as a suitable baseline for evaluating the effect of adjective insertion within extended contexts.

wl-coref (Dobrovolskii, 2021) is a lightweight model that predicts word-level links using a RoBERTa(Zhuang et al., 2021) backbone and subsequently extracts mention spans, adopting a different strategy from the mention-ranking paradigm. Despite its structural simplicity, it achieves high accuracy in mention detection and is valued for its efficiency.

4.3 Evaluation Metric

MUC (Mention-Unicon Cross) (Vilain et al., 1995) is a metric that evaluates coreference resolution based on the precision and recall of coreference links. Calculated by comparing the number of links between clusters and assessing how accurately the predicted cluster connections align with the gold standard clusters.

B³ (B-Cubed) (Bagga and Baldwin, 1998) evaluates coreference resolution by measuring the precision and recall of individual mentions and computing a weighted average to assess how consistently each mention is assigned to the correct cluster. A model achieves a high score only if it excels in both accurate classification (precision) and error-free retrieval (recall) of mentions.

CEAFe (Constrained Entity Alignment F-Measure) (Luo, 2005) evaluates coreference resolution based on a one-to-one mapping between clusters. If a gold-standard cluster is split into multiple predicted clusters or merged into a single predicted cluster, the score penalization is significant.

CoNLL-2012 F1 Score is calculated as the mean of three F1 scores of above metrics.

The detailed formulas for the evaluation metrics are provided in Appendix C.

4.4 Setup

We utilized DeBERTa-v3(He et al., 2023) as the document encoder for the discriminative language model. DeBERTa improves upon the existing BERT architecture by introducing a disentangled attention mechanism and enhances contextual understanding through an improved lexical embedding method. For optimization, Adafactor (Shazeer and Stern, 2018) was employed with weight decay set to 0.01. The number of training epochs was set to 300, with a learning rate of 3e-4 for the linear layers and 2e-5 for the pretrained encoder. All training was conducted on an RTX 4090 GPU with 24GB of VRAM.

5 Results

We evaluate coreference resolution using MUC, B³, CEAFe, and CoNLL-F1 metrics, each capturing different aspects of performance. Table 5 summarizes the results across various training settings, original, fully augmented, and combined datasets. Table 6 presents the experimental results on the PreCo dataset to evaluate the generalization ability of our models on out-of-domain data. In Table 7, we compare the performance of different discriminative models on the LitBank dataset using three variations of training data.

5.1 MUC (Link-based Evaluation)

Among the three models, Maverick-mes exhibited the highest relative improvement in the MUC score, with an increase of 1.7%. The MUC metric evaluates coreference resolution by counting the correctly predicted links between mentions within each cluster, emphasizing structural integrity over mention accuracy. Since the proposed augmentation strategy introduces more syntactically and semantically rich expressions, without increasing the number of mentions, this qualitative enrichment enhances the model's ability to capture the underlying link structures. Maverick-mes, which leverages part-of-speech-informed features and linguistic cues, benefits significantly from this, especially in resolving pronouns and named entities. These elements are highly sensitive to contextual and syntactic nuances, aligning well with the characteristics of the augmented data. The pronounced gain in MUC thus suggests that Maverick-mes capitalizes more effectively on the descriptive augmentation.

5.2 B³ (Mention-based Evaluation)

The Maverick-incr model demonstrated the greatest improvement in the B³ metric, with a notable increase of 3.8%. B³ evaluates the precision and recall of each individual mention, emphasizing fine-grained mention-level alignment across predicted and gold-standard clusters. Incremental models like Maverick-incr construct clusters by progressively incorporating mentions in sequential order,

