MERGE: Minimal Expression-Replacement GEneralization Test for Natural Language Inference

Anonymous ACL submission

Abstract

In recent years, many generalization benchmarks have shown language models' lack of robustness in natural language inference (NLI). However, manually creating new benchmarks is costly, while automatically generating high-quality ones, even by modifying existing benchmarks, is extremely difficult. In this paper, we propose a methodology for automatically generating high-quality variants of original NLI problems by replacing open-class words, while crucially preserving their underlying reasoning. We dub our generalization test as MERGE (Minimal Expression-Replacements GEneralization), which evaluates the correctness of models' predictions across reasoning-preserving variants of the original problem. Our results show that NLI models' perform 4–20% worse on variants, suggesting low generalizability even on such minimally altered problems. We also analyse how word class of the replacements, word probability, and plausibility influence NLI models' performance.

1 Introduction

Generalizability is models' ability to reliably adapt their knowledge to new scenarios (Dutt et al., 2024; Hupkes et al., 2023; Yang et al., 2023), being especially important for their real-life deployment where data is unpredictable (Hupkes et al., 2023; Budnikov et al., 2025). Considering this ability, evaluating models only on a held-out test set – so-called *in-distribution* evaluation – is incomplete for models evaluation, as they might learn shallow heuristics to exploit systematic gaps of the test set (Gardner et al., 2020; Hupkes et al., 2023; Dutt et al., 2024). Thus, *out-of-distribution* (OOD) benchmarks have been developed by introducing a different **text genre** or by slightly changing training data items (Hupkes et al., 2023). One such task where this was applied is Natural Language Inference (NLI), where models need to classify the inference relationship between a premise and a hypothesis. Some NLI benchmarks are SNLI (Bowman et al., 2015) or MNLI (Williams et al., 2018), which contain mainly in-distribution data (Petrov, 2025). As a solution, adversarial benchmarks (ANLI, Nie et al., 2019), or minimally modified existing benchmarks (Verma et al., 2023; Srikanth et al., 2024, among others) have been developed, both showing models' poor generalization skills. Nevertheless, such benchmarks are costly and difficult to scale if created **manually** (Li et al., 2020), while their quality and (dis)similarity with the training data is unreliable if datasets are automatically generated.

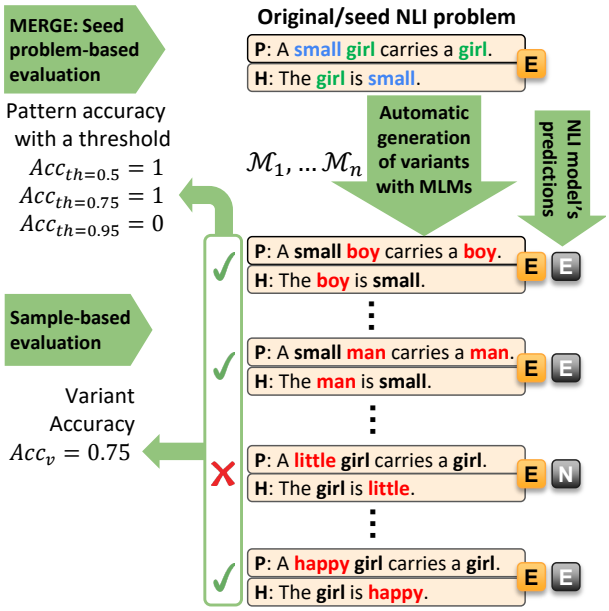

Figure 1: MERGE vs. standard sample-based evaluations: while in the former each variants is an independent example, in MERGE performance is measured as the proportion of correctly classified variants for an NLI problem. Thus, whether a model classifies at least x amount of variants (threshold number) for each NLI problem.

We propose the Minimal Expression-Replacement GEneralization (MERGE) test for NLI, a methodology by which we can reliably

create variants of NLI problems, only minimally differing from the originals. As shown in Figure 1, we achieve this by automatically replacing open-class words shared between the premise (P) and hypothesis (H) – usually irrelevant for the problems' underlying reasoning especially in NLI datasets – with *contextually probable* replacements suggested by masked language models (MLMs). Thus, variants preserve (i) the underlying reasoning of the original problem; and ii) model exploited-heuristics such as sentence lengths and word overlap size (Naik et al., 2018; McCoy et al., 2019; Bernardy and Chatzikyriakidis, 2019; Rajaee et al., 2022); Furthermore, we filter out replacements not adhering to certain quality criteria to obtain plausible variants, without requiring human validation at a large scale.

Importantly, as shown in Figure 1, MERGE assesses NLI models' generalizability by their performance on amounts of variants rather than on individual ones (Abzianidze et al., 2023; Srikanth et al., 2024). We aim to answer the following questions:

1. Are NLI models robust against minimal variants of NLI problems?
2. How reliable is our methodology for **automatically** obtaining variants of NLI problems?
3. Do factors such as the likelihood, word class, semantic plausibility, or the MLM of replacement words influence performance?

