

Analyzing Robustness and Context Use in Clinical Natural Language Inference

Stanford CS224N Custom Project

Ryan Le

Department of Computer Science
Stanford University
ryanmtle@stanford.edu

Abstract

In this project, we aim to 1) benchmark a range of models on clinical NLI (MedNLI) to establish strong baselines and 2) analyze whether high-performing models actually reason over clinical context or exploit surface-level details in the hypothesis. For the first part, we evaluate seven models on MultiNLI and MedNLI. These include bag-of-word classifiers, frozen and fine-tuned BERT-style cross encoders, biomedical transformers, and embedding models. For the second part, we test each model's consistency under negation, uncertainty, and paraphrased hypotheses to see if the models are relying on surface level lexical patterns. We find that PubMedBERT achieves the strongest performance at 84.8% on MedNLI, but all models drop sharply under negation perturbation— between 23 and 51 percent. This suggests that with MedNLI, high accuracy masks reliance on surface-level lexical patterns rather than robust semantic reasoning.

1 Key Information to include

- Mentor: Arpandeeep Khatua
- External Collaborators (if you have any): No
- Sharing project: No
- Late Days (if you are using any):

Team member	Late days remaining
Ryan Le	6
Total pooled	6
Team size	1
Late days applied for the report	3

2 Introduction

Natural language inference (NLI) aims to predict whether a hypothesis is entailed by, contradicts, or is neutral based off of a given premise. While natural language inference has been relatively successful in broad domains, there hasn't been as much success when it comes to specific domains due to domain-specific terminology and limited training data. One of which is clinical data.

[1] introduces MedNLI to study inference in the clinical domain by utilizing a dataset of real clinical notes annotated by medical professionals to include a premise and hypothesis (input) and label (output). One of the key findings of the study is that there can be strong performance on clinical NLI benchmarks from shallow lexical patterns instead of robust semantic reasoning. Even with high accuracy, there is still a gap as linguistic phenomena such as negation, uncertainty, and temporal language can cause errors. This suggests that the way we evaluate natural language inference is

imperfect and does not capture how models interpret or reason appropriately (at least in this specific study).

The goal of this project is to analyze how language models use both the premise and the hypothesis (found in MedNLI) when interpreting the clinical text, and to understand when the reasoning breaks down. That is, the project will analyze whether model predictions remain consistent when hypotheses are rephrased. This analysis is significant because clinical text can be ambiguous so understanding the model behavior in these different cases can reveal limitations that might not be obvious from accuracy alone.

3 Related Work

The development of large NLI datasets has been important to the progress of understanding natural language. [2] introduced SNLI, the first ever large annotated corpus for NLI. It consists of 550k sentence pairs derived from Flickr image captions. However, because it was limited to a single genre, there was less coverage of linguistic phenomena and its accuracy as a representation of human language was flawed. [3] addressed these issues by creating MultiNLI. It expanded the task to 10 genres of written and spoken English from government documents, telephone conversations, fiction, etc. Beyond addressing the issues of diversity from SNLI, it allowed for explicit benchmarking of cross-genre transfer. That is, could models train on niche genres expand to unseen ones? Baseline results showed that this was difficult, with most models performing much worse than on SNLI.

The tendency of NLI models to utilize shallow lexical cues rather than actual reasoning was uncovered by [4]. [4] showed that a hypothesis only classifier could correctly annotate 67% of SNLI (Stanford Natural Language Inference) and 52% of MultiNLI. They claim that crowd workers would adopt consistent annotation heuristics: entailed hypotheses were more generic language, neutral hypotheses had purpose clauses, contradictions disproportionately contain negation words. This finding established the framework for finding models that actually reason over premise-hypothesis pairs instead of leveraging annotation artifacts. [5] extended on this finding by showing the models fine-tuned on SNLI and MultiNLI (fail on examples that specifically required syntactic reasoning rather than lexical overlap struggled. In 2020, [6], created ANLI (Adversarial ANLI), by creating examples that came from a human and model in the loop process designed to defeat strong pre-trained models. Like the other studies these examples were aimed to resist hypothesis only shortcuts. This dataset currently represents one of NLI's most direct attempts at requiring models to have robust reasoning.

