Beyond Accuracy: Revisiting Out-of-Distribution Generalization in NLI Models

Zahra Delbari
Tehran Institute for Advanced Studies
Khatam University, Iran
z.delbari@khatam.ac.ir

Mohammad Taher Pilehvar
Cardiff University
United Kingdom
pilehvarmt@cardiff.ac.uk

Abstract

This study investigates how well discriminative transformers generalize in Natural Language Inference (NLI) tasks. We specifically focus on a well-studied bias in this task: the tendency of models to rely on superficial features and dataset biases rather than a true understanding of language. We argue that the performance differences observed between training and analysis datasets do not necessarily indicate a lack of knowledge within the model. Instead, the gap often points to a misalignment between the decision boundaries of the classifier head and the representations learned by the encoder for the analysis samples. By investigating the representation space of NLI models across different analysis datasets, we demonstrate that even when the accuracy is nearly random in some settings, still samples from opposing classes remain almost perfectly linearly separable in the encoder's representation space. This suggests that, although the classifier head may fail on analysis data, the encoder still generalizes and encodes representations that allow for effective discrimination between NLI classes.

1 Introduction

With the rise of pre-trained language models (PLMs), NLI models have surpassed human performance on several benchmarks. However, this raises questions about whether these models truly understand the NLI task or merely exploit shortcuts and superficial patterns to achieve high accuracy without genuine linguistic comprehension. To address these concerns, researchers have developed analysis and controlled datasets to expose the limitations of NLI models, revealing their reliance on spurious correlations rather than deep linguistic understanding (McCoy et al., 2019; Ravichander et al., 2019; Naik et al., 2018a). For example, models often struggle with numerical reasoning or generalize poorly to adversarial datasets like HANS (McCoy et al., 2019). However, does evaluating a model solely based on its predicted labels provide a complete picture of what it has learned? If a model performs poorly on an out-of-distribution (OOD) dataset, can we conclusively argue that it lacks the essential knowledge for the task? Prior work challenges these assumptions. Studies show that classifier accuracy can be highly sensitive to decision thresholds (Yaghoobzadeh et al., 2021; Zhao et al., 2021a), and representation-space analyses reveal rich task-relevant structures even when classifier predictions fail (Marks and Tegmark, 2024; Amini and Ciaramita, 2023). This highlights that the representation space contains a meaningful structure beyond what accuracy captures. Similar findings exist in computer vision, where models trained on digit recognition datasets—even with some labels withheld—still cluster unseen categories meaningfully (Dyballa et al., 2024).

This paper revisits the generalization of NLI models on OOD datasets[1]: Does poor performance on OOD datasets truly indicate a lack of knowledge, or is it a symptom of misalignment between the encoder's representations and the classifier's decision boundaries? Our findings reveal that the latter could also be true in some settings. We analyze the representation space of NLI models (Section 3), focusing on linear separability (LS) across OOD datasets. The encoder representations exhibit strong LS for all datasets—even those where classifier accuracy is poor. For instance, on the Stress Test Numerical subset, the encoder representations show near-perfect LS (>96%), despite the classifier head achieving only 42% accuracy. This stark contrast suggests that the encoder captures task-relevant knowledge that the classifier fails to exploit. In Section 4, we further examine whether LS can serve as a reliable indicator of a model's knowledge in NLI, exploring encoder's behavior

[1]In this paper, we use *OOD* and *analysis* datasets interchangeably.

across different scenarios.

2 NLI Task and Analysis Datasets

NLI task requires determining the logical relationship between two input sentences: the premise and the hypothesis. The goal is to classify whether the hypothesis entails the premise, contradicts it, or is neutral (neither entailing nor contradicting). The Stanford Natural Language Inference (Bowman et al., 2015, SNLI) and Multi-Genre Natural Language Inference (Williams et al., 2018, MNLI) dataset are among the most widely used benchmarks for this task. Although fine-tuned PLMs achieve high performance on these benchmarks, their performance on analysis datasets suggests that these high results do not necessarily indicate a deep understanding of the task. In this section, we introduce the analysis datasets we selected for this study. These datasets are among the most popular and relatively large evaluation benchmarks for NLI, each designed to target different aspects of linguistic knowledge.

2.1 SICK

Sentences Involving Compositional Knowledge (Marelli et al., 2014, SICK) is a benchmark dataset designed for evaluating compositional distributional semantics models. Comprising over 10,000 pairs of sentences labeled as entailment, contradiction, or neutral, SICK serves as a benchmark for evaluating models' ability to handle compositional meaning and inference (see examples in Appendix Table 6).

2.2 HANS

The Heuristic Analysis for NLI Systems (McCoy et al., 2019, HANS) is a synthetic dataset created to expose the reliance of NLI models on the overlap heuristic. It features premise-hypothesis pairs where all words in the hypothesis appear within the premise. The dataset is divided into three heuristic categories based on word overlap patterns: lexical overlap, subsequence, and constituent. For each category, half of the examples align with the heuristic and are labeled as "Entailment," while the other half contradict the heuristic and are labeled as "Non-Entailment." Some examples from this dataset are provided in Appendix Table 5. NLI models often incorrectly classify samples that contradict the heuristic as "Entailment," demonstrating their reliance on superficial cues rather than true sentence understanding.

2.3 Stress Test

The Stress Test (**ST**) (Naik et al., 2018b) was designed to uncover weaknesses in models fine-tuned on the MNLI dataset by analyzing their performance on challenging validation samples. It identifies key linguistic phenomena, such as *word overlap*, *negation*, *length mismatch*, *antonyms*, *spelling errors*, and *numerical reasoning*, that frequently caused models to make errors.

