Pakistan Journal of Language Studies
Vol. 8, Issue 1, 2024
ISSN (Online) 2664-1461
ISSN (Print) 2664-1453

Balancing Rare Linguistic Features in Small Datasets through LLM-Augmented Training: A Case Study on Negation Detection

Yangtian Li

Viterbi School of Engineering, University of Southern California.
Email address: liyangti@usc.edu

Abstract

Rare linguistic features, such as implicit cues, structural negation, or low-frequency modifiers, pose persistent challenges for NLP models due to their sparse presence in training corpora. This study explores how large language models (LLMs) can be leveraged to surface and amplify such underrepresented features through targeted data augmentation. Implicit negation is selected as a representative case, and a two-stage augmentation pipeline is introduced: (1) structurally diverse training samples are generated around rare negation cues, and (2) counterfactual sentence pairs are constructed to suppress background biases and highlight model sensitivity to critical linguistic features such as implicit negation. Experiments using RoBERTa on the CONDAQA dataset demonstrate that LLM-augmented training significantly improves performance on inputs containing implicit and structurally complex negation. These findings suggest that LLMs can serve as controlled augmentation tools for rebalancing rare linguistic phenomena in low-resource settings.

Keywords: Implicit Negation, Data Augmentation, Large Language Models, Linguistic Generalization, Low-Resource Phenomena

1. Introduction

Rare linguistic features—such as implicit cues, structural negation, or domain-specific modifiers—remain a persistent challenge for modern NLP models due to their sparse representation in training corpora. In imbalanced datasets, such features occur far less frequently than dominant linguistic patterns, leading models to rely on superficial signals rather than deeper linguistic structures. This results in limited generalization to semantically complex, low-frequency constructions.

Among these phenomena, implicit negation is particularly illustrative: lacking explicit lexical markers, it typically appears through abstract semantics and syntactic variability, making it difficult for models to identify and generalize. Negation detection, therefore, serves as a concrete and focused setting in which to examine these broader generalization challenges. More importantly, it provides a representative case study for exploring how rare linguistic features can be systematically balanced and amplified through data-centric intervention. Rather than treating negation detection as an end in

itself, this study adopts it as a proxy task to investigate general strategies for balancing rare features in low-resource settings.

Traditional corpora such as the SFU Review Corpus heavily emphasize explicit negation cues (e.g., not, no, never), but exhibit little structural variation. Datasets like CONDAQA offer more syntactic diversity and include implicit negation forms (e.g., hardly, unaddressed), yet the number of such instances remains small, limiting their impact during training. Prior work in data augmentation has attempted to address such imbalance, but many approaches rely on shallow transformations such as synonym substitution, often introducing semantic noise or failing to capture deeper cue structures.

To overcome these limitations, a two-stage data augmentation framework powered by large language models (LLMs) is proposed, with the goal of improving the structural and distributional balance of rare linguistic features within training data. In the first stage, syntactically diverse examples are generated based on implicit negation cues; in the second stage, counterfactual sentence pairs are constructed to suppress background bias and enhance model sensitivity to the target features. Rather than evaluating LLMs as downstream classifiers, this framework is designed to examine their capacity to surface, structure, and balance rare linguistic phenomena through controlled data generation.

Experiments using RoBERTa as the base model and CONDAQA as the testbed show that LLM-augmented training significantly improves model performance on inputs involving complex and subtle negation. These results underscore the broader potential of LLMs as controlled augmentation tools for rebalancing rare linguistic signals in small, imbalanced datasets.

2. Literature Review

Rare linguistic features, such as morphological structures or implicit semantic cues,present unique challenges in natural language understanding. These elements are often underrepresented in standard datasets, leading models to overlook their significance during training. Henning et al. (2023) provide a comprehensive survey on class imbalance in deep learning-based NLP systems, highlighting its adverse effects on generalization and fairness. Hofmann et al. (2021) further demonstrate that standard pretrained language models struggle to interpret morphologically complex or derivational word forms, underscoring a structural sparsity problem in model training. Gururangan et al. (2018) reveal annotation artifacts in NLI datasets that distort the distribution of linguistic features, often favoring surface-level correlations over deeper cues. Collectively, these findings underscore the need for more targeted strategies to address the low frequency and uneven distribution of linguistically rich signals in textual data.

