

# How Adversarial is BioNLI?

Shiv Kanani, Adarsh Bharti, Vedamsh Ganta

December 6th, 2024

## 1 Project Overview

In biomedical decision-making, logical connections between hypotheses and experimental data are crucial. Natural Language Inference (NLI), a subset of natural language processing, facilitates this. However, dataset biases and challenges in generating relevant instances hinder the evaluation of NLI models, especially under adversarial conditions. This project analyzes the BioNLI dataset and models to assess their robustness and limitations by identifying dataset biases, and benchmarking existing models.

Previous research, including BioNLI, evaluated model performance using fixed perturbations and adversarial datasets generated through neural and rule-based methods. However, these approaches do not address unobserved or domain-shifted adversarial categories (e.g., using chemistry terms in biological contexts) or dataset biases like sentence length and lexical overlap. These gaps necessitate further examination.

To address these issues, we propose three approaches: benchmarking model performance on the BioNLI dataset, analyzing dataset properties for Distributional Dataset Bias (e.g., label distribution) and Qualitative Dataset Bias (e.g., lexical overlap, sentence length), and testing adversarial robustness. Pre-trained models like BioLinkBERT were evaluated using metrics such as accuracy, F1 score, and recall, with analyses performed using tools like Hugging Face Transformers and Python.

Key findings include BioLinkBERT's strong baseline performance (83% accuracy, 0.792 F1 score) on the BioNLI dataset. However, biases such as class imbalance (80% of test instances labeled as contradictions), shorter sentences, and high lexical overlap disproportionately improved model accuracy. These results highlight the need for further testing and dataset augmentation to enhance robustness beyond dataset-specific patterns.

## 2 Ideas

### 2.1 Idea 1: Benchmarking BioLinkBERT on BioNLI with adversarial examples included.

**Objective:** Evaluate the performance of the pretrained BioLinkBERT model on the BioNLI dataset, which contains biomedical premise-hypothesis pairs labeled as entailment or contradiction. The goal is to fine-tune BioLinkBERT and analyze its performance using metrics like accuracy and F1 score.

#### Approach:

- **Model Selection:** BioLinkBERT, a domain-specific BERT-based model, is pretrained on biomedical corpora, making it suitable for NLI tasks in this domain.
- **Fine-tuning:**
  - BioLinkBERT is fine-tuned using PyTorch with a classification head for binary classification.
  - The tokenized data is loaded into PyTorch Datasets and passed to DataLoaders.
  - Training involves cross-entropy loss, AdamW optimization, and a linear learning rate scheduler.
  - Validation is performed after each epoch to ensure generalization, guiding hyperparameter tuning.

- **Testing:** Post-training, the model is evaluated on the test set using metrics such as accuracy, precision, recall, and F1 score.

## 2.2 Idea 2 & 3: Dataset Bias Analysis

Objective: Identify patterns or biases in the BioNLI dataset that may impact model performance and lead to unfair or inaccurate predictions.

### Key steps:

- **Analyzing Dataset Features:**
  - **Class Distribution:** 80% of test data is labeled as contradiction, creating an imbalance that may bias the model toward predicting contradiction.
  - **Lexical Overlap:** High word overlap between premises and hypotheses often leads to better accuracy, indicating the model may over-rely on surface-level similarities.
  - **Sentence Length:** Examined how the number of words in premises and hypotheses affects accuracy, finding shorter sentences often yields better results.
- **Impact on Model Performance:** Models performed better on sentences with high lexical overlap and favored the majority class (contradiction), even when entailment was correct.
- **Visualizing Trends:** Plots illustrating accuracy variations with sentence length and lexical overlap highlighted dataset biases.
- **Significance:** Recognizing these biases helps refine future datasets and models to improve fairness and reliability.

