Semantic Sensitivities and Inconsistent Predictions: Measuring the Fragility of NLI Models

Erik Arakelyan[†], **Zhaoqi Liu**[†], **Isabelle Augenstein**[*]
Department of Computer Science
University of Copenhagen
Copenhagen Denmark
{erik.a,nbq899,augenstein}@di.ku.dk

Abstract

Recent studies of the emergent capabilities of transformer-based Natural Language Understanding (NLU) models have indicated that they have an understanding of lexical and compositional semantics. We provide evidence that suggests these claims should be taken with a grain of salt: we find that state-of-the-art Natural Language Inference (NLI) models are sensitive towards minor semantics preserving surface-form variations, which lead to sizable inconsistent model decisions during inference. Notably, this behaviour differs from valid and in-depth comprehension of compositional semantics, however does neither emerge when evaluating model accuracy on standard benchmarks nor when probing for syntactic, monotonic, and logically robust reasoning. We propose a novel framework to measure the extent of semantic sensitivity. To this end, we evaluate NLI models on adversarially generated examples containing minor semantics-preserving surface-form input noise. This is achieved using conditional text generation, with the explicit condition that the NLI model predicts the relationship between the original and adversarial inputs as a symmetric equivalence entailment. We systematically study the effects of the phenomenon across NLI models for *in-* and *out-of* domain settings. Our experiments show that semantic sensitivity causes performance degradations of 12.92% and 23.71% average over *in-* and *out-of-* domain settings, respectively. We further perform ablation studies, analysing this phenomenon across models, datasets, and variations in inference and show that semantic sensitivity can lead to major inconsistency within model predictions.

1 Introduction

Transformer-based (Vaswani et al., 2017) Language Models (LMs) have shown solid performance across various NLU tasks (Wang et al., 2018, 2019). These advances have led to suggestions regarding the emergent capabilities of the models in terms of syntactic (Sinha et al., 2020; Hewitt and Manning, 2019; Jawahar et al., 2019; Warstadt and Bowman, 2020), logic (Wei et al., 2022a,b) and semantic (Kojima et al., 2022; Dasgupta et al., 2022) understanding. However, we present novel evidence that indicates that these models are prone to inconsistent predictions induced by inherent susceptibility towards semantic sensitivities.

To probe the models for these discrepancies, we formalise *semantic comprehension* as the ability to distinguish logical relations within sentences through identifying compositional semantics (Jacobson, 2014; Carnap, 1959). This means that negligible semantic variations should not impact the inherent relations implied between the texts, e.g. *"There were beads of perspiration on his brow."* entails both *"Sweat built up upon his face."* and the slight variation *"The sweat had built up on his face."* Authentic comprehension of semantics does allow for such understanding through discovering semantic structures and the inherent relations induced by them (Cicourel, 1991; Schiffer, 1986; Rommers et al., 2013). This means that analysing the emergent semantic understanding within a model should minimally involve testing for sensitivity towards semantics-preserving surface-form variations.

We particularly focus on the task of textual entailment (Dagan et al., 2005), otherwise referred to as Natural Language Inference (Bowman et al., 2015, NLI), which has been widely used to probe how well the models understand language (Condoravdi et al., 2003; Williams et al., 2017; Nie et al., 2019). This is a pairwise input task, where given a premise p and a hypothesis h, the objective is to predict if the premise *entails, contradicts* or is *neutral* towards the hypothesis.

We propose a framework for testing semantic sensitivity within transformer-based models trained for NLI, by creating semantics-preserving surface-

[†]Equal contribution, alphabetical order. [*]Senior author.

