Evaluating Pragmatic Inference in Pre-Trained Language Models: A Dual-Task Approach

Tristan Sharma
tristansharma@uchicago.edu

Jonathan Roven
jroven@uchicago.edu

Ruby Otavka
rotavka@uchicago.edu

Abstract

This study investigates the capabilities of pre-trained language models, particularly BERT, in handling complex pragmatic inferences, focusing on masked word prediction and relationship prediction. It explores how these models interpret and process different types of entailment, crucial for advanced natural language understanding.

1 Introduction

In the dynamic field of computational linguistics, understanding natural language inference (NLI) and the ability of models to handle nuanced aspects of language such as presuppositions remains a fundamental challenge. Recent studies, like those by Gururangan et al. (2018) and Ettinger (2020), highlight the potential biases and limitations in NLI datasets and models like BERT. Inspired by these insights, our study aims to explore two novel tasks: Masked Language Modeling (LM) and Conventional NLI Task, focusing on how well models manage entailments, focusing on presuppositions, when given different amounts of information.

An important distinction to establish is the difference between presupposition and entailment. There is overlap between these two concepts, but they are distinct. Premise A entails hypothesis B if A and B have the same truth conditions. Premise A presupposes hypothesis B if A can be negated and still presuppose the truth of B. For example, "The king of France is in Paris" entails that "The king of France is in France," while "The king of France is not in Paris" no longer entails that "The king of France is in France." However, both the original premise and the negated premise presuppose "There is a king of France."

Our first task, Masked LM, involves masking different words in the premise of premise-hypothesis pairs. We then examine model predictions of the masked words based on the relationship between these pairs. Previous papers have suggested that BERT may rely on syntactic or word-association cues to predict relationships. We flip the task and see how much BERT can rely on relationships to predict words in the premise of a pair. BERT, known for its word prediction capabilities, might be able to use NLI relationships to inform its predictions. By manipulating the relationship indicators and analyzing the model's performance, we aim to uncover the subtleties of how models understand and process linguistic relations.

The second task, Conventional NLI, leverages a pre-trained classifier model to determine the relationship between premise and hypothesis pairs when a word is masked in the premise. This task is grounded in the understanding that models' ability to classify relationships in NLI can be significantly influenced by the absence of key information. Our approach will provide insights into how models like BERT interpret and classify relationships under information constraints.

Both tasks involve creating a well-controlled test set inspired by IMPPRES (FacebookResearch, 2019) and MultiNLI (HuggingFace, 2020), with varying types of entailment and masked words, to ensure a comprehensive analysis. Our methodology includes both quantitative metrics such as accuracy and qualitative assessments to identify areas where models succeed or fail in understanding presuppositions.

This study contributes to the broader discourse in computational linguistics by offering a deeper understanding of the capabilities and limitations of current NLI models in handling presuppositions. The findings aim to inform future model development and dataset creation, steering them towards more accurate and unbiased understanding of nat-

ural language.

2 Related Work

Recent advancements in computational linguistics, particularly in Transformer-based models like BERT, have significantly enhanced natural language understanding (NLU). Our study contributes to this growing body of research, with a specific focus on evaluating pragmatic inference in these models. The work by Jeretic et al. (2020), introducing the IMPPRES dataset, is seminal in understanding the capabilities of NLI models like BERT in generating pragmatic inferences, particularly entailments and presuppositions. Our approach extends their methodology, incorporating masked word prediction in premise-hypothesis pairs to further investigate these models' inferential capabilities.

Ettinger (2020) provides critical insights into the linguistic understanding of BERT, particularly in tasks requiring deeper language comprehension. This study informs our approach in examining how well BERT and similar models handle complex presuppositional inferences, a dimension not fully explored in Ettinger's work.

The research by (Gururangan et al., 2018) highlights annotation artifacts in NLI datasets. This revelation is crucial for our study as it underscores the importance of creating a well-controlled test set free from such biases, especially when evaluating models on tasks involving masked words and relationship predictions.

Additionally, Kabbara and Cheung (2022) investigate the performance of Transformer-based models on presuppositional inferences. Their findings provide benchmarks and methodological insights that are particularly relevant for our study, as we aim to assess how these models process different types of presuppositions under varying contextual constraints.

Together, these studies form a comprehensive backdrop against which we situate our research, aiming to push the boundaries of current understanding of NLI models' capabilities in handling nuanced and complex linguistic phenomena.

3 Model

We utilize a pre-trained language model to tackle complex pragmatic inferences. We focus specifically on the widely used BERT model, due to its proven effectiveness in capturing contextual relationships within text.

4 Methodology

This section provides an in-depth overview of the creation of our datasets and our dual-task approach, designed to rigorously assess the inferential capabilities of pre-trained language models, particularly focusing on BERT. We publicly release our datasets, model evaluations, and data processing scripts (Sharma, 2023).

Our experimental design is centered around a dual-task framework. The first task, Masked Language Modeling (LM), leverages BERT's inherent capability to predict masked words within a sentence. We crafted a dataset consisting of premise-hypothesis pairs, with selected words systematically masked. These masked words were categorized as either critical or non-critical to the presuppositional relationship inherent in the pair. This distinction allows us to probe the model's understanding of the nuanced relationships between sentences. Additionally, we studied the presence of relationship indicators in the input text, such as "This implies that..." for entailment, to observe their impact on the model's prediction accuracy.