## 3 Experimental Setup

### 3.1 Idea 1: Basic Benchmarking

- **Model used: BioLinkBERT**
  - **Type:** Transformer-based model, a variation of BERT pre-trained on biomedical corpora.
  - **Layers:** 12 layers (base version of BERT architecture).
  - The model is pre-trained on biomedical and scientific text, then fine-tuned on the **BioNLI dataset** in this project.
- **Training Settings:**
  - **Batch Size:**
    - Training: 4
    - Validation and Testing: 8
  - **Number of Epochs:** 3
  - **Optimizer:** AdamW with a learning rate of  $2e-5$ .
  - **Scheduler:** Linear decay scheduler without warmup steps.
  - **Loss Function:** Cross-Entropy Loss for binary classification.
  - **Training Time:** Approximately 10 minutes per epoch on a T4 GPU.
- **Fine-tuning Modifications:**
  - Added special tokens (**<re>**, **<el>**, etc.) to handle biomedical text structure.
  - Fine-tuned the top three transformer layers and the classification head, freezing other layers for efficiency.
- **Dataset Used:** BioNLI, a biomedical natural language inference dataset.
- **Input:** Premises (supporting biomedical sentences) and hypotheses (conclusions), tokenized and encoded as model inputs.
- **Output:** Binary labels (**1** for positive entailment, **0** for non-entailment or adversarial examples).
- **Training Instances:** Approx. 50,000 (including generated adversarial examples).
- **Development Instances:** Approx. 10,000.

- **Test Instances:** Approx. 10,000.
- Generated adversarial samples include:
  - Swapping entities, numbers, or lexical terms.
  - Flipping sentence polarity.
  - Adding out-of-context entities.
- These increase dataset diversity and test model robustness.
- **Metrics Used:**
  - **Precision:** Proportion of correct positive predictions out of all positive predictions made by the model.
  - **Recall:** Proportion of actual positives correctly identified by the model.
  - **F1-Score:** Harmonic mean of precision and recall, balancing their trade-offs.
  - **Macro F1-Score:** Average F1-Score across classes, accounting for class imbalance.
  - **Accuracy:** Overall fraction of correct predictions.
  - **Confusion Matrix:** Displays true positives, false positives, false negatives, and true negatives for detailed analysis.

## Summary of Results

- The **BioLinkBERT** model achieves excellent performance on clean biomedical text, with high precision and F1-scores.
- Adversarial examples slightly reduce performance, but the model remains robust, achieving a Macro F1-Score of **0.79** on adversarial data.
- Fine-tuning with adversarial examples enhances the model's generalizability and robustness.

## 3.2 Ideas 2 & 3 Dataset Bias Analysis

- **Model Used:**
  - Same BioLinkBERT model as in Idea 1, trained and fine-tuned on the BioNLI dataset.
  - No additional training for this idea; the focus is on evaluating model performance in relation to dataset biases.
- **Dataset Used:** BioNLI, the same dataset as Idea 1.
- **Additional Preprocessing:**
  - Calculated **sentence lengths** for premises and hypotheses.
  - Computed **lexical overlap** (cosine similarity) between premises and hypotheses using **CountVectorizer**.

### Bias Categories Explored:

- **Distributional Bias:**
  - Investigated label/class distribution to identify imbalances in positive (entailment) vs. negative (non-entailment) samples.
- **Qualitative Bias:**
  - **Lexical Overlap:** Measured word overlap between premises and hypotheses.
  - **Sentence Length:** Analyzed the impact of premise and hypothesis lengths on model performance.

### Metrics Used:

- **Accuracy:** Measures the proportion of correct predictions.
- **Precision:** Focuses on the fraction of true positives out of all predicted positives.
- **Recall:** Examines how many actual positives were identified correctly.
- **F1-Score:** Harmonic mean of Precision and Recall, balancing their trade-offs.
- **Confusion Matrix:** Summarizes true positives, false positives, false negatives, and true negatives.

### Results/Findings:

### 1. Distributional Bias:

- The dataset exhibited some class imbalance:
  - Entailment (positive class) was underrepresented compared to negation (negative class).
  - This imbalance could impact model performance, especially in recall for the minority class.

### 2. Qualitative Bias:

- **Lexical Overlap:**
  - Higher lexical overlap between premises and hypotheses correlated with improved accuracy.
  - Suggests that the model relies heavily on shared words to make predictions.
- **Sentence Length:**
  - Sentence length affected model performance:
    - Extremely short or very long premises and hypotheses were associated with lower accuracy.
    - Indicates that the model may struggle with complex or sparse input.