To create subsets targeting these phenomena, specific strategies were applied: for *word overlap* (ST-WO) and *negation* (ST-N), phrases like "and true is true" and "and false is not true" were appended to the hypotheses. For *length mismatch* (ST-LM), the phrase "and true is true" was repeated five times at the end of the premises. *Numerical reasoning* (ST-NU) was crafted using premises extracted from the AQuA-RAT dataset, paired with generated hypotheses (see examples Appendix Table 7).[2] Except for ST-LM, model performance was significantly lower on these subsets compared to the standard validation set, with particularly poor accuracy on ST-N, where results approached random chance.

3 Representation Space and Linear Separability

Discriminative transformers are composed of two key components: the encoder, which typically uses a pre-trained language model, and the classifier head, which is usually a shallow multi-layer perceptron (MLP). In classification tasks, the `[CLS]` token, representing the entire input sequence, is passed to the classifier head to generate the final prediction. Since the `[CLS]` token encodes all the input information and serves as the primary feature for classification, our investigation centers on understanding its representation within the model.

3.1 Experimental Setup

Baseline models. We explore the representation space produced by the `[CLS]` token across three models: RoBERTa (Liu et al., 2019b), BERT (Devlin et al., 2019), and DistilBERT (Sanh et al., 2020). For consistency, we employ the base versions of all models. While BERT has been the focal point in most analytical works, our study extends this analysis to RoBERTa, known for its robust-

[2]We dismiss anatomy subset because it contains samples of only one class.

Dataset	DistilBERT	BERT	RoBERTa
MNLI-m	82.1±0.2	84.3±0.4	87.5±0.1
MNLI-mm	82.2±0.2	84.4±0.5	87.4±0.2
SICK	54.4±0.6	56.4±0.8	57.5±0.5
HANS+	97.3±0.8	97.7±1.2	98.7±0.1
HANS−	9.6±2.7	32.4±5.5	50.1±2.0
ST-NU	35.1±1.5	42.6±1.7	59.5±2.9
ST-LM	80.1±0.2	82.3±0.3	85.2±0.2
ST-N	54.6±1.0	56.0±0.3	57.1±0.7
ST-WO	60.1±1.3	59.0±1.3	63.0±2.7

Table 1: Accuracy of the three baseline models on NLI analysis datasets SICK, HANS, and Stress Test (ST-X), as well as the standard validation sets MNLI matched (-m) and MNLI mismatched (-mm), reported for five runs.

ness, and DistilBERT, a more lightweight alternative with less capacity to gain knowledge.[3]

Datasets. We fine-tune the baseline models on the MNLI and SNLI datasets. Then, we examine the `[CLS]` token generated by these models for analysis datasets mentioned in Section 2. Since the training datasets have three labels (*entailment*, *contradiction*, and *neutral*), while HANS only has two (*entailment* and *non-entailment*), we map both *contradiction* and *neutral* predictions to *non-entailment* and leave *entailment* unchanged.

Fine-tuning. Each fine-tuning run consists of training the models for 5 epochs with a learning rate of 2×10^{-5}, a batch size of 32, the AdamW optimizer, and a learning rate decay of 0.02.

Dimension reduction. To gain a deeper understanding of the representation space in classification, we visualized it by plotting the representations. Since the embedding space is high-dimensional (768 for base models), we applied Principal Component Analysis (PCA) to reduce the dimensionality to three, allowing for a clearer visualization. The reduced space captures approximately 77% of the total variance, with each remaining component contributing less than 2%, as shown in Figure 8 in the Appendix. Therefore, this three-dimensional representation provides a reasonable approximation of the original high-dimensional space.

[3]We also checked BERT-large, and the LS remains strong despite poor accuracy.

3.2 Representation Space Visualization

The average performance of all baselines models are reported in Table 1 for MNLI and in Table 8 (in the Appendix) for SNLI. Consistent with the purpose of the HANS dataset, the table confirms that all models tend to classify HANS samples as *entailment*, achieving near-perfect results on HANS+ but very poor performance on HANS−, which indicates a strong reliance on overlap heuristics. For the Stress Test dataset, the results for the ST-NU subset are particularly poor, with performance close to random chance for DistilBERT, suggesting that these models struggle to infer anything meaningful from mathematical or equation-based samples.

Figure 1 illustrates the representation space of one trial from each model. Given that HANS has the largest sample size (30000) compared to the other datasets, we find it clearer to visualize its representation in relation to the other analysis datasets. As a result, all visualizations are for HANS unless otherwise specified. To match the number of HANS samples, we selected 30K MNLI (train) samples and plotted the 3D space for all 60K data points.[4]

All the models show distinct regions within the representational space, with each region corresponding to one class of MNLI. This structure enables the classifier head to achieve linear separation. The representational space can be visualized as a three-petaled flower, with each petal representing one of the three classes.

For the HANS dataset, however, the data is positioned beneath these petals. If the model's accuracy (trained on MNLI) on HANS matched its performance on MNLI, we would expect the data points to be similarly organized into distinct petals. Instead, the majority of the HANS data is concentrated in the (blue) petal corresponding to the *entailment* label, which cause the poor accuracy presented in Table 1.

But the interesting point is that despite the clustering of HANS data in the entailment region, the orange and yellow points—representing entailment and non-entailment labels, respectively—are still clearly separated. This suggests that although the HANS data is incorrectly categorized according to the standard regions determined by the classifier head, the opposite labels remain well-separated in

[4]Given the challenges of displaying 3D images, we provide 2D views from different angles to offer a clearer understanding of the 3D representation space.