The second task, Conventional NLI Task, employs a trained classifier model to discern the relationship (entailment, contradiction, or neutral) between a premise and a hypothesis, one of which contains a masked word. This task is crucial for understanding how the absence of key information influences the model's classification abilities. The dataset for this task mirrors that of Task 1, allowing for a direct comparison of the model's performance in both word prediction and relationship classification under similar conditions.

4.1 Dataset Creation and Design

Our study required carefully constructed datasets to assess the inferential capabilities of the BERT model. We drew inspiration from two established datasets in the field: IMPPRES and MultiNLI. Our datasets consist of a series of premise-hypothesis pairs designed to evaluate different types of presuppositions and other relationships.

The IMPPRES-inspired portion of our dataset includes 150 unique premises, each paired with three different hypotheses to represent the three relationships of entailment, contradiction, and neutral, resulting in 450 pairs. For the MultiNLI-

inspired portion, we selected and masked 90 pairs from both the training and validation sets of MultiNLI. These pairs cover a range of sentence types and relationships, including those that are not strictly presuppositional.

In our datasets, each premise is systematically manipulated by masking words that are either critical or non-critical to the relationship. This approach allows us to test the model's ability to understand and process the subtleties of language and inferential reasoning. The choice of which word to mask in each sentence was guided by the role that the word plays in the sentence, ensuring a comprehensive and challenging test for the model.

For instance, consider the premise "Rose's ice did stun the children" paired with the hypothesis "Rose has ice," suggesting an entailment relationship. If we mask the word "ice" (a critical word), the input becomes "Rose's [MASK] did stun the children." The model's task is to infer that the masked word is "ice," especially when an entailment relationship indicator such as "This implies that" is included, making the input "Rose's [MASK] did stun the children. This implies that Rose has ice."

In our study, we refer to the test set derived from the IMPPRES dataset as the *IMPPRES Set*, reflecting its inspiration and design basis. Similarly, we designate the test sets inspired by the training and validation data of the MultiNLI dataset as the *MultiNLI Train Split Set* and the *MultiNLI Validation Split Set*, respectively. These terminologies help in clearly distinguishing between the datasets and their respective origins in our analyses.

4.2 Task 1: Masked Language Modeling (LM)

Task 1 focuses on BERT's Masked Language Modeling (LM) capability. We challenge the model to predict words that have been masked in the premises of our premise-hypothesis pairs. This task is designed to test the model's ability to use contextual cues and its understanding of the relationship between the premise and hypothesis to make accurate predictions.

We categorized the masked words in each pair as either 'critical' or 'non-critical' to the presuppositional relationship. This categorization is essential to explore how the model handles different types of information loss. For instance, masking a 'critical' word directly impacts the relationship between the premise and hypothesis, while masking a 'non-critical' word does not.

To assess the influence of explicit relationship indicators, we included conditions where each pair was concatenated with phrases like "This implies that...", "On the contrary...", or "Additionally...", depending on the relationship type (for entailment, contradiction, and neutral respectively). These indicators serve as cues to guide the model's inference process.

In alignment with established practices in previous work utilizing BERT-based models for natural language inference, we concatenate the premise and hypothesis in our input pairs. This concatenation is separated by the special [SEP] token, which signals to the model the end of one segment and the beginning of another. Additionally, each concatenated input sequence is preceded by the [CLS] token. Through this structure, we ensure that BERT processes and understands the relationship between the premise and hypothesis as a coherent, continuous input, facilitating more accurate inference and prediction.

Our analysis involves comparing the model's performance in predicting the masked words under different conditions: with and without relationship indicators, and with critical or non-critical words masked. This comparison allows us to draw insights into how BERT uses linguistic cues and its understanding of entailment and presupposition relationships to make predictions.

4.3 Task 2: Conventional NLI Task

In Task 2, our objective is to evaluate a trained classifier model's capacity to identify the relationship (entailment, contradiction, or neutral) between a premise and a hypothesis, especially when crucial information is masked in the premise. This task aims to delve into the nuances of relationship classification under constraints of incomplete information.

For this task, we employ a variant of the BERT model, specifically fine-tuned for the Natural Language Inference task. This model, pre-trained on the MultiNLI dataset, has demonstrated proficiency in discerning complex relationships within textual pairs (Gu, 2019). It's capable of understanding subtle linguistic cues to categorize relationships as entailment, contradiction, or neutral.

Utilizing the same dataset as in Task 1, we mask critical and non-critical words in the premises

while keeping the hypotheses intact. This setup allows us to assess the classifier's performance in scenarios where essential information may be missing or obscured, simulating real-world challenges in NLI tasks.

Our approach involves comparing the model's classification accuracy across three distinct scenarios: baseline (no words masked), critical word masked, and non-critical word masked. This comparison is critical to understand how varying degrees of information loss affect the model's ability to accurately classify relationships. The baseline scenario serves as a control to gauge the impact of masking words on the classifier's performance.

In this task, we do not provide explicit relationship indicators unlike Task 1, focusing instead on the model's inherent ability to deduce relationships based on the available linguistic context. This approach offers insights into the robustness and adaptability of the model in dealing with partial or obscured information.

Our analysis includes both quantitative metrics, such as classification accuracy, and qualitative assessments. We utilize confusion matrices to illustrate the model's performance across different categories and examine cases of misclassification to understand the model's limitations. By analyzing the types of errors and their contexts, we gain valuable insights into areas where the model might struggle, informing future improvements in NLI modeling.

5 Results

We calculated BERT's accuracy of prediction on our dataset for both tasks in various conditions of given information.