4 Approach

We aim to apply to same investigations performed on SNLI, MultiNLI, ANLI, etc. on MedNLI to see whether this dataset fails under the same conditions.

4.1 Baselines

We implement three baselines of increasing sophistication. The first baseline is a bag-of-words (BOW) for the hypothesis only that is motivated by [4]. We employ a TF-IDF (Term Frequency–Inverse Document Frequency), where we represent each text as a 50,000-dimensional vector, with each dimension corresponding to a word or two-word phrase, weighted by how informative it is. This is a good first baseline, because it is meant to show how hypothesis-only models can still achieve high accuracy on MedNLI and MultiNLI due to the human bias that is introduced when human annotators write hypotheses. The second baseline we use is the same exact BOW setup as the first, but with the premise also included in the input (with the hypothesis). This baseline tests whether adding a premise provides any benefit. The final baseline we utilize is a Frozen BERT [7] + linear head. Specifically, we take bert-base-uncased and freeze all its weights so that nothing updates during training. We then give it the premise and hypothesis (exactly like the second baseline), take the summary vector it produces, and train a small one-layer classifier on top. This gives us insight into how useful BERT's [7] representations are for NLI, before any task-specific training.

4.2 Cross-Encoder Transformers

For our main Transformer models, we feed both the premise and hypothesis into BERT as a combined input, formatted as:

$$\text{input} = [\text{CLS}] \text{premise} [\text{SEP}] \text{hypothesis} [\text{SEP}]$$

CLS = classification token; *SEP* = separator token

BERT will then process this input, allowing each word in the premise to see each word in the hypothesis and vice versa (self attention). This property is particularly important when it comes to catching things like negation. After this, we take the [CLS] vector, $h \in \mathbb{R}^d$, which BERT produces for the whole input (summary), and pass it through a classifier to predict one of three labels: entailment, neutral, or contradiction.

$$\hat{y} = W \cdot \text{Dropout}(h) + b, \quad W \in \mathbb{R}^{3 \times d}$$

We carry out this process with two different versions of BERT: the standard version [7], which is pretrained on Wikipedia and books; PubMedBERT: a version pre-trained on biomedical papers and PubMed articles, which could be helpful for our specific task that includes medical vocabulary and writing style [8]. Each version will be evaluated in both a frozen and fine-tuned setting.

4.3 Bi-Encoder Embedding Models

We also look into the performance of Bi-Encoder embedding models—models (mGTE and Snowflake Arctic) that encode the premise and hypothesis separately into a single dense vector instead of together. The issue with these models is that they were designed for search/retrieval tasks, not NLI. To account for this, we: encode the premise to get a vector h_p , encode the hypothesis to get a vector h_H , and then combine them into a vector that is meant to capture the relationship (what changed between them; what do they have in common)

$$z = [h_p; h_H; |h_p - h_H|; h_p \odot h_H]$$

The classifier will then be able to use this information to know what each sentence says and how they relate and ultimately predict a label

$$\hat{y} = W_2 \text{ReLU}(\text{Dropout}(W_1 z + b_1)) + b_2$$

We evaluate bi-encoders in both frozen and fine-tuned configurations. In the frozen setting, we train only the classification head on top of fixed embeddings; in the fine-tuned setting, we update the full model end-to-end on MedNLI. This lets us compare how much task-specific training helps representations that were originally designed for retrieval.

4.4 Perturbations for Robustness Evaluation

In order to assess whether the models above use semantic reasoning rather than shallow cues, we apply three types of perturbations to the hypothesis and see how the models perform. These perturbations are inspired by [4].

Negation To introduce negation, we add the following phrase to the front of every hypothesis: "It is not the case that". Then, we swap the gold label: entailment to contradiction; contradiction to entailment; neutral remains the same. The model should correctly update its prediction when the semantics change.