To mitigate data imbalance and enhance model robustness toward rare phenomena, various data augmentation methods have been proposed. Traditional approaches like EDA (Wei & Zou, 2019) and HotFlip (Ebrahimi et al., 2018) introduce lexical-level perturbations, but often lack syntactic or semantic fidelity. Kaushik et al. (2020) propose counterfactual augmentation, in which examples are rewritten with alternate labels while preserving grammaticality and coherence, showing strong benefits in robustness.

More recently, large language models (LLMs) have been used to generate high-quality, diverse training examples(Min et al., 2022; Zhou et al., 2023). Such methods go beyond surface perturbation by generating semantically controlled augmentations, including contrastive or structurally informative variants. Gururangan et al. (2020) further demonstrate that continued pretraining on task-specific data distributions can help models better capture rare or domain-specific patterns. Building on this trend, Dai et al. (2025) introduce *AugGPT*, which leverages ChatGPT to produce paraphrased variants that maintain semantic consistency while increasing structural diversity. Similarly, Qu et al. (2020) propose *CoDA*, a contrast-regularized augmentation framework designed to promote feature-aware diversity through controlled transformations such as back-translation. These approaches collectively emphasize the importance of structure, contrast, and cue-level control for generalizing over rare linguistic features.

Negation, particularly its implicit forms, presents a linguistically rich but low-frequency challenge in NLP. As a prototypical rare linguistic structure, it is well-suited to test how augmentation strategies impact model performance on underrepresented cues. Hossain et al. (2022) offer a systematic analysis of negation in widely-used corpora, revealing how implicit negation cues are often under-annotated and inconsistently modeled. Poliak et al. (2018) and Ravichander et al. (2022) highlight that models can exploit dataset artifacts to perform surprisingly well on negation-sensitive tasks without truly understanding the underlying linguistic structure. Shaitarova and Rinaldi (2021) examine cross-lingual scope resolution in zero-shot settings, suggesting that general-purpose representations often lack sensitivity to negation scope boundaries. Fancellu (2018) investigates multilingual detection of negation scopes, reinforcing the idea that negation remains a structurally elusive phenomenon across languages and data sources. In parallel, Truong et al. (2022) explore a negation-focused pretraining strategy that combines masking and augmentation to improve model sensitivity to subtle forms of negation.

While significant progress has been made in understanding and mitigating data imbalance, current methods still lack a systematic exploration of how LLMs can enhance model sensitivity to rare but linguistically meaningful cues, particularly those that are structurally implicit. This paper addresses this gap by using negation as a case study and

proposing a structured augmentation framework that targets rare features through LLM-based contrastive generation. In doing so, it aims to shed light on how language models can move beyond surface-level token distributions and toward deeper representation of rare linguistic phenomena.

3. Methodology

3.1 The Challenge of Implicit Negation in Imbalanced Corpora

Negation in natural language is realized through a spectrum of linguistic cues. While explicit negation—such as*not*, *never*, and *no*—is syntactically overt and frequently represented in training data, implicit cues tend to be lexically diverse, semantically subtle, and structurally underrepresented. This imbalance causes models to overfit on surface-level expressions and struggle to generalize to semantically equivalent but structurally distinct forms. Prior work has raised similar concerns in the context of rare linguistic phenomena and annotation artifacts in natural language inference datasets (Gururangan et al., 2018; Hofmann et al., 2021).

To examine the impact of such structural skew, this study draws from two corpora with contrasting negation distributions: the SFU Review Corpus (Cruz et al., 2016) and the CONDAQA dataset (Ravichander et al., 2022). The SFU corpus, consisting of over 16,000 annotated review sentences, primarily features explicit negation forms. Approximately 18% of its instances contain negation, with most relying on lexical cues such as not, never, and no. Figure 1 shows the highly skewed cue distribution, indicating limited structural diversity.