5.1 Task 1

In Task 1, we evaluated BERT's accuracy in predicting masked words in various scenarios. The accuracy was measured across different conditions, including whether the masked word was critical or non-critical to the presuppositional relationship and the presence or absence of relationship indicators. The results for the IMPPRES Set, MultiNLI Train Split Set, and MultiNLI Validation Split Set are presented in Tables 1, 2, and 3, respectively.

The accuracy percentages illustrate the model's varying ability to predict masked words under different conditions. In general, the model showed higher accuracy in predicting critical masked words when relationship indicators were present, suggesting that such indicators play a significant role in guiding the model's inference process. However, the lower accuracy in scenarios without indicators and with non-critical masked words indicates challenges in the model's ability to rely solely on contextual information. These findings hint at insights into the aspects of NLI where the model excels and where it requires further improvement.

Condition	Accuracy
Critical Masked, Indicator	0.342
Non-Critical Masked, Indicator	0.022
Critical Masked, No Indicator	0.291
Non-Critical Masked, No Indicator	0.016

Table 1: Word Prediction Accuracy - IMPPRES Set. Model's performance on Task 1, masked language modeling. Four conditions shown are overall accuracy for when the masked word is critical vs. non-critical to the relationship, and when a relationship indicator is given or not given.

Condition	Accuracy
Critical Masked, Indicator	0.411
Non-Critical Masked, Indicator	0.189
Critical Masked, No Indicator	0.378
Non-Critical Masked, No Indicator	0.189

Table 2: Word Prediction Accuracy - MultiNLI Train Split Set. Model's performance on Task 1 with a dataset using sentences on which the model was pretrained.

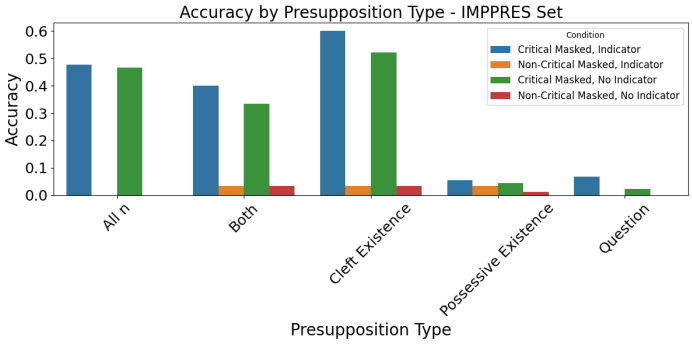


Figure 1: IMPPRES dataset. Model's performance on Task 1, with masked/non-masked and indicator/no indicator conditions. Accuracies are split up depending on the labelled presupposition type.

The accuracies displayed in Figure 4 reveal discernible trends in BERT's predictive performance

Condition	Accuracy
Critical Masked, Indicator	0.400
Non-Critical Masked, Indicator	0.167
Critical Masked, No Indicator	0.389
Non-Critical Masked, No Indicator	0.156

Table 3: Word Prediction Accuracy - MultiNLI Validation Split Set. Model's performance on Task 1 with a dataset using sentences similar to those on which the model was pre-trained.

across different presupposition types within the IMPPRES Set. It is evident that the presence of relationship indicators boosts (by several percent) the model's ability to predict masked words that are critical to the presupposition. For instance, in the 'Cleft Existence' category, there is a difference between the performance with and without indicators. This suggests the possibility that BERT is utilizing the syntactic cues provided by the indicators to inform its predictions. However, the substantially lower accuracy in predicting non-critical masked words, regardless of the presence of indicators, points to the model's reliance on specific lexical items to determine presupposition. It also implies a potential area of improvement in the model's contextual understanding without explicit cues.

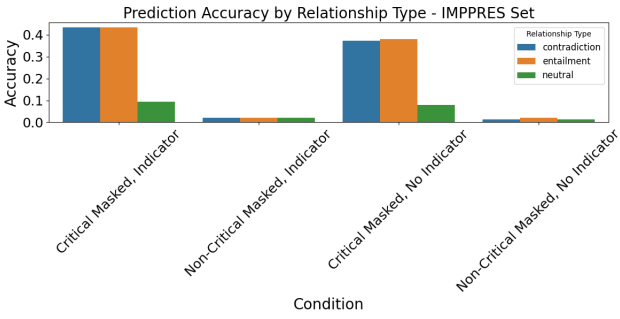


Figure 2: Prediction Accuracy by Relationship Type in the MultiNLI IMPPRES Set. Critical Masked refers to the word carrying the relationship information being masked, while Non-Critical Masked refers to a different word being masked. Indicator refers to the presence of a connecting phrase that indicates the nature of the relationship, while No Indicator refers to the lack thereof.

Across all three sets, the model was more inaccurate in its predictions of non-critical masked words, which was as expected. Additionally, the presence of the indicator did not seem to affect the performance of the model in a majorly significant

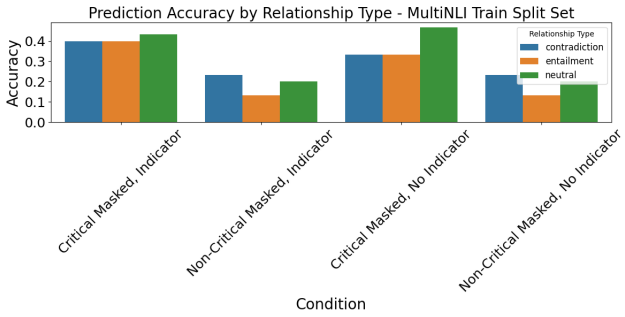


Figure 3: Prediction Accuracy by Relationship Type in the MultiNLI Train Set.