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 432–444

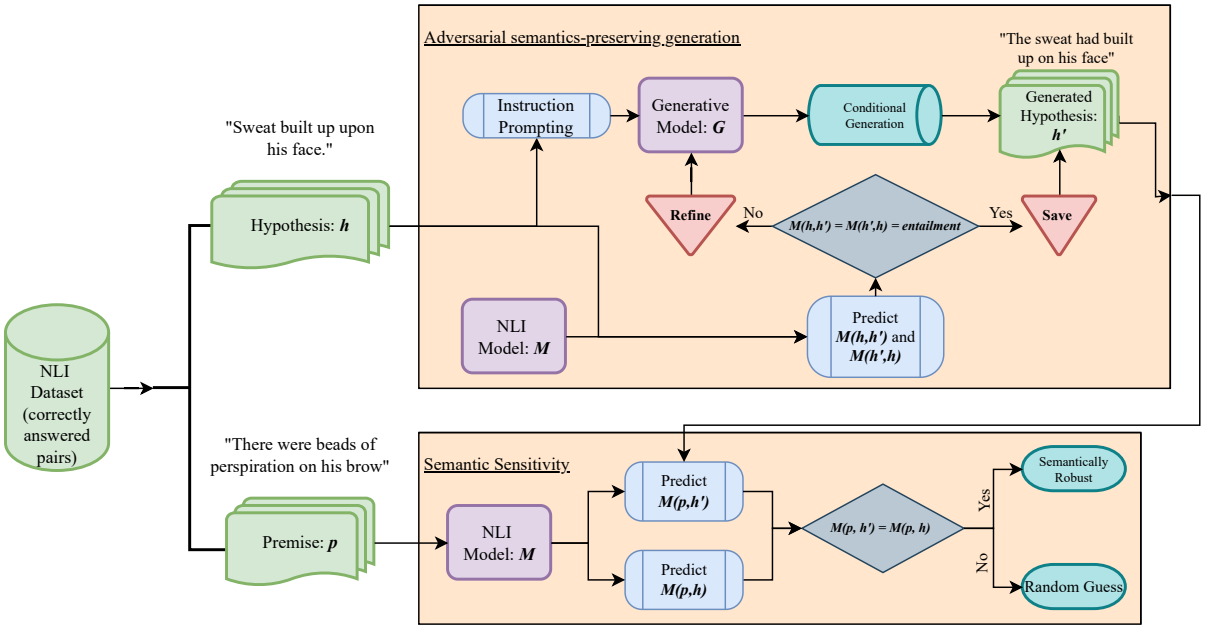

Figure 1: The proposed framework is comprised of two components. (i) a module for generating semantics-preserving surface-form hypothesis variations and (ii) using the generated surface for measuring semantic sensitivity and predictive inconsistency.

form variations of the hypothesis (see Figure 1). These variations are created using conditional generation with Large Language Models (LLMs). We show that proposed candidates do not alter the core meaning or the truth value compared to the original statement. The original and generated sentences maintain denotative equivalence, where two sentences or phrases might be interpreted as having the same truth value or factual content but may carry minor variations of nuances or connotations. To ensure that the relations are preserved within the candidates during conditional generation, we assert that the NLI model predicts the original and generated hypothesis to symmetrically entail each other. This indicates that the model perceives both the generated and original hypothesis as equivalent. After introducing these variations, we evaluate the NLI model by replacing the original hypothesis with the generated candidates. As the candidates are indicated to be equivalent by the same NLI model, this evaluation will indicate whether the model can recover the existent relation between the premise hypothesis pair in the presence of minor semantic-preserving noise. We use the samples where the model identifies the existing relation correctly from the original premise hypothesis pair. This ensures that assessing for semantic sensitivity would not be hindered by the discrepancies in model performance.

We systematically study the semantic sensitivity across transformers that achieve state-of-the-art or similar results when trained on NLI datasets, namely RoBERTa (Liu et al., 2019b), BART (Lewis et al., 2019), DeBERTa (He et al., 2020) and DistilBart (Sanh et al., 2019; Lewis et al., 2019) with different parametrizations. To measure the effect of the phenomenon on the inconsistency of the predictions, we use three popular English datasets - MultiNLI (Williams et al., 2017, MNLI), SNLI (Bowman et al., 2015) and ANLI (Nie et al., 2019). The models are fine-tuned using MNLI, which we choose for *in-domain* testing, as it covers a wide range of topics and is frequently used for zero-shot and few-shot textual classification (Yin et al., 2019). We use the same models for *out-of-domain* evaluation across the other NLI datasets.

Our contributions are as follows: (i) we propose a novel framework for assessing semantic sensitivity within transformer-based language models (ii) we systematically study the influence of this phenomenon on inconsistent predictions across various transformer variants (iii) we show that the effect is persistent and pronounced across both *in-* and *out-of-domain* evaluations (iv) we further complete ablations to assess the severity of the inconsistent predictions caused by semantic sensitivity.

2 Related Work

Semantic comprehension is considered a fundamental building block for language understanding (Allen, 1995). Although attempts have been made to probe language models in terms of compositional semantic capabilities, the conclusions regarding their emergence remain to be discussed.

Models appear to understand semantics Recently a wide suite of tasks has been proposed for testing models for language understanding (Wang et al., 2019; Zellers et al., 2018; Ribeiro et al., 2020) with the credence that a model with strong performance should be able to utilise semantic relations when completing the tasks. In light of these, it has been shown that transformer-based language models can be directly trained (Zhang et al., 2020; Rosset et al., 2020) to utilise semantic structure to gain distributional information within the task. Specifically, NLI models have also been shown to be capable of pragmatic inferences (Jeretic et al., 2020a) with a perception of implicature (Grice, 1975) and presupposition (Stalnaker et al., 1977; Grice, 1975).

Models struggle with semantics Directly probing for a specific aspect of semantic understanding has shown that transformer-based language models tend to struggle with semantics (Belinkov, 2022). It has been indicated that pretraining the language models does not exploit semantic information for entity labeling and coreference resolution (Liu et al., 2019a). Furthermore, transformer attention heads only minimally capture semantic relations (Kovaleva et al., 2019) from FrameNet (Baker et al., 1998). Studies have also shown that NLI models, in particular, tend to struggle with lexical variations, including word replacements (Glockner et al., 2018; Ivan Sanchez Carmona et al., 2018; Geiger et al., 2020), and sequence permutations (Sinha et al., 2021).

Sensitivity in NLI models Probing NLI models for language understanding has been a hallmark testing ground for measuring their emerging capabilities (Naik et al., 2018a; Wang and Jiang, 2015; Williams et al., 2017). A wide range of tests indicates that models trained for NLI are prone to struggling with syntax and linguistic phenomena (Dasgupta et al., 2018; Naik et al., 2018b; An et al., 2019; Ravichander et al., 2019; Jeretic et al., 2020b). It has also been shown that NLI models heavily rely on lexical overlaps (Ivan Sanchez Carmona et al., 2018; McCoy et al., 2019; Naik et al., 2018b) and are susceptible to over-attending to particular words for prediction (Gururangan et al., 2018; Clark et al., 2019). Our line of work is associated with evaluating NLI models for monotonicity reasoning (Yanaka et al., 2019) and sensitivity towards specific semantic phenomenon (Richardson et al., 2020), such as boolean coordination, quantification, etc. However, we systematically test NLI models for their compositional semantic abilities and measuring the degree of inconsistence of their predictions influenced by the phenomenon.