## 4 Results

### 4.1 Idea 1: Basic Benchmarking

#### Main Results:

- The BioLinkBERT model was fine-tuned on the BioNLI dataset for binary classification and evaluated on training, validation, and test sets. Below is a summary of the results:

Metric	Class 0 (Majority)	Class1 (Minority)	Overall
Precision	0.96	0.58	0.88
Recall	0.82	0.88	0.83
F1-Score	0.88	0.70	N/A
Macro F1-Score	N/A	N/A	0.79
Accuracy	N/A	N/A	0.83

#### Observations:

##### Class-Specific Metrics:

- **Class 0 (Majority):** High Precision (0.96) and Recall (0.82), resulting in an F1-Score of 0.88.
- **Class 1 (Minority):** Balanced Precision (0.58) and high Recall (0.88), resulting in an F1-Score of 0.70.

##### Overall Metrics:

- **Accuracy:** 83%
- **Macro F1-Score:** 0.79 (average F1-Score for both classes).

**Thoughts:** The model achieved its best performance in Epoch 3, with a Macro F1-Score of 0.79 and an overall accuracy of 83%. It demonstrated strong precision (0.96) and F1-Score (0.88) for the majority class (Class 0), while maintaining high recall (0.88) for the minority class (Class 1). Although precision for Class 1 was

moderate (0.58), the model effectively balanced predictions across both classes, reflecting its robustness and ability to handle adversarial examples.

---

## 4.2 Idea 2 & 3: Dataset Bias Analysis

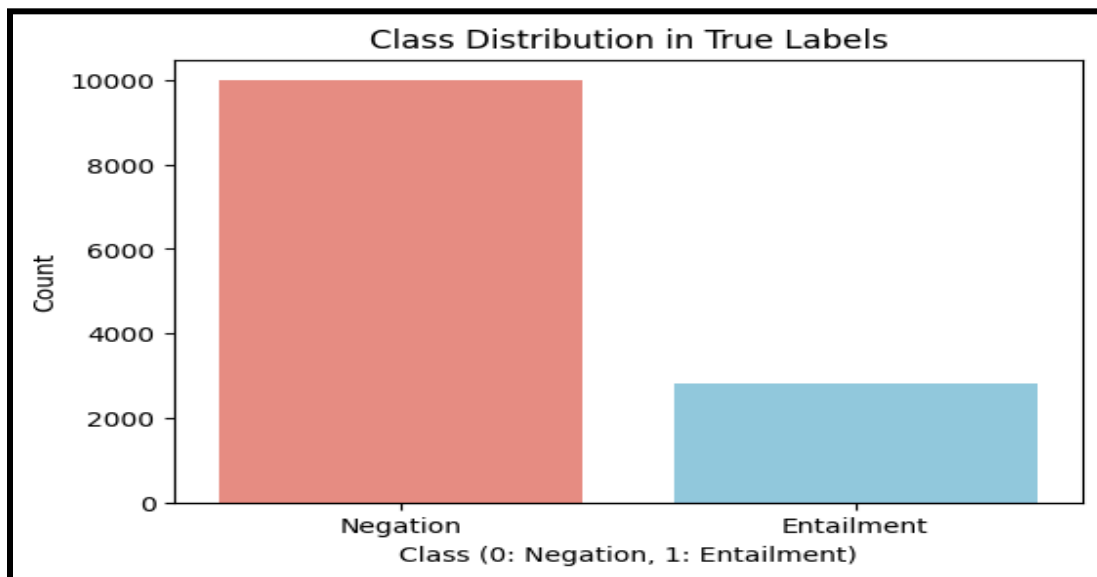
### 4.2.1 Class Distribution

#### Results:

- **Negation (Class 0):** 77.98% of the dataset.
- **Entailment (Class 1):** 22.02% of the dataset.
- Class imbalance favors **negation**, with nearly 4x more examples than entailment.

#### Analysis:

- This imbalance may bias the model towards predicting the majority class (negation), affecting recall for entailment (Class 1).
- **Confusion Matrix:**
  - High true positives (8183) for negation but lower true positives (2487) for entailment.
  - Precision for entailment is only 0.58, highlighting difficulty in predicting the minority class.