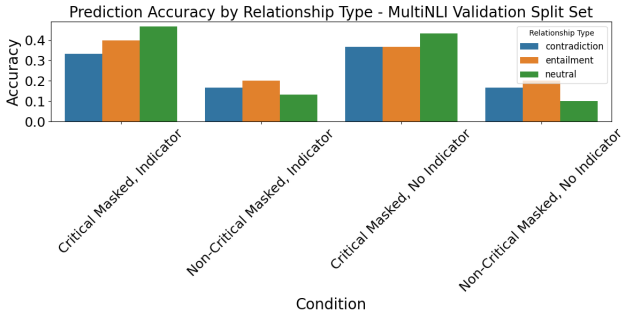


Figure 4: Prediction Accuracy by Relationship Type in the MultiNLI Validation Set.

way across all sets (although there is an increase in accuracy when an indicator is applied–it is only by a few percent on average across our test sets). In the IMPPRES set, the model fared quite well on entailment and contradictions, but had poor results with neutral relationships. In the MultiNLI train set, the model was a bit more accurate on non-critical masked words, but still not as accurate as critical masked words. It generally predicted neutral relationships the best by a small margin. In the MultiNLI validation set, the model also fared better with neutral relationships on critical masked words.

5.2 Task 2

For Task 2, we calculated the model's accuracy for predicting the relationship of a pair. This time, we only varied two elements of the inputs: critical/non-critical masked word and presupposition type.

The confusion matrices provide a visualization that quantifies the performance of our BERT model's baseline predictions on the IMPPRES Set for Task 2. In the matrix, the rows represent the actual categories of the data (entailment, neutral, contradiction), while the columns represent the

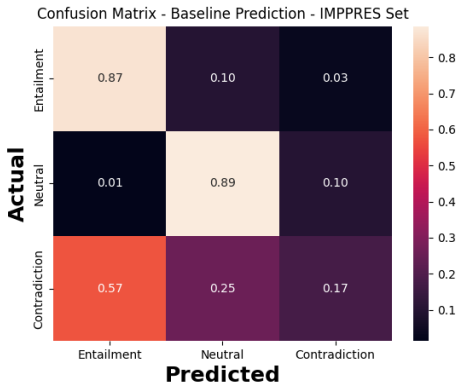


Figure 5: Confusion Matrix for baseline predictions in the IMPPRES Set. Each cell in the matrix represents the proportion of predictions, with rows corresponding to the actual classes and columns to the predicted classes. High values on the diagonal indicate correct predictions, while off-diagonal values represent misclassifications. The matrix shows how well the model distinguishes between entailment, neutrality, and contradiction.

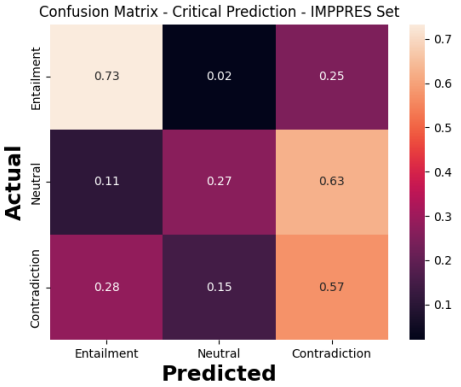


Figure 6: Confusion Matrix for relationship predictions in the IMPPRES Set when a word critical to the relationship is masked.

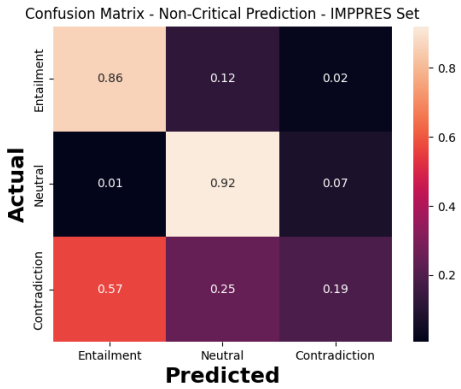


Figure 7: Confusion Matrix for relationship predictions in the IMPPRES Set when a word non-critical to the relationship is masked.

model's predictions for those categories. A high value in a cell indicates a higher proportion of predictions. Ideally, we would expect to see high values along the diagonal, indicating correct predictions, and lower values off-diagonal, which would represent errors or misclassifications. This matrix gives us a granular view of not just the overall accuracy, but more importantly, it helps identify specific areas where the model may be confusing one category for another.

See the appendix for the confusion matrices in our other test sets. We reference them in our discussion later.

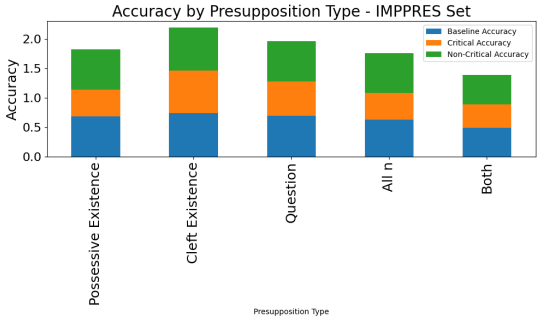


Figure 8: Accuracy of relationship prediction split by presupposition type for IMPPRES dataset. Bars of color represent performance on different NLI relationships (entailment, contradiction, neutral) for each type.