In contrast, CONDAQA offers a structurally rich collection of 1,289 contrastive sentence pairs covering a wide range of negation strategies, including implicit forms such as *hardly*, *unaddressed*, and *absence of*. Figure 2 demonstrates a more balanced distribution between explicit and implicit cues. Designed to test deeper reasoning beyond surface-level negation, CONDAQA provides two key utilities: (1) a testbed with diverse structures and cue types, and (2) a template source for guided data generation. Overlap with SFU is deliberately avoided to ensure structural novelty.

ISSN (Online) 2664-1461
ISSN (Print) 2664-1453

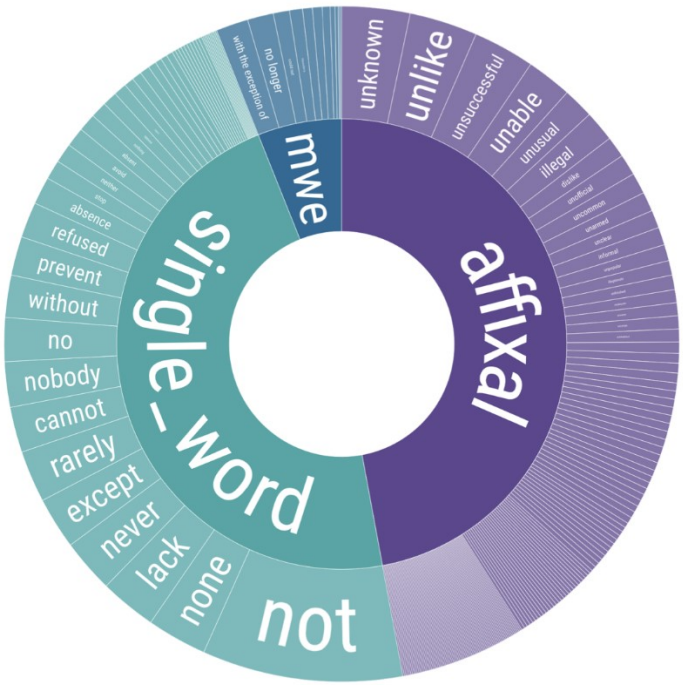

Figure 2. Distribution of negation cues in CONDAQA

To prepare for augmentation, 100 implicit negation cues absent from SFU were sampled from CONDAQA and manually verified. These cues span various semantic domains and serve as seeds for generating 5,000 synthetic sentences using a large language model (details in Methodology). While implicit negation often requires broader context, this setup aims to enrich the structural and semantic space of the training data.

Representative examples illustrate the contrast between the corpora:

Examples from SFU (mostly explicit negation):

- *There is no mystery to the story.*
- *I would definitely not recommend it to anybody.*
- *Damon and Jannie have never been described as anything but angelic.*

Examples from CONDAQA (explicit + implicit):

- *By the time he became emperor he had chronic gout and could hardly walk.*
- *However, Machias Seal Island... was unaddressed by the commission.*
- *...absence from work due to sickness and elevated risk of future disability pension.*

In summary, this section highlights the distributional and structural imbalance, which can be more visibly shown in Table 1, between SFU and CONDAQA and motivates the need for a targeted augmentation strategy. Rather than relying on surface-level rephrasing, the approach introduced here is designed to address structural sparsity in training data through semantic control and cue-level diversity. The next section introduces the augmentation framework in detail.

Table 1

Corpus statistics and negation coverage

Dataset	Size	Negation Type
SFU	16944	18% negation (mostly explicit)
CONDAQA	1289	100% negation (balanced)

3.2 Implicit Cue-Based Augmentation with LLMs

To enrich the training corpus with diverse and underrepresented negation patterns, this study introduces a structured, multi-stage augmentation pipeline leveraging large language models (LLMs). The goal is to expand the training corpus with diverse and underrepresented negation patterns—particularly those involving implicit, low-frequency, and semantically subtle cues. Unlike traditional syntactic augmentation or polarity flipping, this approach targets deep linguistic features that are often absent in standard datasets.