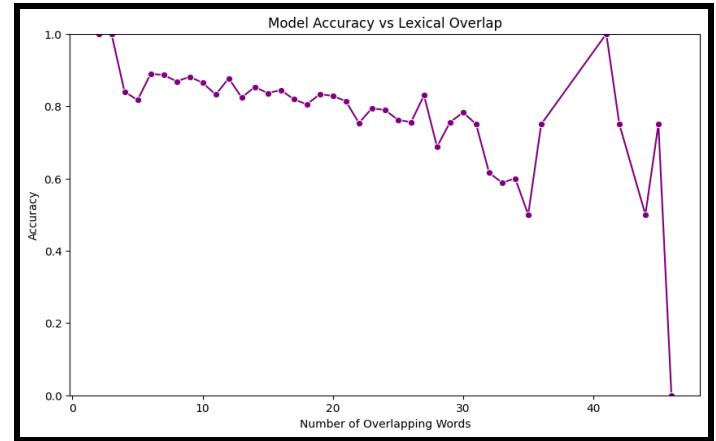
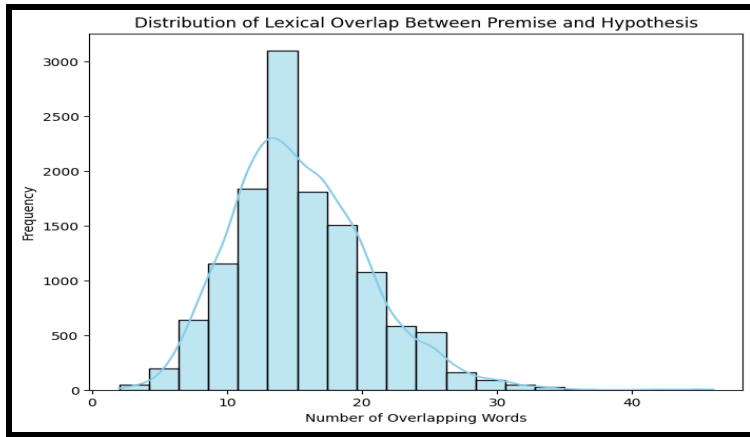
### 4.2.2 Lexical Overlap Impact

#### Results:

- Average accuracy across all overlap levels: **77%**.
- Higher overlap improves accuracy:
  - Overlap = 2 or 3: **100% accuracy**.
  - Overlap = 4 to 6: Accuracy ranges from **81.67% to 88.89%**.

#### Analysis:

- The model relies heavily on shared words between the premise and hypothesis to make predictions.
- Limited lexical overlap (e.g., < 4) correlates with reduced accuracy, indicating a potential reliance on surface-level features rather than semantic understanding.