3 Methodology

We aim to create a framework for assessing semantic sensitivity within NLI models and measure its impact on the inconsistence of model predictions. The first part of the pipeline we propose is an adversarial semantics-preserving generation for introducing variations within the original samples. The second part of the pipeline involves assessment using the acquired generations.

3.1 Semantics Preserving Surface-Form Variations

We formalise NLI as a pairwise input classification task. Given a dataset of premise hypothesis pairs $\mathcal{D} = (p_1, h_1), \ldots (p_n, h_n)$, where $\forall p_i \in P \;\&\; h_i \in H$ are a set of textual tokens $P, H \subseteq \mathcal{T}$, the goal is to classify the pairs as *entailment, contradiction* or *neutrality*, i.e. $\mathcal{C} = \{E, C, N\}$. We are also given a pre-trained language model (PLM) $\mathcal{M}$ that is trained for textual entailment. Before introducing semantic variations, only the samples where model $\mathcal{M}$ predicted the label correctly are filtered, i.e. $D_{\text{correct}} = \{\forall (p_i, h_i) \in \mathcal{D} : \mathcal{M}(p_i, h_i) = \hat{y} = y\}$, where $\hat{y}$ is the prediction and y is the original label. This is completed to ensure that the evaluation of semantic sensitivity is not hindered or inflated by the predictive performance and confidence of the model $\mathcal{M}$. This type of filtering is used when probing for emergent syntactic (Sinha et al., 2021), lexical (Jeretic et al., 2020b), and numerical (Wallace et al., 2019) reasoning capabilities. We can see the original accuracy of NLI models and the number of samples used in the study in Table 1.

To introduce semantics preserving noise within chosen samples, we complete a two-fold refinement process. We utilise a generative LLM $\mathcal{G}$, which has

	bart-l	roberta-l	distilbart	deberta-b	deberta-l	deberta-xl
MNLI$_{(n=10000)}$	90.10%	90.56%	87.17%	88.77%	91.32%	91.44%
SNLI$_{(n=10000)}$	87.55%	86.44%	84.37%	84.39%	88.87%	88.54%
ANLI_r1$_{(n=1000)}$	46.20%	46.40%	41.40%	35.10%	49.70%	53.00%
ANLI_r2$_{(n=1000)}$	31.60%	27.00%	32.80%	29.80%	32.70%	35.40%
ANLI_r3$_{(n=1200)}$	33.08%	26.75%	32.75%	30.50%	35.92%	38.75%

Table 1: The original accuracy on testing/dev sets for various transformers (b-base, l-large, xl-extra large) on *in-domain* MNLI experiments and zero-shot transfers to *out-of-domain* SNLI and ANLI. The number near the dataset name designates the exact amount of original samples in the testing set.

been fine-tuned on natural language instructions (Wei et al., 2021; Chung et al., 2022), and prompt it to paraphrase the original hypothesis h_i, with the following prompt: *Rephrase the following sentence while preserving its original meaning: <h_i>*. This is not sufficient to produce semantics-preserving variations as generative models are prone to hallucinations (Ji et al., 2023) and not assured to produce an equivalent paraphrase. To ensure that the generation h_i' is logically equivalent to the original sample and thus semantics-preserving, we impose the condition that the NLI model should infer the relation between the original and generated hypothesis as a symmetric entailment:

$$\mathcal{M}(h_i, h_i') = \hat{y}_{\mathcal{C}=E} = \mathcal{M}(h_i', h) \tag{1}$$

The bidirectional nature of entailment allows us to claim that sentences are logically equivalent (Angell, 1989; Clark, 1967). We refine the proposed variation candidates using the generator $\mathcal{G}$ until k candidates that satisfy the condition are produced.

Human Evaluation of Surface-Form Variations To further ensure the validity of this variation generation method, we conduct a human evaluation of the generated samples. We randomly sample 100 examples of generated and original hypothesis pairs across all datasets and employ two annotators to assess whether the sentences are semantically and logically equivalent within the pair. Our results show that in 99% of the cases, the annotators marked the samples as equivalent with an inter-annotator agreement measure of Cohen's $\kappa = 0.94$. This further shows the reliability of the method for generating semantics-preserving surface form variations. We provide further token overlap level analysis in Appendix A.

3.2 Evaluating Semantic Sensitivity

After obtaining k semantic variations for each hypothesis, we test the semantic sensitivity of the model by replacing the original hypothesis h_i with the candidates $\{h_i'^1, \dots h_i'^k\}$ and making a prediction with the NLI model $\mathcal{M}$. As the proposed variations are logically equivalent to the original, we want to test if the new model prediction would vary compared to the original.

$$\mathcal{R}(p_i, h_i, h_i'^j, \mathcal{O}) = \\ = \begin{cases} 1, \mathcal{O}(\mathcal{M}(p_i, h_i), \mathcal{M}(p_i, h_i'^j)) = 0 \\ 0, \mathcal{O}(\mathcal{M}(p_i, h_i), \mathcal{M}(p_i, h_i'^j)) = 1 \end{cases} \tag{2}$$