The augmentation process begins with the selection of 100 implicit negation cues manually extracted from the CONDAQA dataset. These cues (e.g., unaddressed, hardly, absence) are chosen under two constraints: they must convey implicit or context-dependent negation, and they must be absent from the SFU Review Corpus. This ensures the injection of novel and semantically challenging constructions into the training set while avoiding overlapping with test data.

To expand the lexical space of negation cues, each selected term is passed to a GPT-4 model, which is prompted to generate five semantically similar alternatives. For example, the cue unlike may yield expansions such as unsubstantiated, untenable, unfathomable, and unprecedented. This step broadens the diversity of linguistic signals without compromising semantic consistency, resulting in 500 additional cues for downstream use.

For each cue—original or expanded—the pipeline then generates ten declarative sentences using GPT-4. Sentence templates drawn from CONDAQA guide the generation, instructing the model to mimic rhetorical structures while shifting the semantic domain (e.g., from media to education or science). Prompting is conducted via a standardized template, with the full prompt specification included in Appendix.

ISSN (Online) 2664-1461
ISSN (Print) 2664-1453

The generation stage yields approximately 5,000 new sentences. These are filtered in two steps: initial grammatical validation using automated tools (e.g., Grammarly API), followed by manual review to verify cue quality and semantic fluency. Only samples passing both stages are retained for fine-tuning.

This process introduces a significant volume of high-quality, structurally diverse implicit negation examples into the training data. Beyond quantity, the augmentation strategy is designed to rebalance underrepresented linguistic features along two dimensions: (1) frequency compensation by increasing the incidence of low-frequency cues, and (2) structural compensation through varied syntactic configurations and domain-specific expressions. Together, these steps aim to strengthen the model's ability to generalize over semantically subtle and structurally implicit negation.

3.3 Counterfactual Augmentation

To enhance the model's sensitivity to polarity-bearing features, this paper applies counterfactual data augmentation across the entire training set. This includes both SFU-originated and LLM-generated samples. Each sentence is treated as a candidate for polarity inversion, regardless of its original label. Two complementary strategies are employed: one based on a self-supervised rationale rewriting framework, and another guided by large language models (LLMs).

3.3.1 Self-supervised Counterfactual

The self-supervised approach, shown in figure 3, is adapted from the work of Plyler1 et al.(2021), which proposes an automated rationale-based framework for counterfactual generation. Given a sentence labeled with its sentiment polarity, the model first identifies minimal spans of tokens—rationales—that are causally associated with the original label. It then modifies these spans using a masked language model (MLM) to generate a contrastive version that flips the sentence's polarity while preserving fluency and syntactic structure.

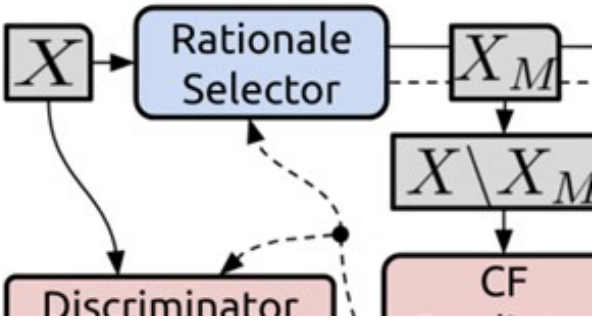

Figure 3. Counterfactual Data Generation Workflow

While this strategy is efficient and annotation-free, it may suffer from limited syntactic diversity and a tendency to rely on explicit negation markers such as not or neverin polarity reversal, which could constrain its ability to model nuanced semantic phenomena.

3.3.2 LLM-based Counterfactual Generation

The second augmentation strategy involves generating polarity-reversed counterparts for all sentences in the merged dataset (approximately 13,000 sentences after combining SFU and LLM-generated augmentation). This process enables us to evaluate whether large-scale polarity transformation can improve the model's understanding of negation at both lexical and structural levels.