Our contributions are the following:

1. We present a model- and dataset-agnostic methodology to test generalizability using the simplest, model-friendliest minimal changes to create reasoning-preserving NLI variants.
2. We evaluate multiple models on MERGE variant datasets, their results suggesting poor generalization and oversensitivity to label-preserving minimal changes.
3. We show that quality criteria for replacements are important: those generated by multiple models and equally likely for the premise and hypothesis are higher quality than those based on plausibility, POS tag, probability, or the specific MLM used.

The other sections of the article review previous works in §2, and present our methodology in §3. §4 describes our experiments, while results are in §5. We discuss our conclusions in §6.

2 Related Work

Generalizability in NLI OOD benchmarks obtained by minimal alterations (or *contrasts sets* in Petrov, 2025; Li et al., 2020, here referred to as *variant datasets*) suggest models severely lack NLI generalization abilities, given their decreased performance (Li et al., 2020; Gardner et al., 2020; Kaushik et al., 2020, among others). We review previous NLI variant datasets studies, considering aspects such as their *modification strategy*, and the *(non)automatic* approach for creating variants.

Modification Strategy To construct variant datasets, previous studies have used various modification strategies, such as **replacing** words, **decomposing**, or **paraphrasing** problems, or a combination of **multiple** operations, described in the Column *Strategy* in Table 1.

Creation Type Operations were achieved **automatically**, **manually**, or by using a **mix** of automatic and manual methods, shown in Column *Creation* in Table 1.

Validation The variants have been validated by i) human validation, **partial** (HVal_p), **full** (HVal_f), or **combined** (HVal_{pf}) where only a randomly chosen subset, the full set, or a combination thereof are manually validated; iii) by a **mix** of automatic and manual validation methods (Mix.); iv) by no validation (*N/A*), described under Column *Validation*.

Meaning, Reasoning, Word Overlap and Syntax The meaning, logical reasoning, syntax of the original problem or word overlap of the premise and hypothesis (PH)[1] can be **preserved** (Y), **changed** (N), or a **mix** of both. The Columns *M*, *R*, *S* and *WO* in Table 1 classify all studies considering these aspects.

Modified Sentence The modifications can be applied only to the *premise* (P), the *update*[2] (U), the *hypothesis* (H), to both of them, but separately (P/H), and to both of them together ($P\&H$), shown under Column *Sentence Mod.* in Table 1.

Evaluation Models' predictions on variants can be evaluated by comparing them (i) individually, with the **gold label** (*V-G*) or the prediction of the

[1] Note that word overlap preservation requires modifying both $P\&H$, without doing so for their syntax.

[2] Some NLI datasets include defeasible inferences where new information, i.e., updates, might change the entailment label between the premise and hypothesis.

Study	Strategy	Creation	Val.	Sentence Mod.	M	R	S	WO	Evaluation	Dataset
Li et al. (2020)	Multiple	Auto.	HVal_p	P	Mix.	Mix.	N	N	Vs-G; Vs-O	SNLI; MNLI
Glockner et al. (2018)	Replace	Auto.	HVal_f	H	N	Mix.	Y	N	V-G	SNLI
Verma et al. (2023)	Paraphrase	Auto.	HVal_f	P/H; $P\&H$	Y	Y	N	N	Vs-O	Pascal RTE1-3 (Dagan et al., 2005)
Srikanth et al. (2024)	Paraphrase	Mix.	HVal_{pf}	H; U	Y	Y	N	N	Vs-Vs; Vs-G	α-NLI (Bhagavatula et al., 2019); δ-NLI (Rudinger et al., 2020)
Arakelyan et al. (2024)	Paraphrase	Auto.	HVal_p	H	Y	Y	N	N	V-O	SNLI; MNLI; ANLI
Petrov (2025)*	Multiple	Auto.	N/A	H	N	N	N	N	V-G	SNLI
Kaushik et al. (2020)	Multiple	Man.	HVal_f	P; H	N	N	N	N	V-G	SNLI
Srikanth and Rudinger (2025)	Decompose	Auto.	Mix.	H	Y	Y	N	N	V-G; Vs-O	SNLI; δ-NLI
MERGE	Replace	Auto.	LMVal	$P\&H$	N	Y	Y	Y	V-G; Vs-G	SNLI