Here, we see that the model predicted presuppositions of the cleft existence type the most accurately by a relatively wide margin. On the other hand, the model performed least accurately on presuppositions of the "both" type.

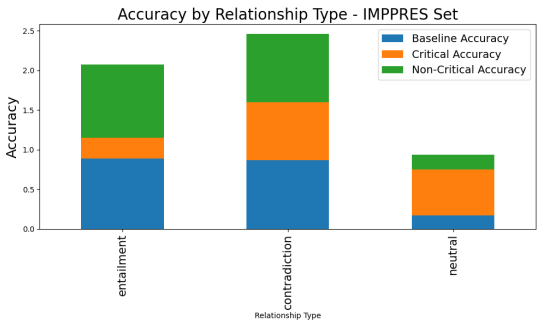


Figure 9: Accuracy by relationship type (entailment, contradiction, neutral) for IMPPRES dataset. Bars are split in color by masked word condition.

The model performed accurately for entailment and contradictions on the IMPPRES set, but was significantly less accurate for neutral relationships. The model performed quite accurately for all three prediction types on the validation set,

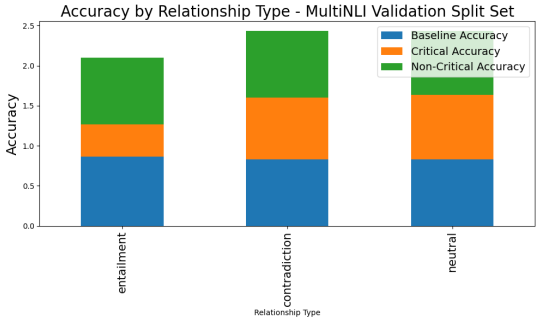


Figure 10: Accuracy by relationship type for MultiNLI Validation Set.

with only small differences between relationship types.

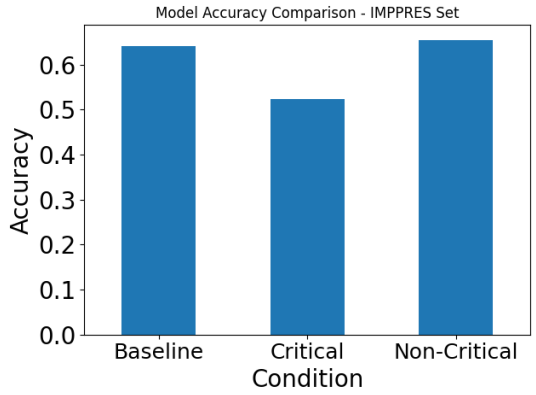


Figure 11: Overall model accuracy in relationship prediction for IMPPRES set. Accuracies are split into the condition of a critical word being masked, a non-critical word being masked, and the baseline of no words being masked.

The model performed fairly accurately in the baseline and non-critical cases on the IMPPRES set. Its performance in the critical case was slightly less accurate than the others, but still slightly above 0.5.

6 Discussion

The model's performance on the two tasks varied depending on the conditions we laid out.

6.1 Task 1

In Task 1, the model predicted a masked word in the premise when it was critical or non-critical to the NLI relationship of the pair. Consider the case where a masked word is critical to the pair's relationship, like in the pair "Rose's [MASK] did stun the children. Rose has ice." Without any indication of whether this is entailment, contradiction,

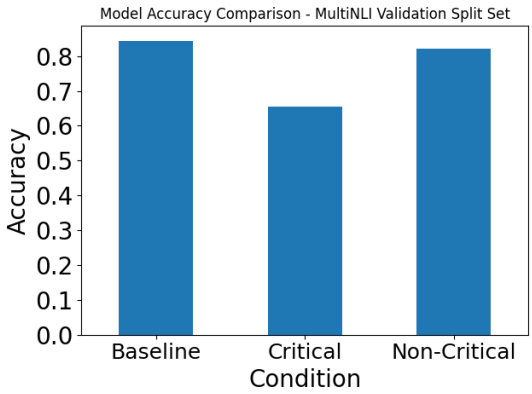


Figure 12: Overall model accuracy in relationship prediction for the MultiNLI validation set.

or neutral, it should be impossible to guess the masked word. On the other hand, if a relationship indicator is given telling us that the relationship is entailment, it should be easy to guess the masked word.

BERT showed a much lower accuracy version of these trends. Looking at the difference between the critical and non-critical masked conditions, the model predicted the critical word at much higher accuracy than the non-critical word for all three datasets, which is what we would expect, at least for when the relationship indicator is given.

Notably, the presentation of the relationship indicator only changed the accuracy result by a few percent. If the model were taking the relationship indicator as the intended indication of the NLI relationship, and if it could then use this relationship to correctly predict the masked word, we would expect that without a relationship indicator, the model would be worse at predicting a masked critical word. Instead, it performs almost the same. This is like equating "Rose's [MASK] did stun the children. Rose has ice." with "Rose's [MASK] did stun the children. This implies that Rose has ice." The model's performance on these conditions may indicate that the relationship indicators we chose are not cuing the model to the NLI relationship as we had intended. It could also indicate that the model is using proximal words to predict the masked word rather than using the given relationship. In general, BERT performed worse in all conditions on the IMPPRES dataset than it did on the MultiNLI datasets, predictably because it was pretrained on the MultiNLI, not IMPPRES.