Here $\mathcal{O} : \mathcal{C} \times \mathcal{C} \rightarrow \{0, 1\}$ is a boolean matching operator between the labels predicted with original hypothesis h_i and the surface-form variations $h_i'^j$. A change in the label would imply that the model is semantically sensitive and the original correct prediction is inconsistent with the label produced for the semantics preserving surface-form variation. A graphical representation can be seen in Figure 5. We use two metrics to measure semantic sensitivity within NLI models, both of which are derivative formulations of a Fooling Rate (Moosavi-Dezfooli et al., 2017), which is used for assessing the success of adversarial attacks (Chakraborty et al., 2018). Given k possible surface-form variations for the hypothesis, we test if at least one of the candidates would be able to cause a label change compared to the original prediction, which can be formalised as:

$$r_r = \frac{\sum_i^{n'} \mathbb{1}\left[\exists j \in [1, k], \mathcal{R}(p_i, h_i, h_i'^j, =) \neq 1\right]}{n'}. \tag{3}$$

r_s/r_r	bart-large	roberta-large	distilbart	deberta-base	deberta-large	deberta-xlarge
MNLI	6.64%/12.35%	5.71%/11.56%	9.20%/ **16.80%**	6.66%/13.81%	5.38%/11.54%	5.89%/11.49%
SNLI	10.11%/15.52%	8.38%/14.98%	15.67%/**23.68%**	9.96%/17.01%	7.83%/13.39%	9.50%/14.69%
ANLI_r1	31.51%/42.89%	28.45%/35.01%	31.48%/**52.30%**	40.0%/48.99%	25.66%/37.88%	22.71%/30.73%
ANLI_r2	34.39%/51.91%	24.62%/42.80%	36.09%/**57.49%**	34.92%/48.47%	28.44%/44.04%	29.46%/46.46%
ANLI_r3	29.11%/51.39%	21.88%/45.00%	29.26%/52.42%	33.88%/**53.17%**	24.88%/44.65%	23.23%/42.37%

Table 2: The strict and relaxed fooling rates of different transformer models across *in-domain* (MNLI) and *out-of-domain* (SNLI, ANLI) evaluations. On average more than half of the labels change towards their logically contrasting counterpart.

Here n' is the number of correctly answered original samples, and the matching operator $\mathcal{O}$ is a simple equality checking operator "=". We refer to this metric as a relaxed Fooling Rate. To measure more drastic label changes, i.e. *entailment* to *contradiction* and vice versa, we also define a stricter version of Equation 3.

$$r_s = \frac{\sum_i^{n'} \mathbb{1}\left[\exists j \in [1,k], \mathcal{R}(p_i, h_i, h_i'^j, =^s) \neq 1\right]}{n'}. \tag{4}$$

We replace standard equality for the operator $\mathcal{O}$ in Equation 3 with a strict counterpart that matches only if the predictions are direct opposites, i.e. *entailment* $\leftrightarrow$ *contradiction*. It must be noted that the *neutral* class does not have a direct opposite; thus, the metric for this label remains unchanged. It can be concluded that the inequality $r_s \leq r_r \leq 1$ trivially holds when using these metrics.

4 Experimental Setup

4.1 Model Details

Semantics preserving Generation To generate and refine semantic variations of the original hypothesis, we chose *flan-t5-xl* as the generation model $\mathcal{G}$. It is an instruction-tuned LLM that has shown close state-of-the-art performance in tasks such as paraphrasing, zero and few shot generation, chain of thought reasoning (CoT), and multi-task language understanding (Chung et al., 2022). For each of the selected hypotheses, we produce $k = 5$ unique semantics-preserving variations. To ensure diversity and consistency of the generated text while avoiding computationally expensive exhaustive search, we use a group beam search (Vijayakumar et al., 2016) with a temperature $t \in [0.3, 0.6]$ and a maximum output of 40 tokens throughout the generation and refinement procedure. We also further diversify the generation by using the recipe from Li et al. (2016).

NLI models We systematically experiment with transformer architectures that are fine-tuned on MNLI, which exhibit state-of-the-art or close predictive accuracy on the dataset. We specifically choose *bart-large* (Lewis et al., 2019), *roberta-large* (Liu et al., 2019b), *deberta-base, deberta-large, deberta-xlarge* (He et al., 2020) and *distilbart* (Sanh et al., 2019). These PLMs are taken without change from their original studies through the Transformers library (Wolf et al., 2020), ensuring the complete reproducibility of the results. To observe the effect in an *out-of-domain* setup, we also evaluate these models on SNLI and ANLI in a zero-shot transfer setting.

5 Results and Analysis

This section presents the results and analyses of our semantic sensitivity evaluation framework along with a suite of ablations analysing the phenomenon across various transformer sizes, domains, and label space. Furthermore, we measure the impact of the phenomenon on the inconsistent predictive behaviour of NLI models.

5.1 Semantic Sensitivity

In-domain We evaluate several PLMs trained on MNLI using our experiments presented in Table 2. The results show that models are limited in their comprehension of compositional semantics as the relaxed fooling rate on *in-domain* experimentation averages at $r_r = 12.9\%$. This is further reinforced by the fact that more than half, $r_s = 6.58\%$ of the label changes occur with strict inequality. This means that minor semantics-preserving changes lead to a sizable shift in model predictions, even prompting towards the opposite decision edge half the time. The behaviour is consistent across all the transformers and leads us to believe that samples

that changed labels after surface-form variations showcase the inconsistent predictive nature of the models. We further elaborate on this in the next section. Consequently, semantically equivalent variations evidently hinder the decision-making of the NLI models, prompting us to believe that models have limited understanding w.r.t. semantic structure and logical relation, even when the model is trained on texts from the same distribution.