Table 1: Overview of minimal generalizability tests. **Strategy**: *Replacing* words, *Paraphrasing*, *Decomposing* sentences, or a combination of *Multiple* operations; **Creation**: *Auto.* – automatic; *Man.* – manual; *Mix.* – *mixing* methods; **Val.**: $HVal_p$ – a subset, $HVal_f$ – the full set, or a combination $HVal_{pf}$ of variants is human-validated; *Mix* – automatic and human validation; *LMVal* – LM validated. **Sentence Mod**: only the premise (P), hypothesis (H), or update (U); or both P and H at the same time ($P\&H$) or separately (P/H) are modified; **M/R/S/WO**: meaning, reasoning, syntax, or word overlap between $P\&H$ are (*Y*), or are not (*N*) preserved, or a mix of these two (*Mix.*); **Eval.**: *individual comparison* – the gold label (*V-G*) or prediction of the original problem (*V-O*) vs. the variant; *group comparison* – the gold label (*Vs-G*), prediction of original sentence (*Vs-O*), or other variants (*Vs-Vs*) vs. variants. **Dataset**: modified dataset(s); * in **Study** column: variant datasets used for fine-tuning models. All studies use labels: *entailment*, *contradiction*, *neutral*.

original NLI problem (*V-O*); (ii) as a group, with the prediction on the original NLI problem (*Vs-O*), the gold label (*Vs-G*), or with each other (*Vs-Vs*). See column *Eval.* in Table 1 for their classification.

Results of previous studies Overall, previous studies showed models have decreased performance on variant datasets by 14 to 30% (Kaushik et al., 2020; Petrov, 2025; Glockner et al., 2018). Furthermore, the models were inconsistent, as they had different predictions for variants in 10-16%, and for variants in comparison to prediction for the original problem (Verma et al., 2023; Arakelyan et al., 2024).

Shortcomings of previous variant datasets Overall, previous lower scores of models on variant datasets cannot be attributed fully to poor generalizability, as they could also be due to: i) new syntactic constructions – syntactic **non-preserving** changes (Li et al., 2020) might be more challenging for models; ii) changed lexical overlap of $P\&H$ – caused by *paraphrased* or *decomposed* NLI problems. Preserving the word overlap of $P\&H$ by using *replacement* words has been partially explored previously (Glockner et al., 2018), however, with a trade-off between creating plausible rather than lexical diverse variants, given the predefined set of replacement words used. Other shortcomings of previous studies concern their creation type, with automatic methods having been criticized due to (i) their potential to bias variants in favor of models deployed (Li et al., 2020; Gardner et al., 2020), with some even arguing benchmarks should be created from scratch manually, thereby avoiding the re-use of previous ones Gardner et al. (2020) and (ii) their potential to construct implausible variants (Dutt et al., 2024). However, manual creation types, besides being time-consuming, by default cannot be faster than the rate at which new models might learn them.

MERGE automatically creates **plausible** variants by replacing shared words with felicitous alternatives. Thus, the lexical overlap of $P\&H$, their syntax, and underlying logical reasoning are preserved, while avoiding the implausibility of constructed variants, unlike in previous studies (Arakelyan et al., 2024; Srikanth et al., 2024; Verma et al., 2023). Lexical diversity is not fixed but can also be increased, for instance, by considering more suggestions from more MLMs, unlike in previous list-restricted replacement studies (Abzianidze et al., 2023; Glockner et al., 2018).

3 Methodology

For a seed NLI problem $\langle P, H, l\rangle$ and a label l, open-class words w shared between P and H, $w \in P \cap H$, are replaced with new words to obtain variant NLI problems $\langle P_i, H_i, l\rangle$.

The new words are collected from a set of MLMs $\mathcal{M} = \{M_1, \ldots, M_n\}$, as follows: for a sentence $S = (s_1, \ldots, s_k)$, an MLM M_j, and a word position $1 \leq i \leq k$, where s_i is an open-class word, $W_j(S, i^c_>)$ is the set of words that (i) are more

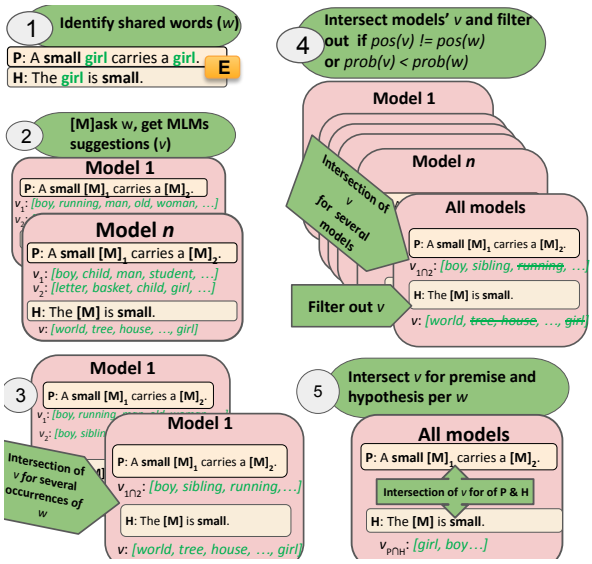

Figure 2: Generating NLI problem variants with MLMs. The suggestions of a shared word between $P\&H$ are excluded if they have different *classes* (*teen* for *little*) than w_i, or if they are already part of the problem (*girl*).