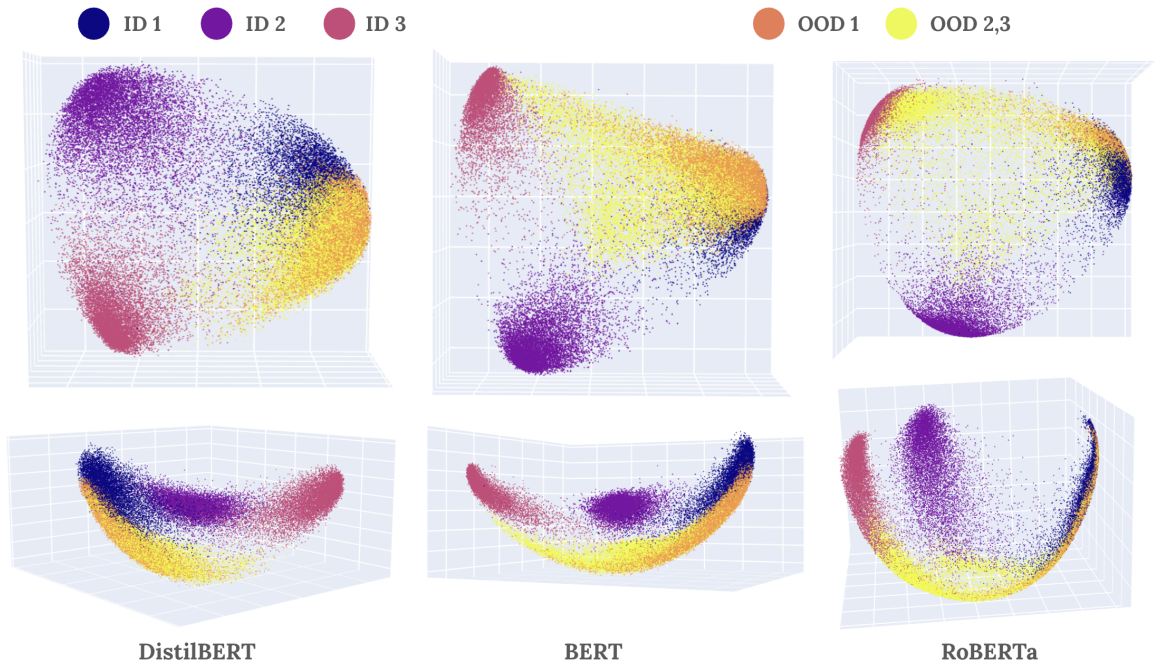


Figure 1: 3D visualization of the [CLS] token representation space for the MNLI (in-distribution, ID) and HANS (OOD) datasets, generated by the three baseline models. Colors indicate the gold labels. In all baseline models, the orange and yellow points (representing the two classes of HANS) are clearly distinguishable. The 3D spaces are visualized from two different perspectives (class 1: Entailment, class 2: Neutral, and class 3: Contradiction). Despite encoder's positioning of the OOD samples towards ID1, they are internally separated for their two classes (OOD1, and OOD2,3), as particularly visible from the BERT visualization (middle, top).

the representational space. For additional clarity, see Figure 2, which compares model outputs (labels given by the classifier head) (2b) with the gold labels (2a) of HANS.

3.3 Linear Separability (LS)

To evaluate whether the encoder's [CLS] embeddings admit **linear separability** between classes, we formalize the problem as follows. Let $\mathbf{h}_i \in \mathbb{R}^d$ denote the last layer hidden state of the [CLS] token for the i-th input sample, and $y_i \in \{1, \ldots, K\}$ its corresponding class label. We assess whether there exists a linear decision boundary that separates classes in the embedding space. This reduces to solving for parameters $\mathbf{W} \in \mathbb{R}^{K \times d}$ and $\mathbf{b} \in \mathbb{R}^K$ such that $\hat{y}_i = \arg\max_k (\mathbf{W}\mathbf{h}_i + \mathbf{b})$ achieves minimal cross-entropy loss over N samples:

$$\mathcal{L}(\mathbf{W}, \mathbf{b}) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbb{I}[y_i = k] \log \left(\frac{\exp(\mathbf{W}_k \mathbf{h}_i + b_k)}{\sum_{j=1}^{K} \exp(\mathbf{W}_j \mathbf{h}_i + b_j)} \right), \quad (1)$$

where $\mathbb{I}[\cdot]$ is the indicator function. High accuracy on held-out data implies the existence of a hyperplane $\mathbf{W}_k \mathbf{h} + b_k = \mathbf{W}_{k'} \mathbf{h} + b_{k'}$ separating classes k and k'.

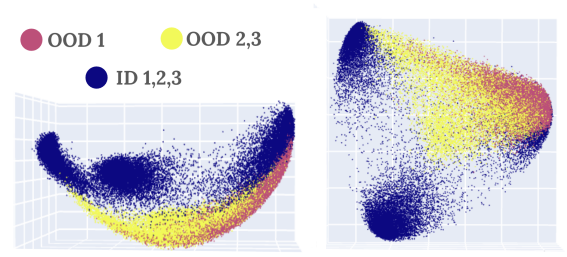
3.4 LS Results

Table 2 quantifies the degree of LS for all analysis datasets across the baseline models fine-tuned on MNLI and SNLI. For comparison, we also present the results of a random experiment, where the labels of the [CLS] token are shuffled randomly, and decision boundaries are then computed (in parentheses). It is important to note that, in higher-dimensional spaces and when the dataset size is small, the accuracy of purely random data can exceed the expected accuracy (50% for two classes and 33% for three classes).