To ensure semantic consistency, structural coherence, and lexical diversity in the generated counterfactuals, a refined rationale-guided rewriting strategy, inspired by Plyler1 et al.(2021), was adopted. The large language model is instructed to follow a constraint-based, three-step transformation process:

- Extract Rationale: Select approximately 15–20% of the sentence (not necessarily contiguous), focusing on the polarity-bearing expression. The rationale must include the original negation cue if present.
- Rewrite Rationale: Rephrase the rationale into its fully affirmative form, strictly removing all negation cues and introducing no new ones. The transformation must preserve the semantic intent while altering the polarity.
- Reintegrate: Embed the rewritten rationale back into the original sentence with minimal changes to the remaining structure. Additionally, the output was required to stay within ±10% of the original sentence length and maintain fluency.

An example of this controlled rewriting is shown below:

- *Original*: This was later adopted in Ancient Greece as the "gamos" and"engeysis" rituals, although unlike in Judaism the contract made in front of witness was only verbal.
- *LLM-generated CF*: This was later adopted in Ancient Greece as the "gamos" and "engeysis" rituals, similar to Judaism, the contract made in front of witnesses was simply oral.

This approach guides the LLM to minimally alter input sentences and produce counterfactuals that are semantically faithful, linguistically fluent, and structurally well-

formed. By enforcing explicit constraints, the method yields high-quality sentence pairs with clearly aligned contrastive polarity.

This rewriting procedure was applied consistently across the entire augmented training corpus, including both SFU-originated and LLM-generated sentences. As a result, each training instance is paired with a polarity-inverted counterpart, providing a rich and balanced contrastive signal for downstream learning.

3.3.3 Experimental Setup

To evaluate the proposed augmentation framework, six experimental configurations were designed using RoBERTa-base as the primary model. Each configuration manipulates the training data composition to assess the impact of feature diversity and counterfactual supervision.

- SFU: 60% of the original SFU corpus, consisting of 3,117 negated and 7,049 affirmative examples.
- SFU + LLM: SFU baseline combined with 5,000 LLM-generated examples constructed using implicit negation cues.
- SFU + LLM + CF (SS): Same as 2, with additional counterfactual examples generated via a self-supervised rationale-based method.
- SFU + LLM + CF (GPT-4): Same as 2, but counterfactuals are generated using GPT-4 guided rewriting.
- SFU + Not-inserted: SFU corpus augmented with 5,000 synthetic examples created by inserting simple negation cues (e.g., “not”).
- NegBERT: Directly evaluating a publicly released pretrained NegBERT model on the CONDAQA test set, without any fine-tuning.

All RoBERTa-based models are fine-tuned using the HuggingFace Transformers library. This paper adopts standard training settings across all experimental groups and report average performance over three random seeds. Training is conducted on a single NVIDIA RTX 4090 GPU with mixed-precision enabled. Early stopping is applied based on validation F1 score.

The base training set consists of 10,166 SFU samples, including all negated sentences and a balanced number of affirmative ones. 5,000 LLM-generated examples emphasizing implicit negation cues were added. For counterfactual experiments, each training instance—original or synthetic—is paired with a polarity-inverted variant, resulting in balanced distributions across both label and cue types.

All models are evaluated on the CONDAQA dataset, which contains 2,578 test sentences evenly split between negated and affirmative forms. To prevent data leakage, none of the test set negation cues appear in the augmented training data. Evaluation uses

standard classification metrics: Accuracy, Precision, Recall, and F1 score. Results are reported as averages over three independent runs.

4. Results and Analysis

The results of our fine-tuning experiments were reported using the RoBERTa model on different data augmentation settings. Evaluation is conducted on the CONDAQA-derived test set, which contains 1,289 pairs of negated and affirmative sentences. Accuracy, F1 score, and AUC are used as metrics. Results are shown in Table 2.