probable than s_i, under M_j, in the masked context $S[s_i := \text{MASK}]$, marked with $_>$, (ii) do not occur in S, and (iii) belong to the same word class as s_i in S, marked with ^{c}.[3]

Constraints (i) and (iii) make sure that suggested words are more felicitous than the original w_i, while preserving syntactic structure and the lexical overlap between $P\&H$. For example, one possible replacement for 'girl' from Figure 2 is 'running', which results in an implausible variant with different syntactic structure that could affect its original label. On the other hand, low probability suggestions corresponding to the original word class might have a similar effect on fluency. Lastly, (ii) decreases the chance of changing the inference label as, for example, 'two dogs and a boy swim' entails 'only one boy swims', but substituting the shared word *boy* with *dog*, already occurring in the premise, breaks the entailment relation.[4]

Note that for a word w with multiple occurrences in S, its suggestion set from M_j is defined as:

$$W_j(S, w^c_>) = \bigcap_{\substack{1 \le i \le k \\ s_i = w}} W_j(S, i^c_>)$$

[3]Nouns, verbs, adjectives, or adverbs. Constraints (i) and (ii) exclude the original s_i from suggestions, while the set of suggestions for s_i will be empty provided it is not part of the M_j vocabulary.

[4]Despite the constraints applied, NLI variants with incorrect inference labels are still possible, although improbable. For example, replacing 'boy' with 'animal' results in an incorrect entailment label.

where the suggested words should comply with the aformentioned constraints at each occurrence position of w in S. When considering a set of MLMs $\mathcal{M}$, their suggestions set for a word w in S are defined as:

$$W_{\mathcal{M}}(S, w^c_>) = \bigcup_{M_j \in \mathcal{M}} W_j(S, w^c_>)$$

where suggested words should be validated by the same MLM at each occurrence position of w in S. Finally, for an NLI problem $\langle P, H, l \rangle$ and $w \in P \cap H$, we define a set of suggestions from a set of MLMs $\mathcal{M}$ as:

$$W_{\mathcal{M}}(\langle P, H \rangle, w^c_>) = \bigcap_{S \in \{P,H\}} W_{\mathcal{M}}(S, w^c_>) \quad (1)$$

and the variant NLI problems $\langle P_{ij}, H_{ij}, l \rangle$ are obtained by replacing the original shared words $w_i \in P \cap H$ with corresponding suggested words $v_{ij} \in W_{\mathcal{M}}(\langle P, H \rangle, w^c_{i>})$:

$$P_{ij} = P[w_i / v_{ij}], H_{ij} = H[w_i / v_{ij}]$$

We will use $W^{d=m}$ to denote a random subset of size m of the suggestion set W. We call d a degree of inflation. If there are k words shared between P and H, and for each word we have a set of suggestions with the inflation degree of d, then the total number of generated variants will be $k \times d$.

4 Experimental Setup

We will now describe the various experimental choices made in building our variant dataset.

Suggestion generation For all w_i being shared nouns, verbs and adjectives between $P\&H$ in SNLI test problems, illustrated in Table 4[5], we generated 200 suggestions v_i with the following MLMs: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2019), Electra (Clark et al., 2020), and BART (Lewis et al., 2019). Their sizes, fully described in Table 5, were base and large, except for ALBERT where we used base and xxl.

Suggestion filtering Suggestions were kept if they were: i) same *class* as w_i; ii) higher probability than w_i; iii) different from words already in $P\&H$, including w_i; iv) validated by at least one model; vi) word tokens (i.e. excluding punctuation marks or subwords). Across open-classes, v_i had more often the same class as w_i (40-80% of v_i),

[5]We limit open-class words to nouns, verbs, and adjectives, given the limited number of NLI problems sharing adverbs between P and H.

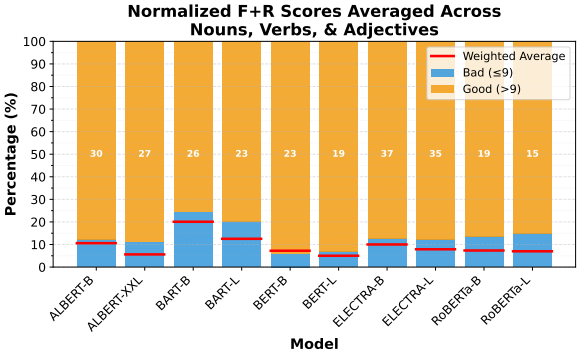

Figure 3: Averaged **F**luency and **R**easoning scores with normalized counts for 100 random variants for Nouns, Verbs, and Adjectives. The red lines are the bar plots weighted considering the distribution of classes in the seed NLI problems (N=67%, V=23%, ADJ=10%). Good variants have a score of $F + R >= 9$.