Out-of-domain We also probe the NLI models in an *out-of-domain* zero-shot setting to assess the transferability of compositional semantic knowledge. Our results in Table 2 show that the discrepancies and limitations in semantic comprehension are even more pronounced in this setting. We see an averaged relaxed fooling rate of $r_r = 23.7\%$, with the maximum at 57.49%, which is only marginally better than a majority voting baseline. It must be noted that because different datasets have varying numbers of samples, the average is weighted w.r.t. the number of sampled instances from the particular dataset in the experiment. The results on *out-of-domain* evaluation once again follow the pattern that more than half, $r_s = 15.8\%$ of the samples switch the labels to their logically contrasting counterparts. This shows that zero-shot transfer further amplifies the limitations that NLI models have for using semantic structures and preserving logical relations. This further suggests that the semantic variations where a label change occurs are likely to be originally predicted correctly as an inconsistent guess. It follows, that although PLMs fine-tuned on MNLI are widely used for zero-shot classification, their effectiveness diminishes if the classification tasks require syntactic understanding. Indeed, model effectiveness declines and the fooling rates rise as the tasks become more challenging, requiring greater syntactic knowledge, as we can see from the comparison of the results from SNLI to ANLI.

Effects of distillation Next, we want to probe if the susceptibility towards semantic noise is transferred during model distillation. Thus, we use *DistilBart* that is distilled from a larger pre-trained BART model. While model accuracy remains comparable to the original model in Table 1, the distilled version struggles sizeably more with surface-form variations. On average, across *in-* and *out-of-domain* evaluation, the distilled NLI model is more sensitive than the original in terms of relaxed fooling rate by $\triangle r_r = 18.4\%$. The effect of supposed inconsistence is amplified when observing the strict fooling rate, where on average $\frac{r_r}{r_s} \leq 1.5$. This indicates that during distillation, models are bound to forget the knowledge regarding compositional semantics making it harder to preserve the logical equivalence during inference.

Effects of model size We also test how semantics-preserving noise affects models of different sizes and parametrization (see Figure 2). Although for *in-domain* setup, the relaxed fooling rate metrics marginally drop as the models get bigger, the same cannot be observed in *out-of-domain* setup. It is evident that bigger PLMs from our study are almost as restricted in semantic comprehension as their smaller counterparts. This indicates that emergent semantic capabilities are not only tied to model size, but also widely depend upon the choice of the training dataset.

5.2 Severity of Inconsistent Predictions

Consistency across label space To analyse the extent of semantic sensitivities within NLI models we test the effect across all the classes in the label spaces, presented in Table 3. The per-class breakdown of the strict and relaxed fooling rate indicates that the effect is consistent across the whole label space. This allows us to conclude that the observed limitations in compositional semantic understanding are not caused by class imbalances and are not specific to a particular set of examples. We see the increased fooling rate across all of the labels when comparing *in-domain* and *out-of-domain* experiments. This reinforces the prior indications regarding models' inability to use semantic structure to preserve inherent relations within the data, as all logical relations attain rather similar amounts of fooling rate during direct evaluation.

Distribution shift in decision making Recall that we want to measure the impact of semantics-preserving surface-form variations on NLI models. We study the predictive distributional shift within the samples that cause a changed model prediction. To do this, we initially split the samples into two categories considering whether the sample induced a change of the original prediction within the NLI model. We further average the probability distribution of labels obtained from the final softmax layer of the model for these two categories. We measure the differences between the two distributions with two statistical tests. To evaluate

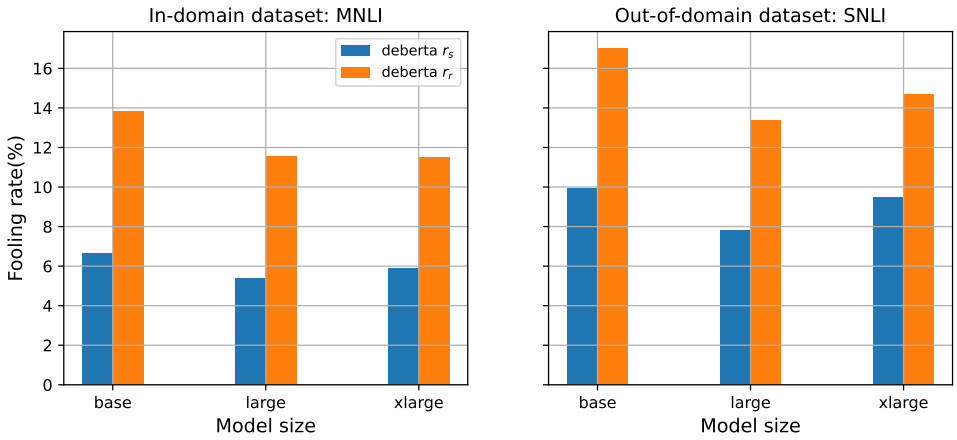

Figure 2: In- and out-of-domain fooling rate of DeBERTa of varied sizes, which are measured on MNLI (left) and SNLI (right). Similarly, r_s and r_r represent the strict and relaxed fooling rates, respectively.