The baseline RoBERTa model without fine-tuning performs at near-chance levels, highlighting its inability to capture negation-specific patterns. Fine-tuning on the SFU corpus boosts performance (60.76% Accuracy, 0.5792 F1), but its explicit cue bias limits generalization to more nuanced forms of negation. Adding 5,000 LLM-generated examples (SFU+LLM) that represent implicit negation significantly improves performance (67.31% Accuracy, 0.6712 F1), confirming the value of exposing the model to structurally diverse, feature-targeted examples.

Table 2

Model performance comparison on negation detection (evaluated on CONDAQA)

	Accuracy	F1	AUC
RoBERTa (w/o FT)	50.06%	0.3340	0.3484
SFU	60.76%	0.5792	0.6379
SFU + LLM	67.31%	0.6712	0.7467
SFU + LLM + CF (SS)	68.15%	0.6805	0.7734
SFU + LLM + CF (GPT4)	75.20%	0.7513	0.7855
NegBERT	60.17%	0.5932	0.6745
SFU + Not	60.56%	0.5774	0.6483

Further gains are achieved by incorporating counterfactual (CF) augmentation. The self-supervised variant (SFU+LLM+CF(SS)) improves upon the LLM baseline by a small but consistent margin (0.6805 F1), suggesting that even automatically generated contrastive examples can help guide the model's attention toward polarity-sensitive spans. However, the benefits appear limited, potentially due to the relatively shallow transformations and reliance on surface-level cues observed in self-supervised counterfactuals.

The highest performance is achieved using GPT-4-generated counterfactual data (SFU+LLM+CF(GPT4)) which are 75.20% Accuracy, 0.7513 F1, and 0.7855 AUC. This configuration demonstrates the effectiveness of combining feature-focused augmentation with semantically consistent polarity shifts, especially for underrepresented linguistic

ISSN (Online) 2664-1461
ISSN (Print) 2664-1453

features. Compared to the self-supervised method, GPT-4 counterfactuals introduce greater structural variation and semantic fluency, contributing to superior generalization.

Notably, simpler strategies like random "not" insertion or the pretrained NegBERT baseline fail to achieve competitive results. These methods lack semantic control, underscoring that negation detection is a feature-centric task, heavily influenced by the quality and variety of negation cues during training.

Interestingly, training solely on the LLM+CF dataset results in an unusually high accuracy (approaching 90%) on the CONDAQA test set. However, based on preliminary evaluations on the SFU corpus (not shown in tables), a substantial performance drop was observed, which is comparable to a non-finetuned RoBERTa. This discrepancy suggests overfitting: the model likely memorizes surface-level negation cues present in the synthetic data, without learning a generalized notion of negation. These findings underscore the importance of cross-domain diversity and cue subtlety in building robust negation detectors.

In summary, the results validate our core hypothesis: augmenting training data with LLM-guided, structurally diverse examples is key to enhancing model sensitivity toward subtle and rare negation patterns. Moreover, the comparison between CF variants highlights that not only the presence of contrastive supervision, but also its semantic and syntactic quality, is crucial for improving generalization.

These findings align with the core objective of this study: to enhance model sensitivity toward underrepresented, implicitly encoded linguistic cues through structured, LLM-driven data augmentation.

5. Conclusion and Future Work

This paper proposes a feature-oriented data augmentation framework based on large language models (LLMs) to address the long-standing issue of inadequate representation of rare linguistic features in small, imbalanced datasets. Using negation detection as a diagnostic task, the study demonstrates that LLMs can generate high-quality training samples through implicit cue expansion and structured counterfactual rewriting, significantly enhancing the model's ability to identify and generalize context-sensitive linguistic features such as implicit negation.

Experimental results indicate that RoBERTa models trained with LLM-driven augmentation achieve considerable improvements in detecting implicit negation, underscoring the effectiveness of structurally rich and semantically nuanced training data over traditional augmentation methods. Moreover, this research highlights a critical issue in existing datasets, where explicit negation expressions vastly outnumber implicit ones.