Training set	LM	Model	MUC	B^3	CEAFe	CoNLL-F1
LitBank$_{\text{original}}$	DeBERT$_{\text{Large}}$	Maverick$_{\text{incr}}$	85.1	71.8	68.3	75.1
		Maverick$_{\text{s2e}}$	88.1	75.0	65.9	76.3
		Maverick$_{\text{mes}}$	86.1	75.7	66.4	76.1
LitBank$_{\text{Augmented}}$	DeBERT$_{\text{Large}}$	Maverick$_{\text{incr}}$	84.2	71.6	67.8	74.5
		Maverick$_{\text{s2e}}$	87.5	74.7	67.3	76.5
		Maverick$_{\text{mes}}$	84.5	71.5	63.1	73.0
LitBank$_{\text{Augmented}}$ (mean ± std)	DeBERT$_{\text{Large}}$	Maverick$_{\text{incr}}$	82.5 ± 0.87	67.9 ± 2.58	66.9 ± 3.06	71.6 ± 1.45
		Maverick$_{\text{s2e}}$	86.9 ± 0.28	73.8 ± 0.55	63.4 ± 1.25	74.7 ± 0.58
		Maverick$_{\text{mes}}$	85.1 ± 0.71	72.5 ± 1.65	62.7 ± 1.82	73.5 ± 1.35
LitBank$_{\text{Synthetic}}$	DeBERT$_{\text{Large}}$	Maverick$_{\text{incr}}$	86.7	75.6	**70.3**	77.5
		Maverick$_{\text{s2e}}$	**88.5**	**76.7**	68.8	**78.0**
		Maverick$_{\text{mes}}$	87.8	76.5	67.6	77.3

Table 5: Performance on four evaluation metrics for the Maverick model on the LitBank dataset, including the original, augmented, average-augmented, and synthetic variants. Average-augmented results are computed via 5-fold cross-validation on randomly sampled subsets. For each fold, we sample 80 documents to match the size of the original training set. The best score for each metric is shown in bold.

Training set	LM	Model	Avg.F1
PreCo$_{\text{Original}}$	DeBERT$_{\text{Large}}$	Maverick$_{\text{s2e}}$	87.4
		Maverick$_{\text{mes}}$	87.1
PreCo$_{\text{Synthetic}}$	DeBERT$_{\text{Large}}$	Maverick$_{\text{s2e}}$	**87.9**
		Maverick$_{\text{mes}}$	**87.4**

Table 6: Performance of CoNLL-F1 for the Maverick model on the PreCo dataset, comparing original and synthetic training sets.

Training set	LM	Model	Avg.F1
LitBank$_{\text{Original}}$	Longformer$_{\text{Base}}$	LingMess	59.0
LitBank$_{\text{Augmented}}$			59.9
LitBank$_{\text{Synthetic}}$			**60.1**
LitBank$_{\text{Original}}$	RoBERTa$_{\text{Large}}$	wl-coref	63.5
LitBank$_{\text{Augmented}}$			63.3
LitBank$_{\text{Synthetic}}$			**66.3**
LitBank$_{\text{Original}}$	DeBERT$_{\text{Large}}$	Maverick$_{\text{s2e}}$	76.3
LitBank$_{\text{Augmented}}$			76.5
LitBank$_{\text{Synthetic}}$			**78.0**

Table 7: Comparison between Discriminative model on LitBank in terms of CoNLL-F1 Score.

making them particularly sensitive to the local coherence and compatibility of mentions. As the augmented data enhances the contextual richness of each mention, this facilitates more accurate matching and disambiguation. Therefore, the substantial gain in B³ for Maverick-incr reflects its improved ability to accurately include relevant mentions within each evolving cluster.

5.3 CEAFe (One-to-One Cluster Alignment)

CEAFe showed its highest improvement of 2.9% in the Maverick-s2e model. This metric computes similarity based on optimal one-to-one alignments between predicted and ground-truth clusters, rewarding holistic cluster-level accuracy. The start-to-end architecture of Maverick-s2e evaluates all mention pairs within a document and directly models their likelihood of belonging to the same cluster, which aligns well with the cluster-level perspective of CEAFe. The augmentation of descriptive modifiers appears to support this model in distinguishing between ambiguous or overlapping mention sets, ultimately leading to more accurate global cluster structures. This suggests that the augmented input not only improves local decisions but also enhances the model's capacity to form cluster assignments that reflect the ground-truth structures more faithfully.