### 4.2.3 Sentence Length Impact

#### Premise Length:

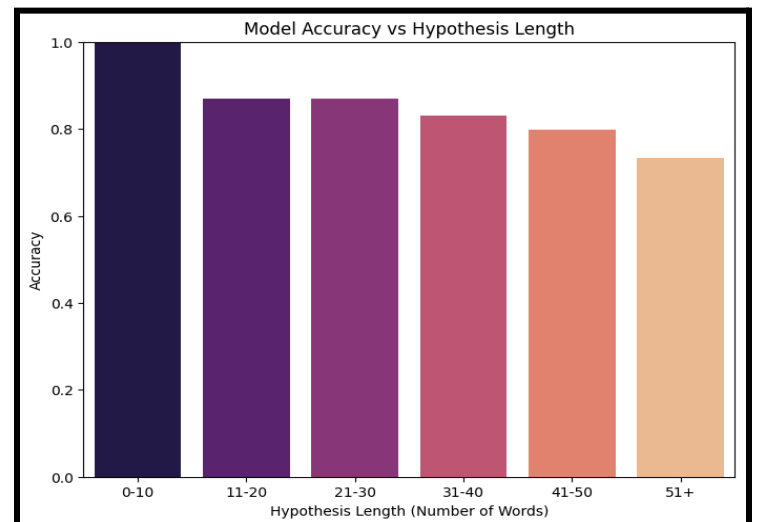
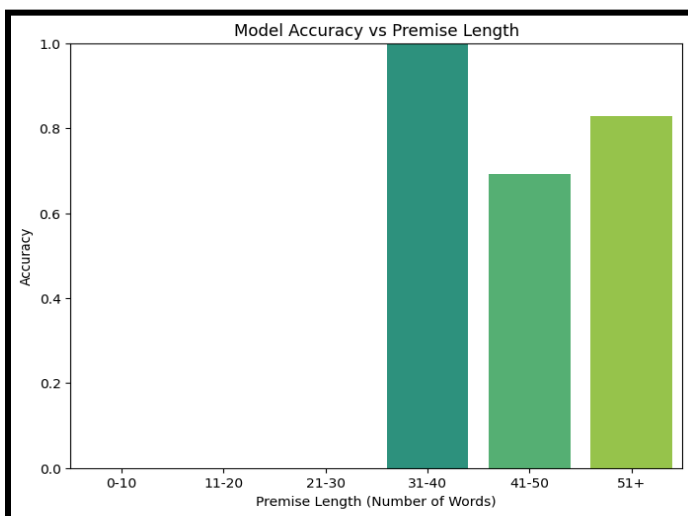
- Binned Accuracy:
  - Premises of 31–40 words: **100% accuracy**.
  - Premises of 41–50 words: **69.23% accuracy**.
  - Longer premises (51+): **82.82% accuracy**.

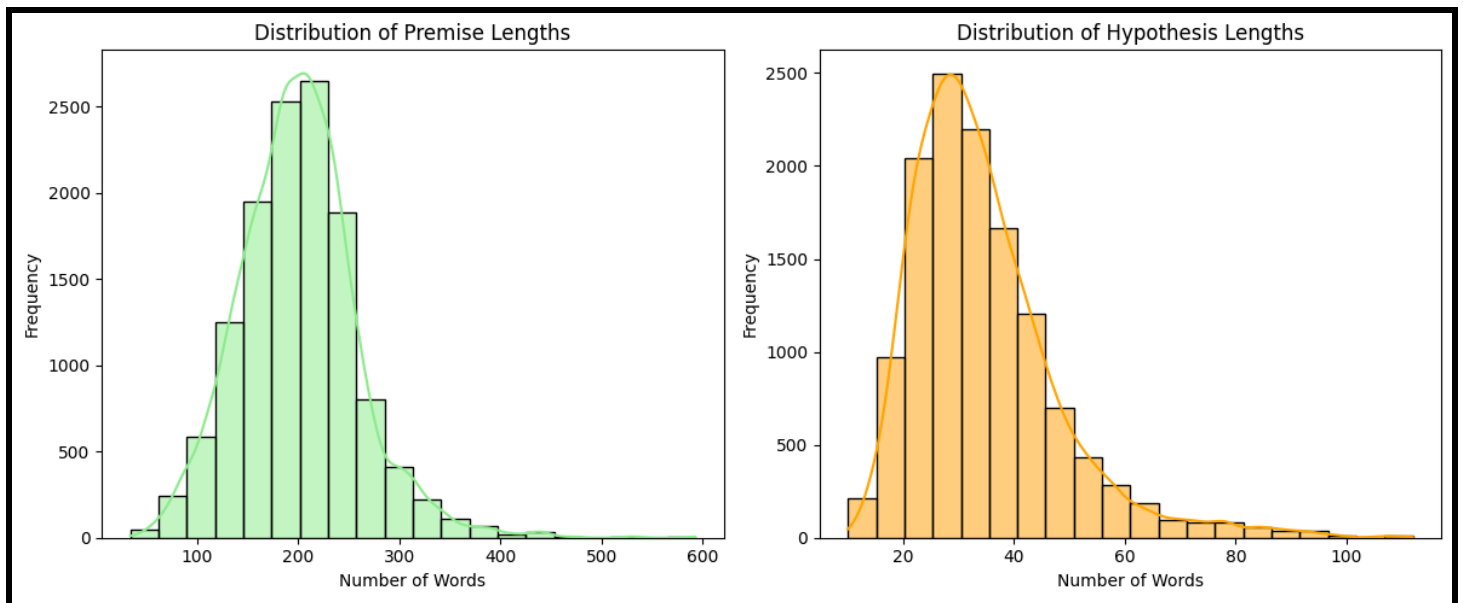
#### Hypothesis Length:

- Binned Accuracy:
  - Hypotheses of 0–10 words: **100% accuracy**.
  - Hypotheses of 31–40 words: **83.04% accuracy**.
  - Longer hypotheses (51+): **73.22% accuracy**.