The trends of higher accuracy for critical masked prediction are maintained in the presup-

positional types all n, both, and cleft existence, but possessive existence and question types show very low accuracy for all conditions. In the question premises, because of our masking protocol, the masked word was often a proper name like "Benjamin," which could have disproportionately lowered the accuracy for the untrained model on these pairs. For possessive existence, the word "had" is included in all the hypotheses but not in the premises, which might have thrown off the model's predictive associations.

In the IMPPRES dataset, the critical and non-critical masked word predictions for contradiction and entailment had a higher accuracy than those for neutral condition, which we would also predict with an understanding of entailment. For "Rose's [MASK] did stun the children. Lissa has ice," the premise and hypothesis are independent from each other so the entailment relationship cannot cue the masked word in the same way it does for entailment or contradiction.

Strangely, the opposite trend was shown for the MultiNLI datasets, with the neutral relationship giving higher accuracy predictions than the entailment or contradiction relationships. This shows a lack of informative information offered by the relationship. This could also show a flaw in the MultiNLI dataset pairs.

6.2 Task 2

In Task 2, the model predicted the relationship type between a masked pair. In this task, if the masked word is critical to the relationship, the model should not be able to predict the relationship type because a key piece of information is missing. If the masked word is non-critical to the relationship, it should be possible to predict the relationship type at the same accuracy as it does for the baseline, non-masked pairs, because the relationship does not depend on the masked word.

For the IMPPRES set, the model may have predicted cleft existence most accurately due to the roundabout constructions of the premises, as they are cleft sentences. These sentences tend to use the same important words as the simpler sentences resembled by the hypotheses, with a slightly more complex structure that often emphasizes one constituent in particular. Compare this construction to the possessive existence sentences, where the word "has" in the hypotheses missing in the premises, instead replaced with the possessor in the genitive case. The presence of the same important words in both the premises and hypotheses of the cleft existence sentences could explain the model's more accurate performance. On the other hand, the model performed least accurately on presuppositions of the "both" type. This may have been caused by the lack of the word "both" in the hypotheses, where it is instead replaced by "two". Compare this construction to the "all n" sentences, where the explicit number is mentioned in both the premises and the hypotheses, potentially allowing the model to perform more accurately.

Regarding the confusion matrices on the MultiNLI train set, the model performed with at least 80% accuracy on all relationship types in baseline and non-critical predictions. It performed worse but not badly in critical predictions, with all relationship types being predicted with more than 50% accuracy. The lowest accuracy was 57% in neutral sentences, which the model often mistook for contradictions. This discrepancy is similar to the one seen in the critical predictions of neutral sentences in the IMPPRES set, suggesting that the same phenomenon of assuming a contradiction from the rest of the premise also occurred in the MultiNLI sets, with which the model was more familiar.

Similarly to the training set, the model performed quite accurately on all relationship types in baseline and non-critical predictions for the validation set, once again with at least 80% accuracy. The results for critical predictions mirror those for the training set, but with even less accuracy. The model fared relatively well on entailment and contradictions but mistook neutral sentences for contradictions 57% of the time, only correctly identifying them with 40% accuracy. These results reinforce the idea that the model has a tendency to assume a contradiction when critical information is missing, perhaps guessing that it is more likely for the masked word to contradict the hypothesis than to be irrelevant or to entail it.

6.3 IMPPRES and MultiNLI

Overall, the model performed differently on the IMPPRES-inspired pairs compared to the MultiNLI-inspired pairs in both tasks. Our original design only included IMPPRES-inspired pairs, but the baseline accuracy for these pairs was only 63% for Task 2 (See Figure 11). We hypothesized that the accuracy was low because our model was

trained on a MultiNLI dataset, which has different vocabulary and structure than the IMPPRES dataset. Additionally, as mentioned in section 4.1, the MultiNLI dataset includes NLI relationships more generally, while our test set for IMPPRES specifically included presuppositional relationships. This means that our model was not trained on the specific types of relationship we were testing it on. Therefore, we also calculated the accuracy of the model's word and relationship predictions when the inputs are masked pairs inspired by the MultiNLI dataset. Predictably, the model's accuracy was highest when it was tested on its training sentences, second-highest when it was tested on the validation sentences for the set it was trained on, and lowest for the IMPPRES set, which it was not trained on. Still the 63% accuracy for IMPPRES shows some capability for generalization beyond the training dataset, so we still examined the difference between accuracies in different conditions for IMPPRES.

7 Conclusion and Future Work

Overall, BERT made predictions based on a masked premise in a way that shows that it can capture some elements of NLI. The trends of accuracy shown in our results indicate that generally, BERT reflects some patterns of consistency in NLI but also performs at a generally low accuracy and performs even worse on certain tasks. Therefore, BERT lacks the general knowledge that a human has about NLI that allows us to make generalizations and predictions based on understood relationships. Instead, BERT relies on things like context cues and word order to form its predictions.

In order to put these results in context, it would be useful to test human participants on both of our tasks. Since humans are able to gain a very high comprehension of NLI relationships, comparing BERT's performance to humans' performance could tell us more about what BERT's performance means in the context of understanding NLI. This could also help us evaluate the efficacy of our tasks and datasets, because if there are places humans make mistakes where they shouldn't, it could mean that our task is flawed.

Because our model was pre-trained on MultiNLI tested on an IMPPRES-inspired dataset, we were unable to focus as much on presuppositional types as we had originally intended. If we had the time and resources to pre-train a model on IMPPRES, we might have been able to get more accurate results on the model's capabilities with this dataset for Task 2.