Uncertainty To introduce uncertainty, we add the following phrase to the front of every hypothesis: "It is possible that". The gold label stays unchanged. This tests model sensitivity to epistemic modality that doesn't alter the logical relationship.

Paraphrase For each hypothesis, we prompt an LLM (gpt-4.1-nano) to paraphrase the hypothesis in a way that still preserves its meaning. The gold label remains unchanged. This checks whether the model relies on surface form or on semantics.

For each perturbation, we measure *accuracy* (correct predictions on the perturbed set), *consistency rate* (fraction of examples where the model’s prediction on the perturbed pair matches its prediction on the original), and *flip rate* ($1 - \text{consistency}$).

5 Experiments

5.1 Data

Dataset	Train	Validation	Test
MedNLI	11,232	1,395	1,422
MultiNLI	392,702	9,815	9,832

Table 1: Dataset details

We use MedNLI [1], a clinical NLI benchmark derived from radiology report. It contains 11,000 premise hypothesis pairs of clinical notes annotated by physicians. We also use MultiNLI [3], as a proxy dataset for out-of-domain evaluation. MultiNLI contains 433k premise-hypothesis pairs across 10 genres, labeled as entailment, neutral, or contradiction. For robustness, we generate perturbed variants via rule-based negation and uncertainty, and GPT-4.1-nano paraphrase as described above; we restrict evaluation to the perturbed subset so original and perturbed accuracies share the same denominator.

5.2 Evaluation method

For each of the experiments, we evaluate the accuracy and macro-averaged F1 on the validation set derived from the particular dataset we use. These two metrics are important because they show whether a model is balanced in its entailment, neutral, and contradiction predictions, or if it collapses towards a dominant class. For robustness, we evaluate the perturbed accuracy.

5.3 Experimental details

All Transformer models use the HuggingFace `transformers` library. Training runs on a single A10G GPU (24GB VRAM) per model via Modal. We use AdamW (weight decay 0.01) with a linear warmup over 10% of training steps. For frozen models we train only the classification head for 5 epochs with learning rate 10^{-3} and batch size 32. For fine-tuned models we update all parameters for 5 epochs with learning rate $2 \cdot 10^{-5}$ and batch size 16. Bi-encoders use the same fine-tuning schedule. We save the checkpoint with best validation accuracy.

5.4 Results

Table 2: Dev set accuracy and macro-F1 by dataset.

Dataset	Model	Acc (%)	Macro-F1
MedNLI	PubMedBERT (fine-tuned)	84.8	0.848
	BERT (fine-tuned)	79.4	0.794
	GTE-large (fine-tuned)	73.1	0.731
	Snowflake (fine-tuned)	72.8	0.728
	BOW (hypothesis only)	67.5	0.675
	BOW (premise + hypothesis)	62.9	0.628
	GTE-large (frozen)	60.6	0.604
	Snowflake (frozen)	58.0	0.576
	BERT (frozen)	57.2	0.562
MultiNLI	PubMedBERT (fine-tuned)	79.2	0.791
	BERT (fine-tuned)	78.2	0.782
	GTE-large (frozen)	52.0	0.510
	BERT (frozen)	51.0	0.500
	BOW (hypothesis only)	51.0	0.508
	Snowflake (frozen)	48.3	0.476
	BOW (premise + hypothesis)	45.6	0.453

We find that fine-tuned cross-encoders outperform the baselines and frozen models for both MedNLI and MultiNLI. Specifically, PubMedBERT reaches 84.8% on MedNLI and 79.2% on MultiNLI, while BERT reaches 79.4% on MedNLI and 78.2% on MultiNLI. However, beyond these models performing the best for their respective datasets, we also see the gap between PubMedBERT and BERT to be greater for the MedNLI dataset than the MultiNLI. These results are in line with the idea that joint encoding the premise and hypothesis and fine-tuning for specific tasks matter for NLI tasks.