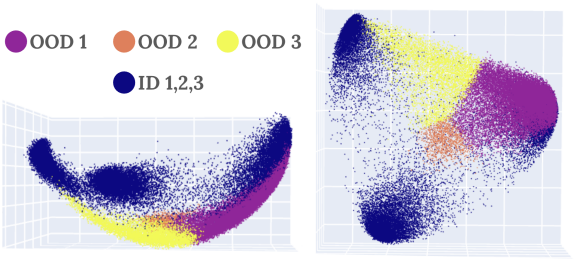
Universal Linear Separability (LS) All models achieve high LS scores (77–98% for models fine-tuned on MNLI), confirming that learned representations inherently encode task-relevant features rather than relying on superficial patterns. This is particularly evident in HANS, where LS exceeds 90% (e.g., BERT: 95.6%). Despite the high degree of overlap between entailment and non-

	MNLI			SNLI		
Dataset	**DistilBERT**	**BERT**	**RoBERTa**	**DistilBERT**	**BERT**	**RoBERTa**
MNLI	80.1±0.2 (48.2)	82.4±0.1 (48.0)	86.4±0.3 (47.7)	70.3±0.2 (48.5)	74.6±0.3 (48.4)	79.4±0.3 (48.5)
SNLI	77.8±1.2 (49.0)	82.0±1.2 (48.8)	85.6±0.2 (48.1)	85.6±0.5 (48.7)	88.2±0.6 (48.9)	89.3±0.8 (49.0)
SICK	84.4±0.8 (59.7)	87.4±0.8 (60.0)	89.4±0.6 (59.5)	86.6±1.5 (59.9)	87.8±0.8 (59.9)	89.7±0.5 (59.8)
HANS	91.1±0.5 (56.1)	95.6±0.7 (56.2)	95.4±0.7 (55.9)	88.4±0.8 (56.0)	93.9±0.6 (56.3)	95.6±0.4 (56.1)
ST-LM	77.7±0.2 (48.0)	80.3±0.2 (47.7)	84.3±0.3 (47.5)	69.0±0.3 (48.6)	72.7±0.2 (48.6)	77.2±0.3 (48.1)
ST-N	77.6±0.5 (48.3)	80.1±0.3 (47.9)	84.0±0.3 (47.3)	67.5±0.3 (48.0)	71.6±0.7 (48.3)	76.1±0.5 (47.7)
ST-WO	78.4±0.1 (48.0)	80.9±0.3 (48.0)	84.8±0.3 (47.5)	69.0±0.3 (48.3)	72.8±0.5 (48.6)	77.5±0.5 (48.5)
ST-NU	96.3±0.4 (51.5)	97.4±0.6 (51.5)	98.4±0.9 (50.3)	94.2±1.3 (51.5)	96.3±0.6 (51.6)	98.4±0.2 (51.7)

Table 2: Results of linear separability for analysis datasets, based on models fine-tuned on MNLI and SNLI. The linear separability is the accuracy of linear boundaries reported for the evaluation set of HANS, the mismatched subsets of the ST datasets (ST-X), and MNLI, as well as the validation sets of SICK and SNLI. The numbers in parentheses represent results from random experiments.



(a) Gold Labels



(b) Predicted Labels

Figure 2: A comparison of (a) the scatter of [CLS] tokens for two HANS classes in space and (b) how a fine-tuned BERT model classifies them into three classes, with MNLI data included for reference.

entailment data points in HANS, the models do not treat them as identical—contrary to what accuracy in Table 1 suggests. For ST-LM, ST-N, ST-WO, and SICK, LS is slightly less pronounced compared to HANS. The most striking result comes from ST-NU (numerical reasoning), where all baseline models achieve over 95% LS. Although the classifier head's poor accuracy suggests that models struggle with numerical reasoning, the high LS indicates that they effectively capture the necessary information for this task.

Accuracy Paradox While classifier head accuracy suggests that ST-NU and HANS are difficult, and ST-WO and ST-N are easier, the representation space reveals the opposite. ST-WO and ST-N are as challenging as the MNLI validation set, while ST-NU and HANS are much easier. Notebly, ST-WO, ST-LN, and ST-LM, which are derived from the MNLI validation set with some modifications, exhibit LS values similar to MNLI itself. This is an interesting finding, as it suggests that these subsets, being structurally similar to MNLI, pose comparable challenges for the model. Since they are as difficult as MNLI, their LS does not exceed MNLI accuracy or reach the high LS values observed in easier datasets like HANS, SICK, and ST-NU.

4 LS as Evidence of Encoder Knowledge

In the previous section, we observed that despite the NLI model's poor accuracy on the analytical dataset, their encoder's outputs remain nearly linearly separable. In this section we argue that low accuracy does not necessarily indicate a lack of NLI or linguistic knowledge. Instead, our results highlight a misalignment between the encoder's learned representations and the classifier head's decision boundaries.

4.1 LS and Training Dynamics

In traditional machine learning, feature engineering was guided by domain experts who carefully crafted features based on their deep understanding of the task. These features were designed to effectively differentiate between classes, making them easy to separate with a simple MLP. In contrast, transformer models delegate this responsibility to

Model	MNLI	H+	H-	HANS
BERT_{Full}	84.7±0.2	97.7±1.2	32.4±5.5	65.0±2.6
$\text{BERT}_{Balanced}$	81.6±0.4	79.1±3.5	48.8±3.4	63.9±1.2

Table 3: Comparison of BERT model accuracy when fine-tuned on the full MNLI dataset (with $392K$ samples) and the balanced dataset (with $235K$ samples). The mean accuracy is reported over 5 different seeds.

the encoder, which is tasked with generating meaningful representations from raw input data. The classifier head, on the other hand, merely maps these representations to labels without any inherent understanding of the task itself. If a model truly grasps the underlying task, this understanding should be reflected in the features produced by the encoder. The fact that the encoder can generate linearly separable features, even for datasets that differ significantly from the training data, suggests that it has captured genuine, task-relevant knowledge. Moreover, we demonstrate that this LS is not just an artifact of the model's representation but also correlates with its process of acquiring knowledge during training. By varying the amount of training data and limiting the number of update steps, we explored the relationship between task understanding (as reflected by standard validation set accuracy) and LS of analysis dataset, with the following findings:

- **Effect of Training Data Size:** Fine-tuning BERT on varying proportions of the MNLI dataset (from 5% to 100%) revealed a clear trend, as the amount of training data increased, LS improved for both the MNLI validation set and the HANS dataset (Figure 3).
- **Effect of Training Iterations:** Similarly, tracking the model's performance on the full dataset at 500-step intervals (Figure 9 in the Appendix) showed that as validation accuracy increased, the LS of analysis datasets also improved.