Uncontrolled merging of such data can inadvertently dilute the representation of implicit cues, a concern future dataset construction should carefully avoid.

Future work can extend the proposed method to other linguistically sparse phenomena, such as modality, coreference resolution, or figurative language, which similarly require targeted structural and semantic augmentation beyond surface-level representations. Additionally, adapting these augmentation strategies to cross-domain and multi-genre corpora could further enhance model generalization in practical applications. Finally, it is worth investigating the impact of feature-balanced datasets on broader downstream tasks, such as sentiment analysis, question answering systems, and fact verification, promising significant research and practical value.

References

Cruz, N. P., Taboada, M., & Mitkov, R. (2016). A machine-learning approach to negation and speculation detection for sentiment analysis. *Journal of the Association for Information Science and Technology*, *67*(9), 2118–2136. https://doi.org/10.1002/asi.23533

Dai, H., Liu, Z., Liao, W., Huang, X., Cao, Y., Wu, Z., … Li, X. (2025). AugGPT: Leveraging ChatGPT for Text Data Augmentation. *IEEE Transactions on Big Data*, *11*(3), 907–918. https://doi.org/10.1109/TBDATA.2025.3536934

Ebrahimi, J., Rao, A., Lowd, D., & Dou, D. (2018, May 24). *HotFlip: White-Box Adversarial Examples for Text Classification*. arXiv. https://doi.org/10.48550/arXiv.1712.06751

Fancellu, F. (2018). *Computational models for multilingual negation scope detection* (University of Edinburgh). University of Edinburgh, Great Britain. Retrieved from http://hdl.handle.net/1842/33038

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020, May 5). *Don't Stop Pretraining: Adapt Language Models to Domains and Tasks*. arXiv. https://doi.org/10.48550/arXiv.2004.10964

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S. R., & Smith, N. A. (2018, April 16). *Annotation Artifacts in Natural Language Inference Data*. arXiv. https://doi.org/10.48550/arXiv.1803.02324

Henning, S., Beluch, W., Fraser, A., & Friedrich, A. (2023, February 22). *A Survey of Methods for Addressing Class Imbalance in Deep-Learning Based Natural Language Processing*. arXiv. https://doi.org/10.48550/arXiv.2210.04675

Hofmann, V., Pierrehumbert, J. B., & Schütze, H. (2021, June 2). *Superbizarre Is Not Superb: Derivational Morphology Improves BERT's Interpretation of Complex Words*. arXiv. https://doi.org/10.48550/arXiv.2101.00403

Hossain, M. M., Chinnappa, D., & Blanco, E. (2022, March 16). *An Analysis of Negation in Natural Language Understanding Corpora*. arXiv. https://doi.org/10.48550/arXiv.2203.08929

Kaushik, D., Hovy, E., & Lipton, Z. C. (2020, February 14). *Learning the Difference that Makes a Difference with Counterfactually-Augmented Data*. arXiv. https://doi.org/10.48550/arXiv.1909.12434

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., & Zettlemoyer, L. (2022, October 20). *Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?* arXiv. https://doi.org/10.48550/arXiv.2202.12837

Plyler, M., Green, M., & Chi, M. (2021). Making a (Counterfactual) Difference One Rationale at a Time. *Advances in Neural Information Processing Systems*, *34*, 28701–28713. Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2021/hash/f0f800c92d191d736c4411f3b3f8ef4a-Abstract.html

Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R., & Durme, B. V. (2018, May 2). *Hypothesis Only Baselines in Natural Language Inference*. arXiv. https://doi.org/10.48550/arXiv.1805.01042

Qu, Y., Shen, D., Shen, Y., Sajeev, S., Han, J., & Chen, W. (2020, October 16). *CoDA: Contrast-enhanced and Diversity-promoting Data Augmentation for Natural Language Understanding*. arXiv. https://doi.org/10.48550/arXiv.2010.08670