Class	Seed	Average	N(%)	C(%)	E(%)	Subs
N_{Var}	1808	90.8	29.4	20.2	50.3	36160
V_{Var}	506	86.8	30.6	18.6	50.8	10120
A_{Var}	259	77.6	32.8	22.4	44.8	5180
ALL_{Var}	**2222**	**77.4**	**30.0**	**20.4**	**49.5**	51,460

Table 2: Total number of variants after filtering across classes: **Seed**– number of seed problems; **Average**d – number of variants per seed problem across classes, not computed as **Subs/Seed**, since variant counts vary by the latter; **N/C/E (%)** – percentages per label across seed problems: *neutral*, *contradiction*, and *entailment*; **SubS**– variants subsampled 10 times.

than a bigger probability (10% of v_i), as shown Table 6 of the Appendix.

We also excluded SNLI test problems with fewer than 20 v_i ($d = 20$) corresponding to all their summed w_1 after our filtering.[6].

Variants Manual Annotation We formed NLI problems' variants by replacing w_1 with v_i. We randomly sub-sampled 100 variants per class for manual annotation, which two authors independently rated on a scale from 1–5 (1 – good; 5 – poor) considering their: a) plausibility of the replacement; and b) label preservation of the original NLI problem, check subsection A.2 for the full annotation guidelines. Variants with $a + b \geq 9$ were classified as good, and generally few variants received scores lower than 3 on b) (i.e. 5%), due to ungrammaticalities and one change of label.[7]. As shown in Figure 3, we plotted the average proportions of good and bad suggestions per model across classes. Given **BART** models had higher scores for bad variants – especially when weighting by class distribution per seed problems—we excluded them, which led to a minor reduction of 318 seed problems across classes. To asses the necessity of different model sizes considering their potential overlapping suggestions, we plotted a confusion matrix in Figure 9 with model-specific suggestions on the diagonal. Overall, models of different sizes generated more unique than overlapping suggestions, with some (e.g., BERT, Electra) showing nearly equal proportions of both, supporting the inclusion of all models' sizes. To evaluate variants post-exclusion, we sampled again 100 random variants from all classes together, out of which 91% had a good score, plotted in Figure 10.

Final Variant Dataset From all variants, we randomly selected 20 per seed problem to prevent imbalance across seeds. This subsampling was repeated 10 times to avoid randomization artifacts, resulting in the dataset All_{Var}. Table 2 shows the distribution of variants across open classes, along with seed counts and label distributions. Note that some problems share a w_i across classes, so the seed total is non-summative; for example, in Figure 2, where $P\&H$ share the noun 'girl' and adjective 'small'.

Evaluation Metrics For evaluation, we used standard accuracy – hereby referred to as Sample Accuracy (SA); and Pattern Accuracy (PA) (Abzianidze et al., 2023) – a problem $\langle P, H \rangle$ is considered correctly predicted only if at least an x threshold amount of variants $\langle P_{ik}, H_{ik} \rangle$ are as well, capturing a model's degree of generalizability of an NLI problem. See Figure 1 for a visualization of the differences between the two metrics. We refer to the threshold x also as the Accuracy Threshold, defined as the averaged scores obtained across the 10 subsamples of a dataset. Considering the 91% qualitative variants, we expect models to get at least 90%[8] of variants correctly (i.e. the quality threshold).

NLI Models We evaluated several models fine-tuned for NLI: BERT, RoBERTa, DeBERTa (He

[6]We chose $d = 20$ as it is easily interpretable (e.g., 95% is 19 out of 20, while out of 10 it would be 9.5).

[7]The replacement 'short' in the hypothesis 'The short player is tall' was ranked R1 given it became a contradiction from its original neutral label. Despite some grammatical errors in the original SNLI, we flagged any that could affect reasoning as poor.

[8]While 91% of variants are qualitative, we chose 90% as it is more intuitive to interpret (e.g. equals 18 out of 20 variants).

et al., 2021), BART, ALBERT, Electra, XLNet (Yang et al., 2019), OPT (Zhang et al., 2022) and GPT-2 (Radford et al., 2019), with the sizes and training datasets shown in Table 3, and model cards shown in Table 9 of the Appendix. Different model sizes were chosen to account for potential effects of models' sizes on their scores.