These findings suggest that as the model refines its understanding of the NLI task, it simultaneously enhances its ability to produce clearer and more distinguishable representations, reinforcing the connection between knowledge acquisition and LS.

4.2 Re-evaluating the Lexical Overlap Bias in NLI Models

One common argument against NLI models achieving true linguistic mastery is their poor performance on heuristic-based datasets. This is often cited as evidence that these models rely on shortcuts in the training data rather than acquiring genuine linguistic knowledge. HANS, as a prominent example of such datasets, is frequently used to support this claim due to its design, which specifically targets lexical overlap heuristics. Since we argue that the model does acquire sufficient linguistic knowledge, we challenge this assumption by conducting an experiment to remove the potential influence of lexical overlap bias and examine whether the model's performance improves.

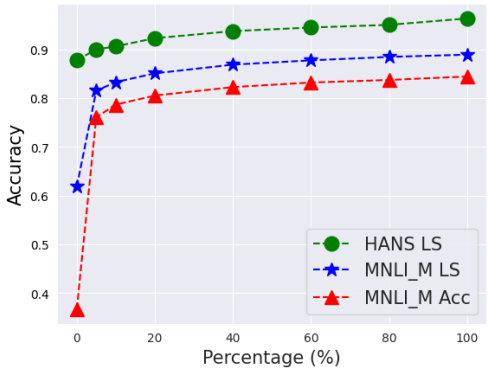


Figure 3: LS of the HANS and MNLI (matched) datasets for BERT fine-tuned on different percentages of the MNLI dataset, along with model accuracy. The consistent rise in LS alongside accuracy shows that improved LS is not incidental but emerges as the model learns the task more deeply with more data; reflecting the accumulation of generalized, task-relevant knowledge in the encoder.

To explore this, we calculated the overlap percentage for all training examples and grouped them into 100 bins, each representing a 1% range (e.g., [88, 89) overlap). Within each bin, we ensured an equal distribution of examples across all three labels by selecting a balanced number of samples from the least frequent label. This process eliminated label imbalance across different levels of lexical overlap, as shown in Figure 4. Using this balanced dataset, we fine-tuned a BERT model for five epochs, with results reported in Table 3. While accuracy on HANS− improved, this came at the cost of decreased accuracy on HANS+, leading to an overall drop in HANS performance compared to the model trained on the full MNLI dataset. Figure 5 visualizes the representation space of the `[CLS]` tokens from the model trained on the balanced dataset. The HANS representations remain largely clustered together within the entailment region, rather than forming distinct groups. If the

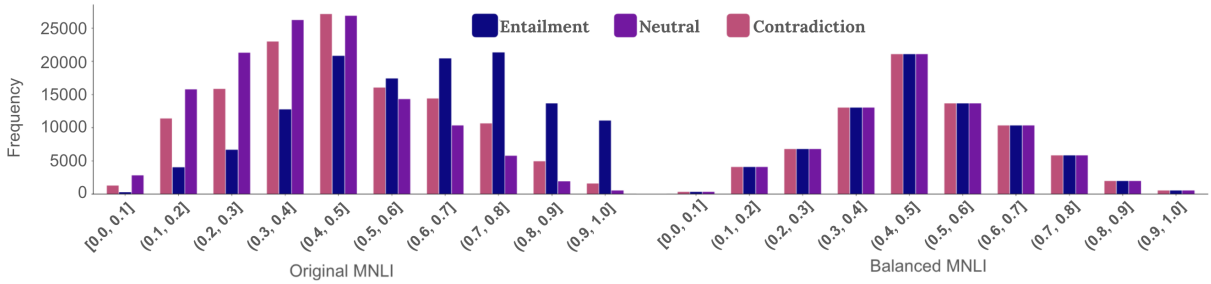


Figure 4: Histograms of label frequency across different overlap percentages, before and after balancing the dataset. The original experiment used 100 bins, but for the sake of space, we present both histograms with 10 bins.

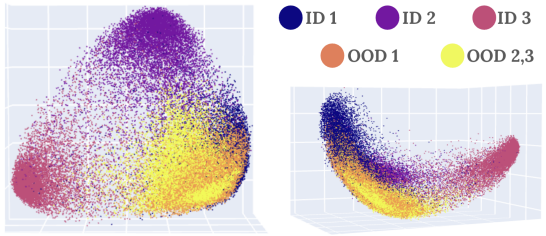


Figure 5: Visualization of the [CLS] representation space for BERT fine-tuned on the blended MNLI dataset. Colors indicate the gold labels.

overlap heuristic was the primary cause of the bias, balancing the dataset should have improved the results.

4.3 Effect of Random Seed on Performance

Prior works (McCoy et al., 2020; Zhou et al., 2020) have reported that models trained on standard NLI datasets exhibit consistent in-domain (ID) validation performance across different random seeds, yet their performance on challenge datasets (OOD cases) such as HANS fluctuates significantly. In some subsets of HANS, accuracy varies between 0% and 66% depending on the seed. As shown in Table 1, accuracy variance is large for HANS and ST-NU, whereas the MNLI validation set shows almost no variance. Notably, these results are based on only five random seeds; increasing the number of trials would likely reveal even greater variance.

Based on these results, prior work suggests that while the model consistently learns patterns that perform well on the validation set, its generalization to OOD or adversarial cases is unstable. However, a closer analysis points to an alternative explanation. The encoder, which encodes linguistic knowledge, exhibits a high degree of consistency across random seeds. Its representations maintain LS even for adversarial inputs, regardless of initialization. In Table 4 we compare the accuracy and LS of two BERT models with two different initial seeds, one with very poor HANS performance and one with very strong performance, yet their encoder representations remain distinguishable in the same way. This suggests that the encoder reliably captures task-relevant linguistic features which are preserved across seeds.