5.4 Performance on General Purpose Data

PreCo is a dataset that differs from LitBank in both structure and domain, and serves as a comparative setting to evaluate whether the adjective insertion-based augmentation technique proposed in this study generalizes across diverse linguistic environments. According to the results presented in Table 6, both Maverick-s2e and Maverick-mes exhibit a slight improvement in performance when trained on the augmented PreCo dataset, as measured by CoNLL-F1. Although the magnitude of the improvement is modest, the consistency observed across both models provides meaningful evidence supporting the generalizability of the proposed augmentation technique. Moreover, while PreCo tends to emphasize explicit mention links during training, the insertion of modifiers appears to guide the model toward learning complementary semantic cues. This suggests that the augmentation strategy remains effective even in domains with relatively low structural variability.

5.5 Evaluation on Additional Models

Table 7 presents results from additional experiments conducted on two models, LingMess, which is based on Longformer, and wl-coref, a lightweight model using RoBERTa, to assess whether the proposed augmentation method generalizes beyond the Maverick. Both models exhibit performance improvements when trained on the augmented PreCo dataset, as measured by the CoNLL-F1 metric. Notably, wl-coref achieves a 2.8% gain. These findings suggest that the modifier insertion-based augmentation technique is effective across diverse model architectures, including those with fixed-length input handling and simplified linking mechanisms.

5.6 Semantic Data Validation

We conduct follow-up analyses to ensure that the augmented data is not biased toward specific syntactic patterns or semantic categories, and that it preserves the expressive diversity and logical consistency necessary for effective model training. First, to assess whether the original meaning is preserved after modifier insertion, we apply a natural language inference (NLI) model to determine the logical relationship between each original sentence and its augmented counterpart. Cases classified as "entailment" or "neutral" are treated as meaning-preserving. Additionally, to verify that the augmented data is not structurally or semantically concentrated around specific patterns, we visualize the distribution of sentence embeddings using dimensionality reduction techniques, including PCA, t-SNE, and UMAP. These analyses confirm that the augmented sentences are broadly and evenly distributed across the embedding space, without clustering around particular expressions or semantic classes. Full results and visualizations are provided in Appendix D.

5.7 Discussion

In Table 5, the average-augmented dataset was constructed by randomly sampling examples from the augmented pool. This allows us to isolate and evaluate the impact of augmentation itself.

Interestingly, the performance of models trained on the augmented data alone is slightly lower than those trained on the original data. This is due to the adversarial filtering process, which retained more challenging examples (e.g., adjectival modifiers of proper nouns) and filtered out easier ones (e.g., adjectives added to common nouns). Despite containing only these harder cases, the augmented dataset achieves performance comparable to the original, demonstrating the robustness of the proposed augmentation approach. Finally, the synthetic dataset, which includes all types of examples, achieves the best overall performance, confirming that the full augmentation strategy is effective in improving coreference resolution models. Cross-validation is not applied to the synthetic dataset. Similar to the augmentation setting, sampling a subset of the data may lead to training that overfits specific augmented examples. This undermines the intended purpose of augmentation, namely promoting diversity and generalizability. To mitigate this issue, all augmented examples are integrated and used collectively during training.

6 Conclusion

We propose a benchmark dataset for coreference resolution that integrates challenging descriptive phrases through a prompting-based adversarial filtering pipeline. This approach combines few-shot prompting using GPT-4o-mini with adversarial filtering to generate linguistically diverse and underrepresented patterns. Contextually appropriate phrases are inserted through interaction with a discriminative language model and validated via human evaluation, resulting in a high-quality synthetic dataset.

Experimental results demonstrate consistent performance improvements across all evaluation metrics, with CoNLL-F1 gains of up to 2.4% on the LitBank dataset. Notably, similar improvements are observed on the PreCo dataset, highlighting the generalizability of the proposed approach across domains. Model-specific strengths are also identified: Maverick-incr yields the highest gain in B^3, while Maverick-s2e performs best on CEAFe. Furthermore, additional models, including wl-coref and LingMess, also benefit from the synthetic data, confirming the robustness of the augmentation method across diverse model architectures.

In summary, the proposed pipeline enhances both the accuracy and generalization capacity of coreference models by reducing their dependence on overly simplistic patterns and promoting linguistic diversity. Future work will focus on scaling up augmentation with more varied descriptive phrases, broader part-of-speech coverage, and extensions to multilingual corpora and other NLP tasks.