The BOW baselines also give us some interesting insights into the nature of the datasets. For MedNLI, BOW-hypothesis-only reaches 67.5% and BOW-with-premise reaches 62%. The BOW-with-premise, in particular, is surprisingly high, outperforming all of the frozen models. A model with zero understanding of syntax, context, or clinical meaning should not perform this well. This is not the case with MultiNLI where BOW-hypothesis-only (51%) only outperforms one frozen model and BOW-with-premise outperforming none. This is what we expect from general-domain NLI as the greater lexical and genre diversity makes it harder for BOW to exploit. More importantly, it suggests that the MedNLI dataset’s clinical phrasing is most likely more consistent and easier to capture.

Table 3: Robustness on MedNLI: accuracy drop $\Delta = \text{Orig}(\text{restricted}) - \text{Perturbation}(\%)$.

Model	Negation Δ	Uncertainty Δ	Paraphrase Δ
PubMedBERT (fine-tuned)	44.0	0.9	2.3
BERT (fine-tuned)	41.0	1.9	6.7
GTE-large (fine-tuned)	35.1	12.5	5.0
Snowflake (fine-tuned)	50.7	8.4	6.5
GTE-large (frozen)	23.4	2.4	1.9
Snowflake (frozen)	25.4	0.6	1.4
BERT (frozen)	24.3	3.6	2.5

This idea that domain-specific NLI’s consistent phrasing is being exploited can be seen in the drop in accuracy across all models across all different types of perturbations. Under negation, all models drop sharply—between 23.4 and 50.7 points. Although the extent is surprising, it is known that models tend to underweight negation and rely on lexical overlap. The biggest takeaway is that even the best performing model, PubMedBERT, is not robust to negation with one of the highest drop offs at about 35%. Among the fine-tuned models, Snowflake and GTE still perform the worst with Snowflake

having the highest drop off. This intuitively makes sense as they are bi-encoders that are optimized for retrieval and not three-way entailment discrimination.

When looking at uncertainty and paraphrase, the accuracy drop is less drastic—between 0.6 and 12.5 points. Compared to negation, this is relatively surprising—the changes feel relatively similar in complexity semantically. However, when diving into what they actually change it does make sense why the drop is much less. Uncertainty, or hedging, doesn’t actually change the logic label, and paraphrasing is meant to preserve meaning. A model that is actually using semantics instead of surface cues shouldn’t collapse. In other words, the main failure mode when it comes to most of these models is negation handling.

6 Analysis

To understand why negation performs so poorly, we perform a qualitative analysis on the performance of our best performing model, PubMedBERT fine-tuned, and its predictions on a subset of negation-perturbation failures.

Table 4: Predicted label vs. gold label in the annotated negation-perturbation failures ($n = 40$)

Gold label	Predicted label			
	Entailment	Neutral	Contradiction	Total
Entailment	0	0	17	17
Neutral	2	0	20	22
Contradiction	2	0	0	2
Total	4	0	37	40

It is clear from the confusion matrix above that when the model is wrong, it is often predicting contradiction when it is entailment or neutral. We hypothesize this is most likely due to the fact that whenever we add "It is not the case that", it most likely frames the sentence in a way that can seem ambiguous. That is, **just because something isn’t true doesn’t necessarily mean its inverse is true**. This can be extended to most other negations. To confirm the validity of this hypothesis, we **manually label** each of these samples as one of three types: negation confusion, medical knowledge, and ambiguous premise. These types were determined after an initial survey of the subset.

Table 5: Distribution of failure types in the annotated sample ($n = 40$).

Failure Type	Count	%
negation_confusion	21	52.5
medical_knowledge	17	42.5
ambiguous_premise	2	5

Negation Confusion We label a sample as "negation confusion" when the model appears to misread or ignore the negation. For example, the premise "At that time, a CBC and Chem-7 was drawn and the Chem-7 was within the normal range for all of his electrolytes as well as his renal function" supports the hypothesis "patient has normal renal function". When negating it, the hypothesis becomes contradictory. However, the model predicts entailment meaning it ignored the negation and likely matched "normal renal function" to the premise. In other words, the model places so much weight on lexical overlap that it overrides even an explicit negation signal. This ties into the findings above with the BOW models—the model is relying on surface patterns. Another example labeled as negation confusion is "Received GI cocktail for h/o GERD, ?esophageal spasm" vs "It is not the case that Received a proton pump inhibitor". The gold here is neutral and the model predicts contradiction. While less obvious than the previous example, we know that this doesn’t fall into "medical knowledge" or "ambiguous premise" (described in detail below), so we add it under here as it is a fundamental mishandling of negation.