Trial	HANS Accuracy	Linear Separability
High-performing	67.6	95.5
Low-performing	52.8	95.0

Table 4: Comparison of HANS accuracy and LS for a high-performing and a low-performing trials.

Instead, the classifier head—a shallow, randomly initialized MLP—is highly sensitive to weight initialization. Different random seeds result in divergent decision boundaries within the encoder's representation space. While these boundaries work well for ID validation data (MNLI), they fail to generalize to OOD datasets like HANS. This is because the classifier is primarily optimized for MNLI's feature distribution, which does not necessarily align with the structure of adversarial or OOD samples.

Thus, rather than instability arising from differences in learned knowledge, it stems from the classifier's inconsistent mapping of the encoder's representations, leading to poor generalization beyond the training domain.

5 Discussion

We have shown that despite the poor and unstable performance of NLI models on OOD datasets, the encoder representations of these datasets remain consistently and highly discriminative with respect to class labels. This suggests that the model acquires core linguistic knowledge relevant to the

NLI task that generalizes beyond the training distribution. If this were not the case, it would be unclear why the encoder organizes unseen data in a way that permits linear separation. Notably, this behavior does not universally occur for all types of OOD data; it stands in contrast to tasks such as paraphrase detection, where the encoder often fails to produce similarly structured representations.

For example, QQP is a standard benchmark for paraphrase detection, while PAWS (Zhang et al., 2019) was introduced to challenge models that rely on shallow heuristics such as word overlap. A BERT model fine-tuned on QQP performs poorly on PAWS, misclassifying most examples as paraphrases, despite nearly half being non-paraphrases. In this case, LS is close to random—61.4% compared to 57.2%—and the PAWS examples appear scattered within the QQP duplicate region in the representation space, as shown in Figure 6.

As discussed in Section 4.3, one potential explanation for why encoder representations can be discriminative for OOD datasets despite low accuracy is that, from the perspective of the PLM, the MNLI dataset occupies a distinct and well-defined region in the representation space, whereas analytical datasets reside elsewhere (Figure 7). During fine-tuning, the encoder and classifier head are updated jointly to establish decision boundaries. However, this optimization process focuses only on MNLI training examples, which are explicitly supervised. As a result, the encoder is shaped to structure MNLI data effectively while ignoring how these changes affect other parts of the space. Since OOD data are not included during training, misalignments in those regions incur no penalty, leading to reduced generalization performance.

It is important to emphasize that the LS values we report are not the result of any additional training. Rather, they reflect the decision boundaries already present in the representation space after fine-tuning. This distinction is critical, as it rules out multitask learning as a source of the observed patterns. In multitask learning, the encoder is jointly trained on multiple objectives, encouraging knowledge sharing across tasks. In our setting, however, the encoder is fine-tuned solely on MNLI, and the analysis datasets are never seen during training. We simply train a linear classifier on frozen representations using cross-entropy loss, thereby probing the task-relevant structure already encoded by the model.

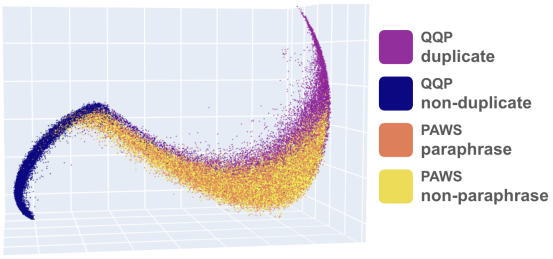


Figure 6: Representational space of the [CLS] token generated by the BERT model fine-tuned on the QQP dataset.

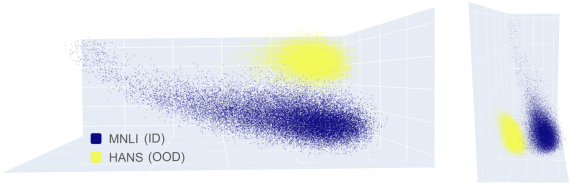


Figure 7: Visualization of the [CLS] representation of the MNLI training set and HANS from the perspective of pre-trained BERT.

6 Related Work

6.1 Probing Knowledge

Probing the representation space of PLMs has been central to understanding the knowledge they encode. Early studies analyzed layer-wise representations to identify where syntactic and semantic information is captured, revealing a hierarchical organization of linguistic features (Liu et al., 2019a; Jawahar et al., 2019; Tenney et al., 2019). Follow-up work examined attention mechanisms, showing that specific attention heads specialize in tasks such as coreference and syntax (Clark et al., 2019; Voita et al., 2019). Other approaches explored the geometry of the representation space, finding that upper layers tend to produce more context-specific, anisotropic embeddings (Ethayarajh, 2019). While initial work focused on static PLMs, later studies investigated how fine-tuning alters representations, showing that core structural properties often remain stable despite task-specific adaptations (Merchant et al., 2020; Zhou and Srikumar, 2022).

6.2 Discrepancies Between Final Predictions and Model Representations

Yaghoobzadeh et al. (2021) showed that adjusting the classification threshold for HANS data can significantly impact BERT's accuracy. This phenomenon is not exclusive to encoder models, gen-

erative models also exhibit discrepancies between what they learn and what their final outputs imply. Zhao et al. (2021b) highlighted a similar issue in generative models, showing that the structure of a prompt can influence the threshold required for classification tasks such as sentiment analysis. By calibrating models with a null input, they achieved more reliable results. Amini and Ciaramita (2023) argue that the sensitivity of encoder-decoder model to instruction phrasing stems from the constraint that models must verbalize their predictions. By bypassing the decoding step and directly probing the encoder representations, they achieved more stable and improved results. Marks and Tegmark (2024). Furthermore, Marks and Tegmark (2024) found that LLMs encode the truth or falsehood of factual statements in a linear manner, despite their tendency to generate incorrect information.