Limitations

Limitations in human evaluation arise due to resource constraints. While manual assessment plays a critical role in ensuring data quality, the involvement of only three annotators limits the generalizability of the findings. Similarly, for the large-scale PreCo dataset, only a subset is augmented due to limited annotation capacity.

In addition, due to constraints in our experimental setup, we were unable to include the Maverick-incr model in the PreCo experiments. Although structurally comparable to the other Maverick variants, the incremental clustering architecture of Maverick-incr requires sequential state updates during inference, resulting in higher computational and memory costs. These resource demands created a bottleneck that prevented scaling to the full PreCo dataset within our infrastructure, thereby limiting the completeness of our cross-model evaluation.

Future work may address these limitations by increasing the number of human evaluators for more reliable qualitative assessment and by optimizing the augmentation ratio to enhance linguistic diversity and dataset balance.

Despite these constraints, the proposed dataset and augmentation pipeline represent a meaningful contribution to improving both linguistic diversity and model robustness in coreference resolution. They also offer valuable insights for the development of more sophisticated NLP models.

References

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An annotated dataset of coreference in English literature. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 44–54, Marseille, France. European Language Resources Association.

Prajjwal Bhargava and Vincent Ng. 2022. DiscoSense: Commonsense reasoning with discourse connectives. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10295–10310, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E. Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. *CoRR*, abs/2002.04108.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Hong Chen, Zhenhua Fan, Hao Lu, Alan Yuille, and Shu Rong. 2018. PreCo: A large-scale dataset in preschool vocabulary for coreference resolution. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 172–181, Brussels, Belgium. Association for Computational Linguistics.

Kevin Clark and Christopher D. Manning. 2016. Improving coreference resolution by learning entity-level distributed representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653, Berlin, Germany. Association for Computational Linguistics.

Vladimir Dobrovolskii. 2021. Word-level coreference resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7670–7675, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Eraldo Rezende Fernandes, Cícero Nogueira dos Santos, and Ruy Luiz Milidiú. 2014. Latent trees for coreference resolution. *Computational Linguistics*, 40(4):801–835.

JL Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378—382.

Abbas Ghaddar and Phillippe Langlais. 2016. WikiCoref: An English coreference-annotated corpus of Wikipedia articles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 136–142, Portorož, Slovenia. European Language Resources Association (ELRA).

Aria Haghighi and Dan Klein. 2010. Coreference resolution in a modular, entity-centered model. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 385–393, Los Angeles, California. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Lauri Karttunen. 1969. Discourse referents. In *International Conference on Computational Linguistics COLING 1969: Preprint No. 70*, Sånga Säby, Sweden.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Giuliano Martinelli, Edoardo Barba, and Roberto Navigli. 2024. Maverick: Efficient and accurate coreference resolution defying recent trends. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13380–13394, Bangkok, Thailand. Association for Computational Linguistics.

Shon Otmazgin, Arie Cattan, and Yoav Goldberg. 2023. LingMess: Linguistically informed multi expert scorers for coreference resolution. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2752–2760, Dubrovnik, Croatia. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4596–4604. PMLR.

Yulia Tsvetkov, Nathan Schneider, Dirk Hovy, Archna Bhatia, Manaal Faruqui, and Chris Dyer. 2014. Augmenting English adjective senses with supersenses. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4359–4365, Reykjavik, Iceland. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

Sam Wiseman, Alexander M. Rush, and Stuart M. Shieber. 2016. Learning global features for coreference resolution. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 994–1004, San Diego, California. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

A Prompt Template for Adversarial Filtering

When using GPT-4o-mini to augment descriptive phrases, it is essential to identify coreference mentions in a given sentence and add modifiers only to those mentions. In doing so, the following considerations should be taken into account when selecting modifiers:

- Avoid repeating the same modifier within a sentence.

- Do not use overly generic modifiers.

- Modifiers should not alter the original meaning of the sentence.

The first issue arises from repeating the same word, which can make the sentence structure awkward and potentially grammatically incorrect. Nevertheless, we excluded repeated modifiers during human evaluation to maintain naturalness. The second issue is that overly generic modifiers fail to contribute meaningfully to identifying coreference mentions, contradicting the purpose of our augmentation strategy. To address this, we instructed annotators to select contextually relevant modifiers derived from the given sentence that do not compromise its original meaning. Detailed prompts for modifier generation are provided in Table 8.