5 Results

5.1 Are models robust against variants?

Models' close SA scores on ALL_{Seed}, $SNLI_{test}$ and $SNLI_{Var}$ in Table 3 indicate that our subset of seed problems and their variants are not much more difficult than the overall $SNLI_{test}$ dataset. If we consider the quality threshold of 90%, scores under Column **90** already decrease somewhere between 4–9%, continuing to decrease even further after, as shown in Figure 4. Most importantly, column **MT** in Table 3 shows that obtaining PA scores on All_{Var} comparable to the SA scores on ALL_{Seed} would required setting the accuracy threshold at around 60%.

Models' scores decrease considerably after the 80% threshold, showing less generalization the more variants are considered, as shown in Figure 11. Initially, BERT-L-S performs worst and OPT-1-3b-S among the best, but

beyond the 86% threshold, OPT-1-B drops to the lowest rank, while BERT scores higher. In line with previous studies, both RoBERTa variants do

Model	$SNLI_{Test}$	ALL_{Seed}	ALL_{Var}	90	MT
BERT-B-S	90.5	89.6	88.9	-4.9	59
BERT-L-S	87.1	87.2	87.4	-4.5	47
RoBERTa-B-S	90.1	90.1	89.2	-4.5	47
DeBERTa-v3-B-S	91.7	92.1	90.7	-4.9	58
DeBERTa-v3-L-S	91.7	91.9	91.0	-4.9	54
BART-B-S	90.6	90.2	89.4	-4.3	57
OPT-1-3b-S	91.0	90.5	89.1	-8.6	58
GPT-2-L-S	90.9	90.9	89.5	-7.7	55
RoBERTa-L-SMFA	91.8	91.4	90.5	-5.0	59
BART-L-SMFA	92.0	91.9	90.5	-6.0	55
Electra-L-SMFA	91.1	90.6	90.0	-6.5	56
XLNet-L-SMFA	91.7	91.4	90.6	-5.4	55
ALBERT-XXL-SMFA	91.9	92.2	91.2	-4.8	57

Table 3: SA scores of models on *$SNLI_{test}$*, seed problems (ALL_{Seed}) and their variants (ALL_{Var}). For model names, the first acronym marks their size, while the second the NLI datasets used for fine-tuning, i.e. *S* – only SNLI; *SMFA* – SNLI, MNLI, FEVER (Thorne et al., 2018), and ANLI. **90**: PA difference at threshold 90 (ALL_{Seed}–ALL_{Var}); **MT**: nearest **m**atching **t**hreshold where PA on $All_{Var} \approx$ SA on All_{Seed}.

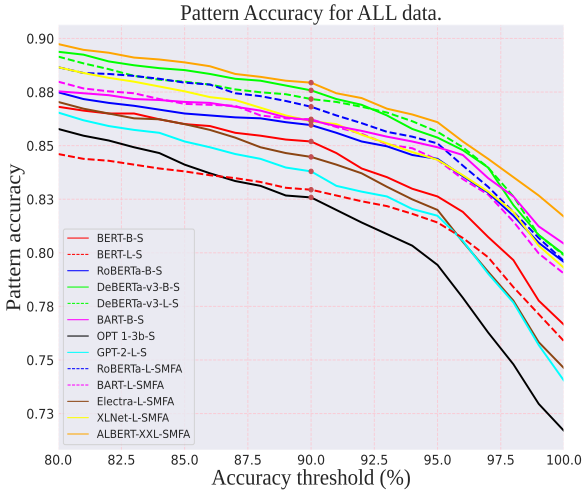

Figure 4: PA scores of models on ALL_{Var} from 80% threshold on. The red dots on are the scores models get at 90% – the quality threshold of ALL_{Var}.

better than BERT; but not better than DeBERTa (Li et al., 2020; Srikanth and Rudinger, 2025). However, RoBERTa-L-SMFA, has the highest thresholds in Column **MT**, suggesting its degradation in generalizability is slower. In terms of best-performing models, DeBERTa-v3-L-S and ALBERT-XXL-SMFA have best PA scores, with the latter scoring best on ALL_{Seed} as well. Models trained on more data tend to do better, as generally they rank among the highest-performing models with the exception of Electra-L-SMFA.

Overall, such results indicate models are sensitive to word-level changes that are irrelevant for reasoning, in line with previous studies showing their lack of robustness against variant datasets (Verma et al., 2023; Arakelyan et al., 2024; Glockner et al., 2018, , etc). Lower PA scores might be caused by variants disrupting spurious correlations via replacements of w_i, such as the associations between certain words and labels exploited for better classification $P\&H$(Poliak et al., 2018; Gururangan et al., 2018).

5.2 Which word classes are more difficult?

We test models on the ALL_{Var} and its classes from Table 2: N_{Var}, V_{Var}, and A_{Var}, to observe their effect on models' scores. Figure 5 shows averaged model scores across open-class categories. Until the 80% threshold, adjectives are the most difficult in the green line, after which nouns become hardest as shown by the blue line. Verbs remain the easiest throughout, as indicated by the higher scores of the red line.