6.3 Instability in OOD Generalization

Models that appear stable and performant on standard ID test sets often exhibit significant variability when evaluated on OOD datasets (McCoy et al., 2020; Zhou et al., 2020), raising concerns about their generalization capabilities. Similarly, Zhao et al. (2021b) demonstrated that even powerful generative models like GPT-3 suffer from notable instability in few-shot learning scenarios. This instability has been attributed to several factors, including catastrophic forgetting during fine-tuning (Lee et al., 2020), limited size and diversity of available datasets (Dodge et al., 2020), and optimization difficulties such as vanishing gradients in deeper architectures (Mosbach et al., 2021). In addition to these architectural and data-related challenges, the structure of prompts and the order in which training examples are presented have also been shown to significantly influence performance in few-shot settings (Zhao et al., 2021b), highlighting the sensitivity of model behavior to seemingly minor variations in input.

7 Conclusion

In this paper, we revisited the performance of fine-tuned PLMs on challenging NLI datasets. Our experiments revealed that, despite poor classifier accuracy, the encoder's representation space often demonstrates clear linear separability between classes. This suggests that the models possess relevant task-specific knowledge, but there is a misalignment between the classifier's decision boundaries and the knowledge embedded in the encoder's representations. While we proposed some hypotheses for this misalignment, further in-depth investigation is required, which we leave for future work.

8 Limitations

One limitation of this study is that the analysis was limited to three pretrained language models—DistilBERT, BERT, and RoBERTa. While these models are widely used, they do not represent the full spectrum of transformer-based models, and therefore, the findings may not be fully generalizable to newer or more specialized models. Additionally, this study does not provide a direct solution for improving classification accuracy. Although we demonstrate the existence of linear boundaries, determining the optimal decision boundaries for each dataset still requires access to the full dataset, which may not be efficient or feasible for OOD datasets. Furthermore, relying on linear separability as a proxy for model knowledge may oversimplify the complexity of how models truly understand the nuances of inference. There is room for further exploration using alternative probing techniques to assess and deepen our understanding of model comprehension.

References

Afra Amini and Massimiliano Ciaramita. 2023. In-context probing: Toward building robust classifiers via probing large language models. *Preprint*, arXiv:2305.14171.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages

4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *Preprint*, arXiv:2002.06305.

Luciano Dyballa, Evan Gerritz, and Steven W. Zucker. 2024. A separability-based approach to quantifying generalization: which layer is best? *Preprint*, arXiv:2405.01524.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Cheolhyoung Lee, Kyunghyun Cho, and Wanmo Kang. 2020. Mixout: Effective regularization to finetune large-scale pretrained language models. In *International Conference on Learning Representations*.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Samuel Marks and Max Tegmark. 2024. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *Preprint*, arXiv:2310.06824.

R. Thomas McCoy, Junghyun Min, and Tal Linzen. 2020. BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 217–227, Online. Association for Computational Linguistics.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. 2020. What happens to BERT embeddings during fine-tuning? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 33–44, Online. Association for Computational Linguistics.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. On the stability of fine-tuning {bert}: Misconceptions, explanations, and strong baselines. In *International Conference on Learning Representations*.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018a. Stress test evaluation for natural language inference. In *The 27th International Conference on Computational Linguistics (COLING)*, Santa Fe, New Mexico, USA.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018b. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Abhilasha Ravichander, Aakanksha Naik, Carolyn Rose, and Eduard Hovy. 2019. EQUATE: A benchmark evaluation framework for quantitative reasoning in natural language inference. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 349–361, Hong Kong, China. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head

self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *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. Association for Computational Linguistics.

Yadollah Yaghoobzadeh, Soroush Mehri, Remi Tachet des Combes, T. J. Hazen, and Alessandro Sordoni. 2021. Increasing robustness to spurious correlations using forgettable examples. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3319–3332, Online. Association for Computational Linguistics.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021a. Calibrate before use: Improving few-shot performance of language models. *Preprint*, arXiv:2102.09690.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021b. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Xiang Zhou, Yixin Nie, Hao Tan, and Mohit Bansal. 2020. The curse of performance instability in analysis datasets: Consequences, source, and suggestions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8215–8228, Online. Association for Computational Linguistics.

Yichu Zhou and Vivek Srikumar. 2022. A closer look at how fine-tuning changes BERT. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1046–1061, Dublin, Ireland. Association for Computational Linguistics.

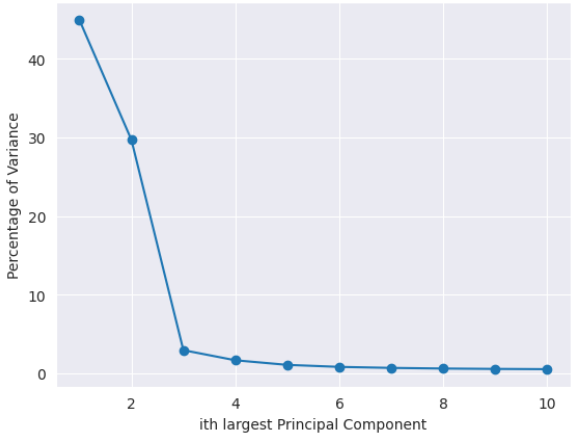


Figure 8: The percentage of variance explained by each of the 10 greatest principal components of PCA applied to the `[CLS]` tokens from the HANS and MNLI training sets.