Ravichander, A., Gardner, M., & Marasović, A. (2022, November 1). *CONDAQA: A Contrastive Reading Comprehension Dataset for Reasoning about Negation*. arXiv. https://doi.org/10.48550/arXiv.2211.00295

Shaitarova, A., & Rinaldi, F. (2021). Negation typology and general representation models for cross-lingual zero-shot negation scope resolution in Russian, French, and Spanish. In E. Durmus, V. Gupta, N. Liu, N. Peng, & Y. Su (Eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop* (pp. 15–23). Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-srw.3

Truong, T. H., Baldwin, T., Cohn, T., & Verspoor, K. (2022, May 9). *Improving negation detection with negation-focused pre-training*. arXiv. https://doi.org/10.48550/arXiv.2205.04012

Wei, J., & Zou, K. (2019, August 25). *EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks*. arXiv. https://doi.org/10.48550/arXiv.1901.11196

Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., … Levy, O. (2023). LIMA: Less is more for alignment. *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 55006–55021. Red Hook, NY, USA: Curran Associates Inc.

Appendix A

Prompt Design

Prompt 1: Generating Sentences with Implicit Negation

The following prompt was used to guide the generation of declarative sentences containing implicit negation cues from the CONDAQA corpus:

> You are now playing the role of a professional writing assistant. Your task is to help construct well-written declarative sentences that showcase linguistic features related to negation.
>
> ✧ Below is a source sentence that uses the implicit negation cue 'unlike':
>
> *This was later adopted in Ancient Greece as the "gamos" and "engeysis" rituals, although unlike in Judaism the contract made in front of witnesses was only verbal.*
>
> - Please generate 10 new declarative sentences that retain a similar syntactic and rhetorical structure but shift to different topics or domains (e.g., law, science, education, or culture).
> - In each sentence, replace 'unlike' with a semantically similar implicit negation cue (e.g., unsubstantiated, untenable, unfathomable, unprecedented).
> - Use formal language, and ensure each sentence is coherent and contextually meaningful.

Prompt 2: Counterfactual Rewriting (Negated → Affirmative)

This prompt was used to guide LLM-based counterfactual rewriting of negated sentences:

> Hi, chatgpt. Please help me to transform a sentence containing the negation cue 'unlike' into a fully affirmative version under the following constraints:
>
> - Character count of the result must be within ±10% of the original.
> - 15–20% of the original words must be changed (via replace, delete, or insert).
> - Remove all negation markers. Do not introduce new negation.
> - Output must be fluent and natural.

For any input, reply with:

1. Original metrics (character and word count)

2. Changed tokens (words replaced/added/removed)

3. Final rewritten sentence

Example input:
This was later adopted in Ancient Greece as the "gamos" and "engeysis" rituals, unlike in Judaism the contract made in front of witness was only verbal.

Example output:
This was later adopted in Ancient Greece as the "gamos" and "engeysis" rituals, similar to Judaism, the contract made in front of witnesses was simply oral.

Your help are deeply appreciated.

Prompt 3: Counterfactual Rewriting (Affirmative → Negated)

This prompt was used to guide LLM-based counterfactual rewriting of affirmative sentences:

Hi ChatGPT, please create a counterfactual version of the following affirmative sentence under these rules:

✧ Character count of the result must be within ±10% of the original.

✧ 15–20% of the original words must be changed (via replace, delete, or insert).

✧ The result should remain negated, but do not introduce any explicit "no","not" or similarexplicit structural negation tokens.

✧ Output must be fluent and natural.

For any input, reply with:

1. Original metrics (character and word count)

2. Changed tokens

3. Final sentence

Example input:
She always arrives early to every meeting.

ISSN (Online) 2664-1461
ISSN (Print) 2664-1453

Example output:
She hardly comes early to every meeting.

Appendix B

Hyper parameter

Table B1

Training hyperparameters used across all RoBERTa-based models

Hyperparameter	Value
Epochs	20
Batch size	16
Optimizer	AdamW
Learning rate	2e-5
Schedule	Linear decay
Early stopping	F1-based (patience = 2)
Precision	Mixed-precision training