Another possible follow-up for these tasks could be adding more conditions in order to continue to narrow down the contexts in which BERT succeeds or makes mistakes. For simplicity's sake, we used only the unembedded premises from IMPPRES, but it could be interesting to see how BERT performs when given a negative or interrogative premise, as well as other presuppositional triggers.

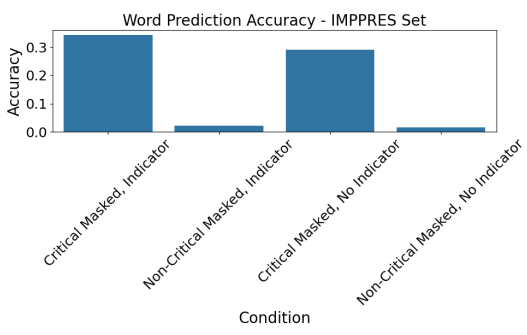
For our analyses of Task 1, we only looked at k=1, classifying the model's accuracy on whether the correct prediction was the model's top word prediction. However, further study could compare the accuracies with different k values.

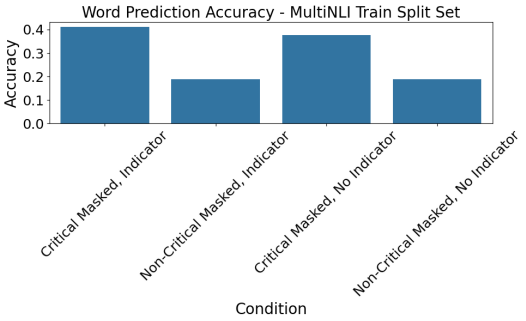
Various relationship indicators could be tested to better determine the effect that they had on the prediction accuracies. We could also have included the presence of relationship indicators as a condition in task 2 as well as task 1, which could have been another check for the efficacy of the indicators.

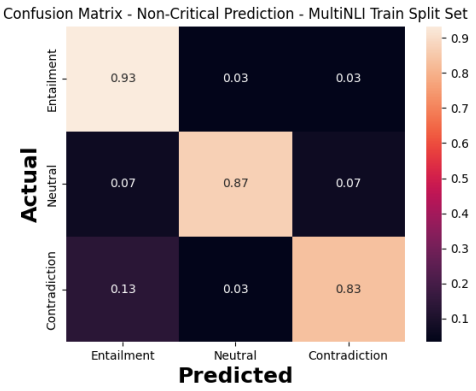
Finally, we had to choose only a subset of roles for the critical vs non-critical masked words, but further study could examine the importance of the specific placement of these words. Because the masked-word paradigm indicates the model's reliance on specific words and sentence structure, further study in this direction could continue to elucidate the mechanism by which BERT makes NLI determinations.

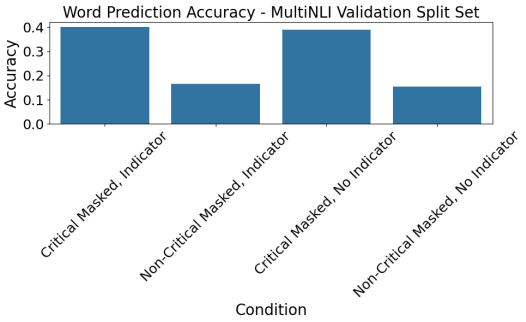
8 Appendix

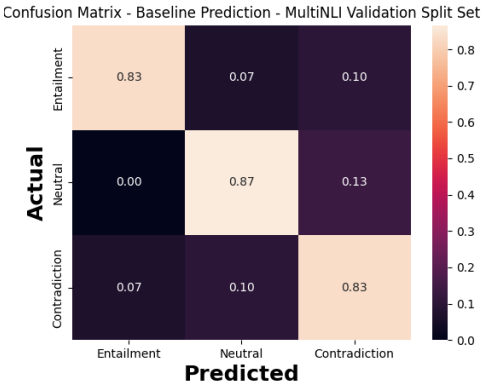
8.1 Task 1 Figures

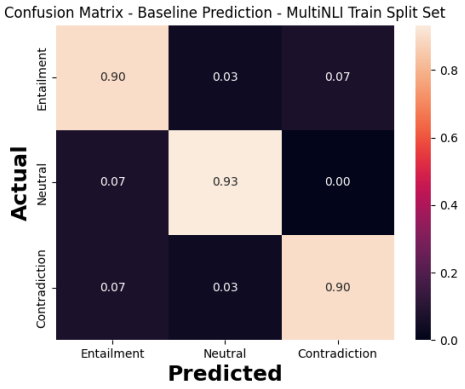


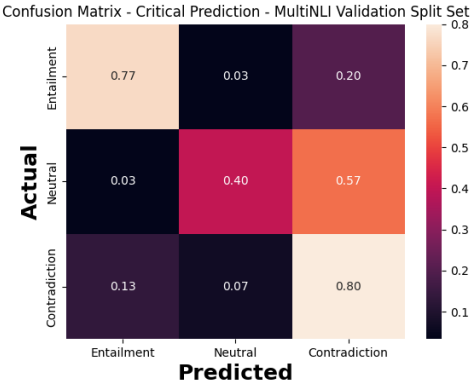

Word Prediction Accuracy - MultiNLI Train Split Set
Accuracy
0.0
0.1
0.2
0.3
0.4
Critical Masked, Indicator
Non-Critical Masked, Indicator
Critical Masked, No Indicator
Non-Critical Masked, No Indicator
Condition