#### Analysis:

- The model performs best with moderate-length inputs, struggling with very short or excessively long premises and hypotheses.
- Longer sentences may introduce irrelevant details or complexity, reducing model accuracy.





#### 4.2.4 Key Inferences

- **Biases Identified:**
  - **Class Imbalance:** Negation dominates the dataset, affecting recall for entailment.
  - **Lexical Overlap:** High overlap correlates with better accuracy, indicating reliance on surface-level features.
  - **Sentence Length:** The model performs best on moderate-length premises and hypotheses, struggling with extremes.
- **Recommendations:**
  - Balance the dataset to improve entailment recall.
  - Introduce adversarial examples with minimal lexical overlap to encourage deeper semantic understanding.
  - Use longer contexts sparingly and ensure relevant information dominates.

#### 4.2.5 Hypotheses and Testing

1. **Hypothesis 1:** The model struggles with long premises requiring evidence integration.
  - Test: Simplified premises by extracting relevant evidence.
  - Result: Correctly predicted entailment, validating the hypothesis.
2. **Hypothesis 2:** The model underperforms on long, complex, or ambiguous hypotheses.
  - Test: Shortened hypotheses to focus on core claims.
  - Result: Correct predictions, supporting the hypothesis.
3. **Hypothesis 3:** The model relies on surface-level patterns and struggles with low lexical overlap.
  - Test: Paraphrased hypotheses to increase overlap.
  - Result: No change in prediction, suggesting semantic reasoning is needed.


### 5 Summary of Findings

The manual analysis revealed several insights into the BioLinkBERT model's reasoning abilities:

1. **Strengths:**
  - The model performs well when premises and hypotheses are concise and lexically aligned.
  - It demonstrates reasonable understanding of biomedical concepts in simpler contexts.
2. **Weaknesses:**

- **Handling Complexity:** Struggles with lengthy, multi-faceted premises that require integrating multiple pieces of evidence.
  - **Ambiguity:** Long or ambiguous hypotheses pose a challenge for semantic reasoning.
  - **Semantic Reasoning:** Minimal lexical overlap leads to failure in cases requiring deeper semantic alignment.
3. **Recommendations:**
- **Data Augmentation:** Introduce more training examples with long premises and complex hypotheses to improve generalization.
  - **Attention Mechanisms:** Enhance attention mechanisms to better capture relationships across long texts.
  - **Preprocessing:** Consider automatic extraction of relevant portions of premises to reduce complexity.
  - **Semantic Models:** Explore integrating semantic models (e.g., entailment reasoning modules) to handle low-overlap cases.

## 6 Code

Link to google drive with code files for idea1, 2 & 3. It also contains the dataset used along with our saved model: (smart chip:  354 FINAL PROJECT )

(link: <https://drive.google.com/drive/folders/1Zbq-pxpLFGgeeU9uM9BYHjuTZxUphHyn?usp=sharing>)

## 7 Contributions

Adarsh Bharti - Slides, debugging and project report

Vedamsh Ganta - Project report and debugging

Shiv Kanani - Debugging and coding and project report

## 8 References:

1. Mohaddeseh Bastan, Mihai Surdeanu, and Niranjan Balasubramanian. 2022. [BioNLI: Generating a Biomedical NLI Dataset Using Lexico-semantic Constraints for Adversarial Examples](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5093–5104, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
  2. Yasunaga, Michihiro, et al. *LinkBERT: Pretraining Language Models with Document Links*. arXiv:2203.15827, arXiv, 29 Mar. 2022. *arXiv.org*, <https://doi.org/10.48550/arXiv.2203.15827>.
-