B Human Evaluation Criteria for Augmented Data

The Human evaluation criteria are provided in Table 9.

C Equations of Evaluation Metric

C.1 MUC (Mention-Unicon Cross)

$$MUC_{Precision} = \frac{TP}{TP + FP}$$

$$MUC_{Recall} = \frac{TP}{TP + FN}$$

$$MUC_{F1} = 2 \cdot \frac{MUC_{Precision} \cdot MUC_{Recall}}{MUC_{Precision} + MUC_{Recall}}$$

- TP (True Positives): Correctly predicted links in coreference clusters.

- FP (False Positives): Predicted links that do not exist in the gold standard clusters.

- FN (False Negatives): Links that exist in the gold standard clusters but are missing in the predictions.

C.2 B³ (B-Cubed)

$$B^3_{Precision} = \frac{1}{N} \sum_{i=1}^{N} \frac{|C_i \cap G_i|^2}{|C_i|}$$

$$B^3_{Recall} = \frac{1}{N} \sum_{i=1}^{N} \frac{|C_i \cap G_i|^2}{|G_i|}$$

$$B^3_{F1} = 2 \cdot \frac{B^3_{Precision} \cdot B^3_{Recall}}{B^3_{Precision} + B^3_{Recall}}$$

- C_i: Predicted cluster containing the i-th mention.

- G_i: Gold cluster containing the i-th mention.

- N: Total number of mentions.

- $|C_i \cap G_i|$: Number of mentions shared between the predicted and gold clusters.

C.3 CEAFe (Constrained Entity Alignment F-Measure)

$$Similarity(C, G) = \sum_{(c,g) \in OptimalMatching} \phi(c, g)$$

$$CEAFe_{Precision} = \frac{Similarity(C, G)}{|C|}$$

$$CEAFe_{Recall} = \frac{Similarity(C, G)}{|G|}$$

$$CEAFe_{F1} = 2 \cdot \frac{CEAFe_{Precision} \cdot CEAFe_{Recall}}{CEAFe_{Precision} + CEAFe_{Recall}}$$

$$\phi(c, g) = \frac{2 \cdot |c \cap g|}{|c| + |g|}$$

- C: Set of predicted clusters.

- G: Set of gold clusters.

- $|C|$: Number of predicted clusters.

- $|G|$: Number of gold clusters.

- $\phi(c, g)$: Similarity between a predicted cluster c and a gold cluster g.

C.4 CoNLL-2012 F1 Score

$$CoNLL - 2012_{F1} = \frac{MUC_{F1} + B^3_{F1} + CEAFe_{F1}}{3}$$

Instructions:

You will be given a sentence in OntoNotes format along with a coreference cluster and its offsets. Your task is to add several adjectives that aligns with the given coreference term. The adjective must be placed immediately before the term within the sentence.

Guidelines:

1. Identify the words in the sentence that correspond to each offset.
2. Updated Coreference Offsets should be calculated step by step.
3. For each remaining term (starting from the second), add several adjectives immediately before the term if it adds meaningful context.
4. Never add articles (‘the’, ‘a’), only adjective.
5. Ensure the adjective does not change the sentence’s original meaning.
6. Avoid repeating the same word multiple times in sequence.
7. Use adjectives that are contextually relevant and meaningful. Avoid using too general adjectives like ‘good’, ‘bad’, ‘nice’, or nonsensical combinations.
8. Adjectives should enrich the meaning or add useful information without making the description redundant or awkward.
9. If no suitable adjective can be added without disrupting the meaning or creating redundancy, do not add an adjective at all.
10. NEVER VIOLATE THE OUTPUT TEMPLATE.