	$r_s/r_r\,(y=E)$	$r_s/r_r\,(y=N)$	$r_s/r_r\,(y=C)$	r_s/r_r
MNLI	2.78%/13.41%	14.33%/14.33%	3.69%/11.17%	6.58%/12.92%
SNLI	9.54%/18.73%	19.42%/19.42%	2.92%/11.82%	10.24%/16.54%
ANLI_r1	21.64%/41.97%	38.62%/38.62%	29.17%/44.57%	29.97%/41.30%
ANLI_r2	20.84%/46.28%	49.41%/49.41%	21.89%/50.80%	31.32%/48.53%
ANLI_r3	11.65%/52.00%	47.18%/47.18%	16.42%/46.50%	27.04%/48.17%

Table 3: Fooling rate averaged over all models. r_s represents the strict fooling rate, in which case the predicted label of the evaluation pair is opposite to the original label y. r_r measures the proportion of label change. $y \in \{E, N, C\}$ group the (p, h) pairs by their semantic relation, representing entailment, neutrality, and contradiction, respectively.

the relative entropy between them, we use Jensen-Shanon Divergence (Fuglede and Topsoe, 2004), a symmetric, non-negative, and bounded metric for assessing the similarity between two distributions, $\text{JSD}(P\|Q) = \frac{1}{2}D(P\|M) + \frac{1}{2}D(Q\|M)$, where D is the Kullback–Leibler divergence (Joyce, 2011). We verify the statistical significance of our findings with the Kolmogorov–Smirnov test (Berger and Zhou, 2014), which shows if the two sets of samples are likely to come from the same distribution.

Our results in Figure 3 show a significant distribution shift when assessing semantics-preserving surface-form variations. The cosine distance in the sentence embedding space between the generated and original samples is negligible at 0.04. As the absolute cosine similarity values possess limited interpretable meaning, we further explore the distributions of cosine distances towards original samples for the examples that do and do not induce label changes. We measure the Jansen-Shannon divergence of these two distributions at 0.001, implying they are strongly similar. This reinforces the hypothesis that surface-form variations produce logically equivalent samples with minor distance in the embedding space regardless of the induced label changes. However, despite minor changes in the semantic composition, we see a sizable change in the final predictive distribution of the NLI models. We see a significant rise both in Jensen-Shannon divergence and Kalmogorov-Smirnov metric, $\triangle\text{JSD} = 0.51$ and $\triangle\text{K-S} = 0.54$, when comparing the examples where the model prediction has changed compared to the original. This indicates that the generated variations do not cause negligible change within model prediction, but rather can be considered adversarial for the model. It shows that the limited capabilities to utilise syntactic information cause the model to significantly change the final prediction given minuscule variations, which is an to inconsistent predictive behaviour. Given that we initially sampled examples that the models answered correctly, these results assert our belief that the models do not display consistent predictive behaviour despite having equivalent inputs. This shows that albeit the strong model performance presented in Table 1, there is

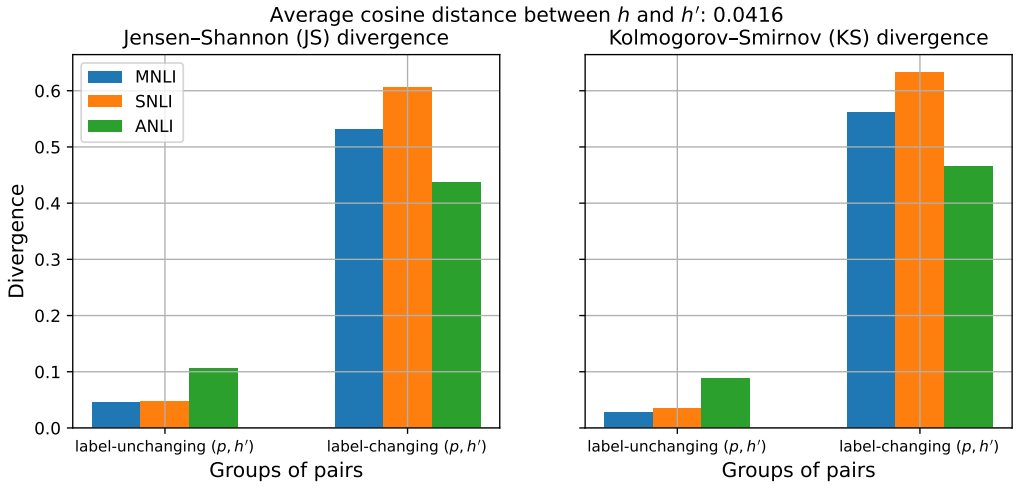

Figure 3: Divergence of predictive probability distribution between (p, h) and (p, h') measured across the datasets (ANLI is averaged over the rounds) and averaged over all models. All evaluation pairs are split into two groups based on whether they manage to flip the original label. Two divergence metrics are shown – JS divergence (left) and KS divergence (right).

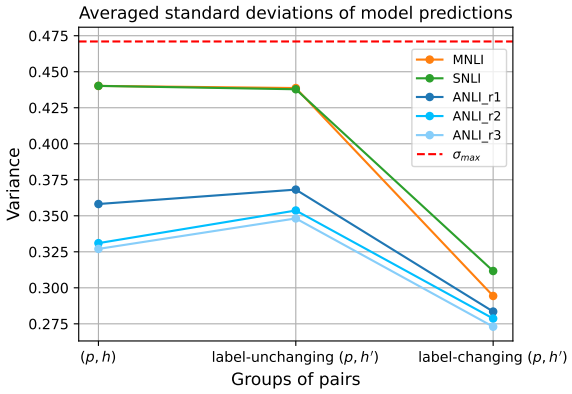

Figure 4: Standard deviation σ of predicted label probabilities (obtained from the final softmax layer of the model) averaged for original premise-hypothesis pair (left), surface-form variations that did not cause label changes (mid) and did induce label change (right). The bigger σ, the more confident the model is w.r.t. the predictions. The results are averaged over all models.

masked degeneration and discrepancies within the NLI models stemming from semantic sensitivity. Our method allows for explicitly quantifying the degree of semantic sensitivity within PLMs and allows to measure the impact of that sensitivity on the decision-making process of the model.