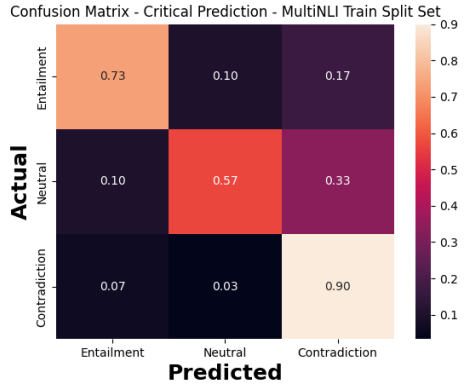

Confusion Matrix - Non-Critical Prediction - MultiNLI Train Split Set
Actual
Entailment
Neutral
Contradiction
0.93
0.03
0.03
0.07
0.87
0.07
0.13
0.03
0.83
Entailment
Neutral
Contradiction
Predicted
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9

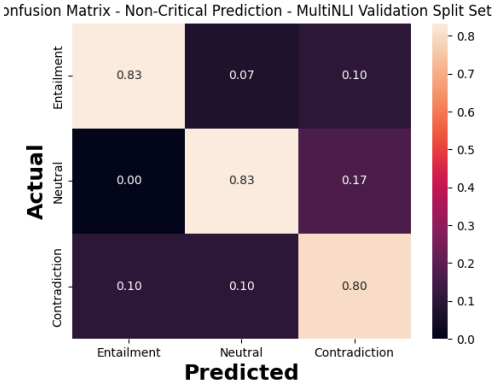

Word Prediction Accuracy - MultiNLI Validation Split Set
Accuracy
0.0
0.1
0.2
0.3
0.4
Critical Masked, Indicator
Non-Critical Masked, Indicator
Critical Masked, No Indicator
Non-Critical Masked, No Indicator
Condition

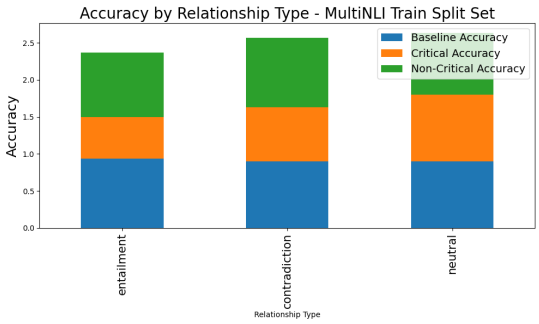

Confusion Matrix - Baseline Prediction - MultiNLI Validation Split Set
Actual
Entailment
Neutral
Contradiction
0.83
0.07
0.10
0.00
0.87
0.13
0.07
0.10
0.83
Entailment
Neutral
Contradiction
Predicted
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8

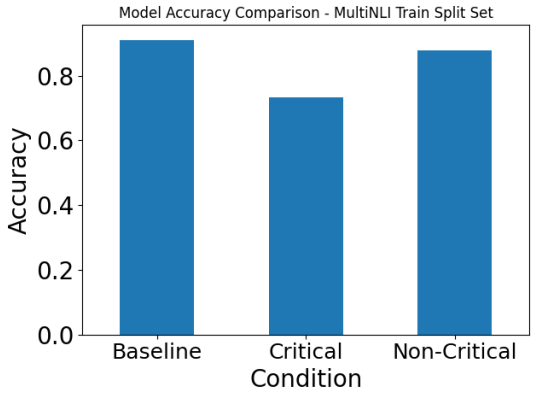

Confusion Matrix - Baseline Prediction - MultiNLI Train Split Set
Actual
Entailment
Neutral
Contradiction
0.90
0.03
0.07
0.07
0.93
0.00
0.07
0.03
0.90
Entailment
Neutral
Contradiction
Predicted
0.0
0.2
0.4
0.6
0.8


Confusion Matrix - Critical Prediction - MultiNLI Validation Split Set
Actual
Entailment
Neutral
Contradiction
0.77
0.03
0.20
0.03
0.40
0.57
0.13
0.07
0.80
Entailment
Neutral
Contradiction
Predicted
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8


Confusion Matrix - Critical Prediction - MultiNLI Train Split Set
Actual
Entailment
Neutral
Contradiction
0.73
0.10
0.17
0.10
0.57
0.33
0.07
0.03
0.90
Entailment
Neutral
Contradiction
Predicted
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9


onfusion Matrix - Non-Critical Prediction - MultiNLI Validation Split Set
Actual
Entailment
Neutral
Contradiction
0.83
0.07
0.10
0.00
0.83
0.17
0.10
0.10
0.80
Entailment
Neutral
Contradiction
Predicted
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8





8.2 Task 2 Figures

References

Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models.

FacebookResearch. 2019. Imppres: Repository for improving neural network models for natural language understanding tasks. https://github.com/facebookresearch/Imppres. Accessed: [12/9/23].

Y. Gu. 2019. BERT for Natural Language Inference. https://github.com/yg211/bert_nli.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. *CoRR*, abs/1803.02324.

HuggingFace. 2020. Multinli dataset. https://huggingface.co/datasets/multi_nli. Accessed: [12/9/23].

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online. Association for Computational Linguistics.

Jad Kabbara and Jackie Chi Kit Cheung. 2022. Investigating the performance of transformer-based NLI models on presuppositional inferences. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 779–785, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

T. Sharma. 2023. Pragmatic Inference in Natural Language Inference Models. https://github.com/tristan-sharma/pragmatic-inference-nli.