Medical Knowledge We label a case as "medical knowledge" whenever the inference requires some sort of background knowledge that is not explicit in the text (in our case its medicine). For example, "History of end-stage renal disease" with "It is not the case that the patient is on dialysis". Typically, end-stage renal disease implies dialysis, but this is not clear from the text. Therefore instead of predicting contradiction, it predicts entailment as it likely doesn't make that link. Similarly with "The patient noted an increase in shortness of breath for a few days and a chest x-ray revealed a new left pleural effusion" vs "It is not the case that the patient has a hepatohydrothorax," the gold is neutral (pleural effusion does not entail hepatohydrothorax). The model predicts contradiction.

Ambiguous Premise We label a case as "ambiguous premise" when the premise is not detailed enough and so the correct label is somewhat datable, even for a human (we used our best discretion). For example, "Hx of substance abuse CHF EF 55% ... Right shoulder pain ..." with "It is not the case that patient has constant shoulder pain". Even when looking at this, it is hard to tell whether or not the patient has constant shoulder pain as the premise only mentions "right shoulder pain". As expected, the gold is neutral and model incorrectly predicts contradiction.

Overall, this suggests the negation robustness in clinical NLI is not a single problem, but rather systematic. It requires targeted training as most of the examples actually fail when it comes to semantic understanding, not anything related to the domain itself. That is, a significant amount of the negation confusions were actually related to the exploiting of lexical regularities. However, the medical knowledge failures do point to a deeper limitation that these surface level interventions can't fix that should also be explored after.

7 Conclusion

Our results show that fine-tuned cross-encoders achieve strong performance on clinical NLI, with domain-matched pretraining offering an advantage for MedNLI. More importantly, our results show that some of this accuracy is illusory. BOW baselines perform really well on MedNLI, outperforming frozen models, suggesting that the dataset's phrasing may be consistent enough for it to be predictable using lexical patterns alone. The robustness evaluation we do makes this more concrete. Under negation perturbation, all models drop significantly, indicating that a high accuracy does not guarantee robustness when it comes to semantic reasoning.

We prove this with a qualitative analysis of a subset of negation failures. The majority of failures fall under "negation confusion" (52.5%), a category in which we define to be irrelevant of domain knowledge and ambiguous premises, but rather indicative of failed semantic misunderstanding. A substantial second category (42.5%), "medical knowledge", points to examples that even shallow lexical patterns could not account for. Together, these two categories point to the idea the robustness in clinical NLI is both a surface-level problem that is addressable with targeted training data or explicit supervision and a domain-related problem which may require external knowledge integration at some point in the training process.

The primary limitation of this project is that our negation and uncertainty perturbations tend to have the same fixed prefix. This prefix in of itself could have played a significant role as mentioned above, though there were still plenty of examples that were predicted correctly with it. More natural perturbations in the future would make this work more concrete whether that be through human modification or (also) utilizing an llm to modify the data for negation (we elected to not use llm for negation as with a limited size dataset, we wanted a sufficient amount of samples for each "negation marker". For paraphrasing, we felt like this did not matter hence use of llm.)

Team contributions (Required for multi-person team)

This project was done individually

References

- [1] Alexey Romanov and Chaitanya Shivade. Lessons from natural language inference in the clinical domain. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*,

pages 1586–1596, Brussels, Belgium, October–November 2018. Association for Computational Linguistics.

- [2] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference, 2015.
- [3] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.
- [4] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.
- [5] R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference, 2019.
- [6] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial nli: A new benchmark for natural language understanding, 2020.
- [7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.
- [8] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, October 2021.