Semantic-Sensitivity and decision variations We lastly analyse the standard deviation within the predicted label distribution produced from the softmax of the model. We compute the standard deviation for the distribution of original premise hypothesis predictions and compare it with a replacement that does not and does cause label changes in PLM classification, see Figure 4. For reference, the upper bound for standard deviation in this 3 class setting happens when the model is greatly confident in one of the classes, i.e. softamx $= [1, 0, 0] \rightarrow \sigma_{\text{max}} = 0.471$. Bigger σ on average implies more confident answers by the PLM. It can be observed that the average predictions with the original samples have a great degree of confidence. We see an interesting phenomenon where the predictive confidence slightly rises across most of the datasets for the cases where the model is able to recover the inherent textual relations. However, when faced with examples that cause label changes, there is a significant drop of $\triangle\sigma = 0.1$ in the standard deviation averaged across the datasets. This signifies that predictive confidence sizably degrades when the model struggles to recover the existent relations because of slight semantics-preserving variations. That further indicates that NLI models are susceptible to semantic sensitivity and have limited knowledge of compositional semantics, which can lead to the degradation of predictive confidence and incidentally inconsistent predictions.

6 Conclusion

We present a novel framework for assessing semantic sensitivity in NLI models through generating semantics-preserving variations. Our systematic study of the phenomenon across various datasets and transformer-based PLMs shows that the models consistently struggle with variations requiring

knowledge of compositional semantics. This performance deterioration happens across the whole label space, almost regardless of model size. We measure the impact of semantic-sensitivity and show that it diminishes models' predictive confidence and can lead to predictive inconsistency.

Limitations

In our work, we cover the semantic-sensitivity that can be found within NLI models. However, the framework can be applied to a wider range of classification tasks. The benchmark can be extended with more datasets and further enhanced with larger human evaluation. Also, we covered PLMs specifically trained for NLI; however, it would be great to cover bigger LLMs, in particular w.r.t. their emergent zero-shot capabilities. Another limitation is that we only cover English-based language models and do not test in multi-lingual or cross-lingual settings.

Ethics Statement

Our work completes an analysis of numerous models w.r.t. their decision inconsistency induced by semantic surface form variations. We show that models are somewhat unable to handle logically and semantically equivalent sentences, which would lead to an inconsistent use across various domains and applications. Our generation method does not induce any further exploitation threat and can only be used for measuring the above-mentioned inconsistencies. We exclusively use open source publicly accessible data and models within our experimentations.

Acknowledgements

Erik is partially funded by a DFF Sapere Aude research leader grant under grant agreement No 0171-00034B, as well as by a NEC PhD fellowship. This work is further supported by the Pioneer Centre for AI, DNRF grant number P1.

References

James Allen. 1995. *Natural language understanding*. Benjamin-Cummings Publishing Co., Inc.

Aixiu An, Peng Qian, Ethan Wilcox, and Roger Levy. 2019. Representation of constituents in neural language models: Coordination phrase as a case study. 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 2888–2899, Hong Kong, China. Association for Computational Linguistics.

Richard B Angell. 1989. Deducibility, entailment and analytic containment. *Directions in relevant logic*, pages 119–143.

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Vance W Berger and YanYan Zhou. 2014. Kolmogorov–smirnov test: Overview. *Wiley statsref: Statistics reference online*.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Rudolf Carnap. 1959. *Introduction to semantics and formalization of logic*. Harvard University Press.

Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. 2018. Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Aaron Cicourel. 1991. Semantics, pragmatics, and situated meaning. *Pragmatics at Issue*, 1:37–66.

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.

Michael Clark. 1967. The general notion of entailment. *The Philosophical Quarterly (1950-)*, 17(68):231–245.

Cleo Condoravdi, Dick Crouch, Valeria De Paiva, Reinhard Stolle, and Daniel Bobrow. 2003. Entailment, intensionality and text understanding. In *Proceedings*

of the HLT-NAACL 2003 workshop on Text meaning, pages 38–45.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, pages 177–190. Springer.

Ishita Dasgupta, Demi Guo, Andreas Stuhlmüller, Samuel J. Gershman, and Noah D. Goodman. 2018. Evaluating compositionality in sentence embeddings.

Ishita Dasgupta, Andrew K Lampinen, Stephanie CY Chan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. 2022. Language models show human-like content effects on reasoning. *arXiv preprint arXiv:2207.07051*.

Bent Fuglede and Flemming Topsoe. 2004. Jensen-shannon divergence and hilbert space embedding. In *International symposium onInformation theory, 2004. ISIT 2004. Proceedings.*, page 31. IEEE.

Atticus Geiger, Kyle Richardson, and Christopher Potts. 2020. Neural natural language inference models partially embed theories of lexical entailment and negation. *arXiv preprint arXiv:2004.14623*.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word 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 4129–4138.

V Ivan Sanchez Carmona, Jeff Mitchell, and Sebastian Riedel. 2018. Behavior analysis of nli models: Uncovering the influence of three factors on robustness. In *NAACL HLT 2018-2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies-Proceedings of the Conference*, volume 2018, pages 1975–1985. Association for Computational Linguistics (ACL).