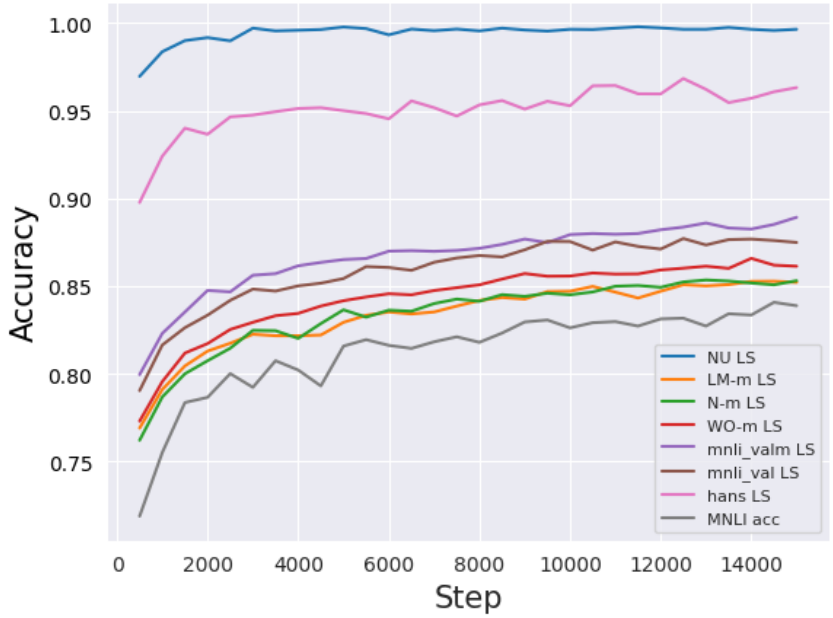


Figure 9: Linear separability (LS) of the analysis datasets along with the accuracy of BERT reported at every 500 optimization steps.

Heuristic	**Hypothesis**	**Premise**	**Label**
Lexical Overlap	The banker near the judge saw the actor.	The banker saw the actor.	E
	The lawyer was advised by the actor.	The actor advised the lawyer.	E
Subsequence	The doctors visited the lawyer.	The lawyer visited the doctors.	N
	The judge by the actor stopped the banker.	The banker stopped the actor.	N
	The artist and the student called the judge.	The student called the judge.	E
	Angry tourists helped the lawyer.	Tourists helped the lawyer.	E
	The judges heard the actors resigned.	The judges heard the actors.	N
	The senator near the lawyer danced.	The lawyer danced.	N
Constituent	Before the actor slept, the senator ran.	The actor slept.	E
	The lawyer knew that the judges shouted.	The judges shouted.	E
	If the actor slept, the judge saw the artist.	The actor slept.	N
	The lawyers resigned, or the artist slept.	The artist slept.	N

Table 5: Examples from the HANS dataset, where E represents the entailment label and N represents the non-entailment label.

Premise	Hypothesis	Label
A woman is painting a man's face	Makeup is being put on the man by the woman	Entailment
The women are singing and dancing	The persons are singing and dancing	Entailment
A group of kids is playing in a yard and an old man is standing in the background	A group of boys in a yard is playing and a man is standing in the background	Neutral
A black, red, white and pink dress is being worn by a woman	A woman is taking off a cloak, which is very large, and revealing extravagant apparel	Neutral
Two large persons are sitting on a park bench and they have a bottle of soda between them	Two large persons are standing near a park bench and they have nothing between them	Contradiction
A man is playing a violin	There is no man playing a violin	Contradiction

Table 6: Examples from the SICK dataset.

Subset	Premise	Hypothesis	Label
LM	The chain wielder smiled at her and true is true and true is true and true is true and true is true and true is true.	The chain wielder was friends with the woman.	Neutral
	The two programs are currently housed in buildings about a block apart and true is true and true is true and true is true and true is true and true is true	The two buildings are on opposite sides of the city.	Contradiction
	oh i did and i laughed real hard when i took it in for the two thousand mile checkup and uh and true is true and true is true and true is true and true is true and true is true	I had a good laugh when I took it for the two thousand mile checkup.	Entailment
WO	I can't do any jumping up and down because it makes it hurt.	The pain is too much after jumping, it needs surgery and true is true.	Neutral
	Both professors soon realized that creating a new language was not an easy task.	Professors realized it was easy to make a new language and true is true.	Contradiction
	Well that's good that's great.	That is not bad, it is actually pretty good and true is true.	Entailment
N	They did this to us.	The practical joke was played on them and false is not true.	Neutral
	No, Dave Hanson, you were too important to us for that.	Yes, Dave Man, you are not important to us and false is not true.	Contradiction
	Then he sobered.	He had sobered up and false is not true.	Entailment
NU	Jose joined him less than 6 months later, investing Rs.	Jose joined him 2 months later, investing Rs.	Neutral
	On every Adidas shoe the restaurant makes a profit of $ 10 while on every Puma shoe it makes a profit of $ 8	On every Adidas shoe the restaurant makes a profit of $ more than 10 while on every Puma shoe it makes a profit of $ 8	Contradiction
	A train leaves Delhi at 9 a.m.	A train leaves Delhi at more than 3 a.m.	Entailment

Table 7: Examples from different subsets of the stress test dataset.

Dataset	DistilBERT	BERT	RoBERTa
SNLI	89.3±0.1	90.9±0.3	91.8±0.1
SICK	53.5±1.0	56.6±0.3	57.1±0.4
HAN	52.9±0.6	58.9±1.1	66.6±1.0
ST-NU	35.3±0.7	37.8±4.6	38.1±2.1
ST-LM	65.1±0.7	70.6±0.6	76.6±0.2
ST-N	45.8±2.3	51.4±2.5	63.4±1.6
ST-WO	56.7±3.6	59.2±2.6	69.8±1.8

Table 8: Accuracy of SNLI fine-tuned models on NLI analysis datasets, SICK, HANS, and Stress Test (ST) alongside the standard validation sets of SNLI.