To better isolate the effect of classes from their

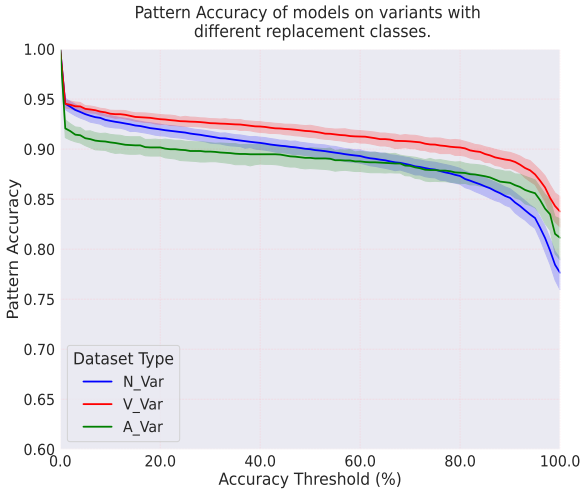

Figure 5: Averaged PA scores on N_{Var}, V_{Var}, and A_{Var}.

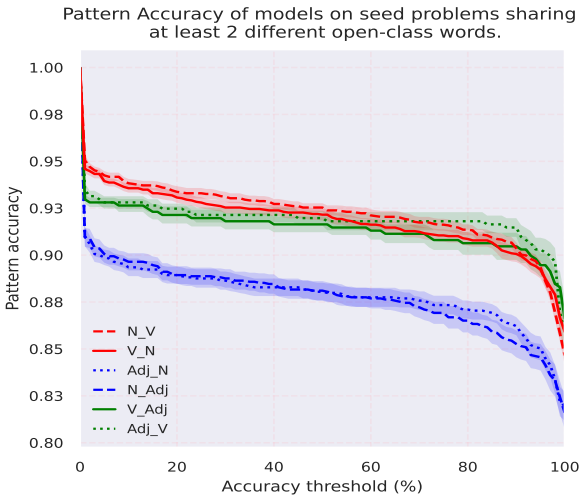

Figure 6: Averaged PA curves of all models on seed problems sharing at least two out of the three open-class words.

imbalanced number of seed problems shown in Table 2, we only select variants for those seed problems that share at least two out of the three open-class words[9]. For example, the problem corresponding to 'A small girl carries a girl' is included in this subset as it shares an adjective and a noun across $P\&H$. The averaged scores across models are plotted in Figure 6. For dataset names, their first letter indicates which variants out of the two classes are we considering, e.g. V_N – variants formed by only replacing verbs in seeds sharing at least a verb and a noun. Figure 6 shows that nouns are initially easier than verbs (red lines), similarly to adjectives in the green lines. On higher accuracy thresholds, nouns and adjectives become more difficult than verbs, with the nouns being most difficult. The results indicate, in the end, a similar trend to Figure 4.

5.3 Do MLMs favor same NLI models?

To test whether MLM-based variant creation favors the performance of its corresponding NLI models, as proposed by (Li et al., 2020; Gardner et al., 2020), we averaged the scores of the latter on their MLM equivalent: BERT-B-S, BERT-L-S, RoBERTa-B-S, RoBERTa-L-SMFA, Electra-L -SMFA, and Albert-XXL-SMFA – and plotted them in Figure 7. The scores came from variants validated: 1) only by the equivalent MLM –Only-MLM; 2) by its equivalent MLM of different size, e.g. variants from BERT-L for BERT-B-S – Cross-Size; 3) by any two MLMs – Multi-MLM; and 4) by another MLM – Cross-MLM. MLMs of different sizes may be classified as either Only-MLM or Cross-Size, depending on the evaluated NLI model (e.g., BERT-B variants are Only-MLM for BERT-B-S and Cross-Size for BERT-L-S). We plotted the averaged number of entries per subsample in Table 8, where the Multi-MLM is shown to be disproportionately larger than others. Until around the 70% accuracy threshold, models did better on Multi-MLM and Cross-MLM data, while around the 80% accuracy threshold they started to perform worse. Comparatively, they constantly performed well on Only-MLM and Cross-Size variants. However, their performance on Multi-MLM variants, to which they have contributed in a lower proportion than in Only-MLM, indicates that the MLM used for generation might be less important than the size of the variant dataset, as models' scores are lowest on variants they have also contributed to.