Pauline I Jacobson. 2014. *Compositional semantics: An introduction to the syntax/semantics interface*. Oxford Textbooks in Linguistic.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020a. Are natural language inference models imppressive? learning implicature and presupposition. *arXiv preprint arXiv:2004.03066*.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020b. 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.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

James M Joyce. 2011. Kullback-leibler divergence. In *International encyclopedia of statistical science*, pages 720–722. Springer.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of bert. *arXiv preprint arXiv:1908.08593*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. A simple, fast diverse decoding algorithm for neural generation. *arXiv preprint arXiv:1611.08562*.

Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew E Peters, and Noah A Smith. 2019a. Linguistic knowledge and transferability of contextual representations. *arXiv preprint arXiv:1903.08855*.

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. *arXiv preprint arXiv:1907.11692*.

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*, Florence, Italy. Association for Computational Linguistics.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. 2017. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018a. Stress test evaluation for natural language inference. *arXiv preprint arXiv:1806.00692*.

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.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*.

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.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. *arXiv preprint arXiv:2005.04118*.

Kyle Richardson, Hai Hu, Lawrence Moss, and Ashish Sabharwal. 2020. Probing natural language inference models through semantic fragments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8713–8721.

Joost Rommers, Ton Dijkstra, and Marcel Bastiaansen. 2013. Context-dependent semantic processing in the human brain: Evidence from idiom comprehension. *Journal of Cognitive Neuroscience*, 25(5):762–776.

Corby Rosset, Chenyan Xiong, Minh Phan, Xia Song, Paul Bennett, and Saurabh Tiwary. 2020. Knowledge-aware language model pretraining. *arXiv preprint arXiv:2007.00655*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Stephen Schiffer. 1986. Compositional semantics and language understanding. *Philosophical Grounds of Rationality: Intentions, Categories, Ends, Oxford*, pages 174–207.

Koustuv Sinha, Prasanna Parthasarathi, Joelle Pineau, and Adina Williams. 2020. Unnatural language inference. *arXiv preprint arXiv:2101.00010*.

Koustuv Sinha, Prasanna Parthasarathi, Joelle Pineau, and Adina Williams. 2021. Unnatural language inference.

Robert Stalnaker, Milton K Munitz, and Peter Unger. 1977. Pragmatic presuppositions. In *Proceedings of the Texas conference on per˜formatives, presuppositions, and implicatures. Arlington, VA: Center for Applied Linguistics*, pages 135–148. ERIC.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. Do nlp models know numbers? probing numeracy in embeddings. *arXiv preprint arXiv:1909.07940*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Shuohang Wang and Jing Jiang. 2015. Learning natural language inference with lstm. *arXiv preprint arXiv:1512.08849*.

Alex Warstadt and Samuel R Bowman. 2020. Can neural networks acquire a structural bias from raw linguistic data? *arXiv preprint arXiv:2007.06761*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al.

2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019. Can neural networks understand monotonicity reasoning? *arXiv preprint arXiv:1906.06448*.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *arXiv preprint arXiv:1909.00161*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. *arXiv preprint arXiv:1808.05326*.

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2020. Semantics-aware bert for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9628–9635.

A Appendix

Dataset	Fuzzy token match %	average length h	average length $h\prime$	average token overlap
mnli	84.83	14.31	14.14	13.25
snli	81.55	10.81	11.21	10.38
anli_r1	87.59	17.3	17.02	13.73
anli_r2	86.49	15.99	15.84	12.8
anli_r3	85.17	14.32	14.29	11.27

Table 4

Evaluation under Label change To assess the extent of the impact of semantic sensitivity, we employ an evaluation under label change. This means we consider the examples that changed the original prediction of the model after a surface-form variation replaced the original hypothesis. A graphical representation of this can be seen in Figure 5. It must be noted that we use only the samples that the model originally predicted correctly to avoid incorrect assessment regarding the reasoning behind the false predictions. Our primary aim is to measure the semantic sensitivity within the model predictions and the extent of inconsistency it causes.

Token Level-Differences of the generated variations We further explore the difference between surface-form variations and original examples by conducting a token-level analysis for each pair $(h, h\prime)$. We compute the average amount of tokens present for the original and generated hypothesis and use fuzzy and exact matching to assess the overlap of tokens on average for each dataset. The results can be seen in Table 4. The results show that the generated and original examples have a high token level overlap which further reinforces the idea that surface form variations are close both syntactically, in the embedding space and logically.

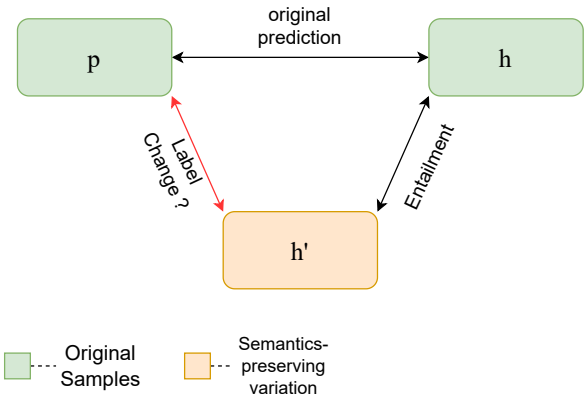

Figure 5: A diagram for assessing semantic similarity. Given the generated semantics-preserving surface-form variation h', we evaluate if a label change occurs when replacing the hypothesis in accordance with Equation 1 .