Input:

- Sentence: ontonotes_sentence
- Coreference Offsets: offsets
- Coreference Words: words

Output Format:

1. Updated Coreference Words : The modified OntoNotes format sentence with adjectives added.

Example:

Input:

- Sentence: [‘Barack’, ‘Obama’, ‘is’, ‘traveling’, ‘to’, ‘Rome’, ‘.’, ‘The’, ‘city’, ‘is’, ‘sunny’, ‘and’, ‘the’, ‘president’, ‘plans’, ‘to’, ‘visit’, ‘its’, ‘most’, ‘important’, ‘attractions’]
- Coreference Offsets: [[5, 5], [7, 8], [17, 17]]
- Coreference Words: [[‘Rome’], [‘The’, ‘city’], [‘its’]]

Correct Output:

1. Updated Coreference Words : [[‘Rome’], [‘The’, ‘picturesque’, ‘city’], [‘its’]]

Explanation:

- ‘picturesque’ was added to ‘city’ to enrich the description without altering the intended meaning.
- No adjective was added to ‘Rome’ or ‘its’ as it was unnecessary.

Table 8: A sample prompt for adversarial filtering.

Case	Criteria	Original Sentence	Augmented Sentence	Explanation
High-Quality (Best)	The sentence must be grammatically correct while incorporating descriptive phrases that are semantically relevant to the coreferential clusters.	"The man went to the store."	"The diligent man went to the store."	A contextually relevant descriptive phrase, ‘diligent,’ was added before the coreferential word ‘man’.
Unacceptable (Worst)	The augmented descriptive phrases are either grammatically incorrect or not suitable for coreference clusters.	"My name is Jim."	"My name is enchanting Jim."	The descriptive phrase ‘enchanting’ is inappropriate, making it difficult to establish a coreference cluster with ‘Jim’.
Acceptable but Semantically Misaligned (Weird)	The sentence is grammatically correct, but the descriptive phrases or synonyms used as a replacement are semantically inappropriate for the coreferential clusters.	"The cat jumped onto the couch."	"The shiny feline jumped onto the couch."	The adjective ‘shiny’ is contextually inappropriate for the coreferential word ‘cat,’ and the original term has been replaced with its synonym ‘feline.’

Table 9: Examples of coreference-based sentence augmentations categorized by quality: best, weird, and worst, with evaluation criteria and explanations.

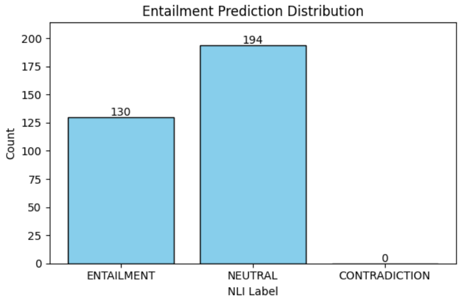

Figure 2: Entailment Prediction Distribution of LitBank Augmented Dataset

D Augmented Data Quality Analysis

To verify whether our augmented data is constructed in a balanced and unbiased manner, we conducted a series of qualitative analyses, including NLI label distribution and dimensionality reduction techniques such as PCA, t-SNE and UMAP. These analyses serve to confirm the semantic soundness and class distribution of the generated examples beyond numerical performance scores.

D.1 NLI Label Distribution

To verify whether the inserted adjectival phrases altered the original meaning, we employed a Natural Language Inference (NLI) model, which classifies the relationship between a premise and a hypothesis into entailment (logical consistency), contradiction (logical conflict), or neutral (independence). In our setting, the original sentence served as the premise and the augmented one as the hypothesis.

We regarded entailment and neutral as acceptable, indicating meaning preservation or harmless addition, while contradiction signified semantic inconsistency and led to data exclusion.

As shown in Figure 2, the label distribution consisted of 130 entailment and 194 neutral cases, with no contradictions, demonstrating that the augmentation maintains logical coherence and avoids semantic noise.

D.2 PCA Projection

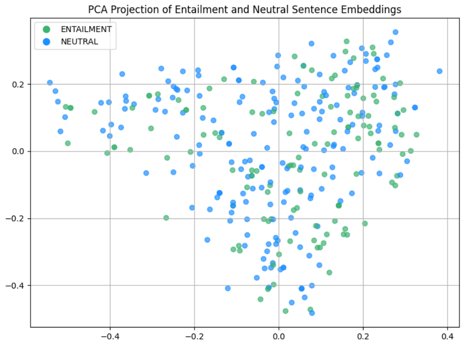

Figure 3: PCA Projection of Entailment and Neutral Sentence Embeddings of LitBank Augmented Dataset

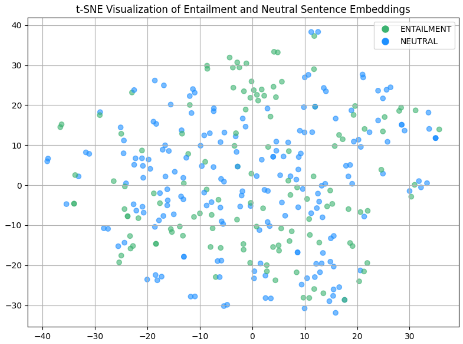

Figure 4: t-SNE Visualization of Entailment and Neutral Sentence Embeddings of LitBank Augmented Dataset

The PCA projection of entailment(green) and neutral(blue) embeddings shows that both categories are evenly distributed across the first two principal components, without strong clustering or skew. The dispersion of points suggests that the model does not encode a dominant pattern for one class over the other in the embedding space. This reinforces the idea that the augmentation process produced a balanced representation between the two classes. The analysis results are shown in Figure 3.

D.3 t-SNE Visualizations

The t-SNE visualizations also demonstrate well-distributed embeddings of entailment(green) and neutral(blue) examples. While there is no sharp boundary between the classes, the fact that the points are broadly and evenly scattered implies that the data captures diverse linguistic and semantic expressions across both classes without introducing structural bias. These findings support the conclusion that the augmentation method preserved semantic variety and avoided overfitting to narrow templates. The analysis results are shown in Figure 4.

D.4 UMAP Representation

The UMAP representation further confirms the even distribution of entailment(green) and neutral(blue) embeddings. Similar to the t-SNE visualizations, the UMAP plot does not show a strict boundary between the two classes, but rather a

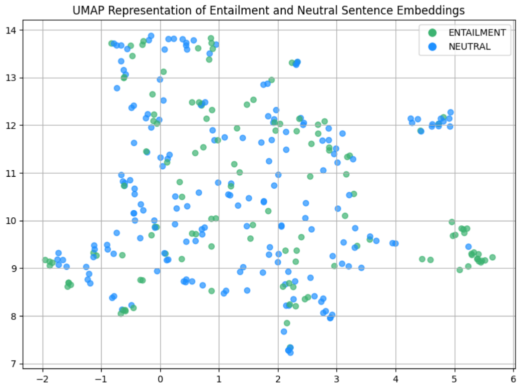

Figure 5: UMAP Visualization of Entailment and Neutral Sentence Embeddings of LitBank Augmented Dataset

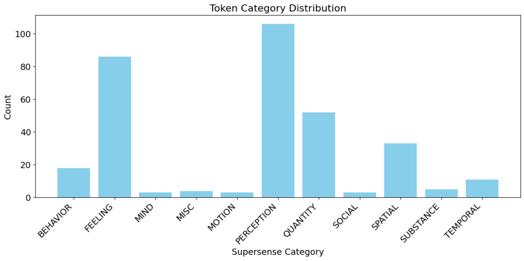

Figure 6: Distribution of inserted modifiers over WordNet supersense categories used during data augmentation

smooth and overlapping layout. This indicates that the semantic space is well-structured and that the data reflects a wide range of sentence-level variations. The consistent dispersion across the embedding space suggests that the augmentation process did not introduce artificial clusters or distortions. The analysis results are shown in Figure 5. In summary, the visual and distributional analyses validate that the augmented dataset is logically sound, semantically rich, and free from major biases, thus providing a strong foundation for downstream model training.

D.5 Supersense Category Analysis

To further analyze the linguistic variety introduced by the augmentation, we categorized the inserted adjectival phrases using supersense labels based on the categorization scheme proposed by the WordNet Domains project (Tsvetkov et al., 2014; Fellbaum, 1998). Figure 6 shows the distribution of token categories. The majority of the added modifiers fall under Perception, Feeling, and Quantity, suggesting that the augmentation effectively introduces human-centric and descriptive features, which are common in literary texts and beneficial for enhancing coreference resolution in such domains.