5.4 Which filtering criteria matter?

To test the effect of different filtering criteria for v_{ij} on NLI models' scores, we subsampled new variants for the seed problems of ALL_{Var} forming datasets: $_{P \cup H}$ – where v_{ij} for $P\&H$ are the result of union, instead of the intersection in Equation 1; $_{POS}$ – only with v_{ij} having w^c; $_{Prob}$ – only with v_{ij} having $w_{i>}$; $_{None}$ – any class or probability for v_{ij}; $_{Scr}$ – v_{ij} have their letters randomly scrambled. Except for $_{Scr}$, the datasets are more diverse due to less strict filtering, while still excluding punctuation signs and v_{ij} already in P or H. We plot the averaged PA scores of all models in Figure 8, highest performance being achieved on Scr, followed by ALLVar, Prob, Pos, None, and $P \cup H$.

[9] Sharing N–V=202; N–A = 130, and V–A =46.

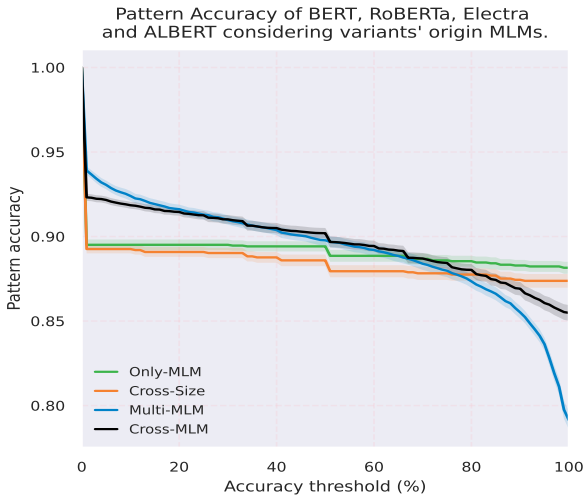

Figure 7: Averaged PA curves for BERT, RoBERTa, Electra and ALBERT, all sizes tested, on variants divided by their origin MLM.

The effect of union in $_{P \cup H}$ is the strongest out of all filtering criteria, as scores drop to 80% PA near the 60% threshold and decrease further with higher thresholds, the large error bars corresponding to some models having very low scores, i.e. OPT-1.3B-S and GPT-2-L-S reaching about 60% on high thresholds.

Comparatively, controlling for factors such as probability in $_{\text{Prob}}$, or plausibility in $_{\text{Scr}}$ might not be as important, as models' score better on scrambled variants, and close to ALL_{Var} in $_{\text{Prob}}$ ones. Replacements with all classes of words lowers scores of models more, but not as much as $P \cup H$.

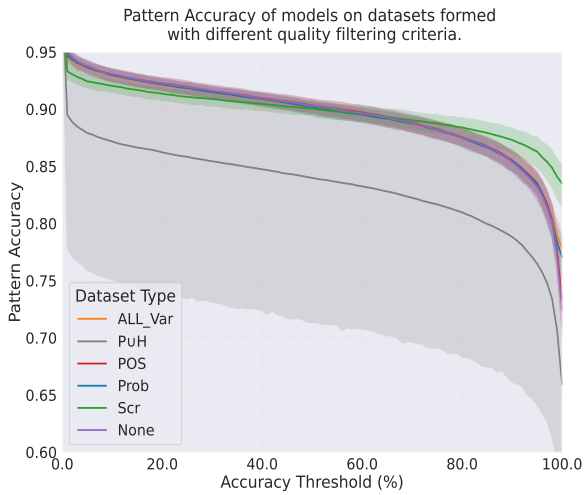

Figure 8: Averaged PA curves of all models on datasets formed with v_{ij} having any class or probability – $_{\text{None}}$, being scrambled – $_{\text{Scr}}$, with higher probability than the original word – $_{\text{Prob}}$, with the same class – $_{\text{Pos}}$, suggested in either P and H – $_{P \cup H}$, and ALL_{Var}.

5.5 What the hard NLI problems look like

We analyzed the NLI seed/original problems (2222) based on how successfully on average the NLI models classify their variants. We found that 0.77% (17) of the seed problems are hardest: no NLI model could correctly classify their single variant. After careful inspection we found out that only 3 out of 17 seed problems have a correct gold inference label (see the problems in Table 10). We further analyzed the seed problems (in total 55) whose variants were classified with $< 5\%$ average accuracy by the models. From them only 9 had correct gold label. The models mistakes can be associated with commonsense knowledge (4) and assuming too much (3). The analysis of the hardest problems, where models' generalization is lowest, showed that only around 15% of them have correct gold labels. This adds evidence to the well-known issue of annotation variation in the context of NLI (Pavlick and Kwiatkowski, 2019; Weber-Genzel et al., 2024).

6 Conclusions

We presented MERGE, a methodology for minimal generalization testing that automatically constructs variants of NLI problems, while preserving reasoning and word overlap. Its strength lies in minimal, controlled replacements: through quality filters for fluency and consistency, MERGE tests models under the simplest generalization conditions.

We show models do not generalize well, as they drop with 4–6% one the quality level of our variant dataset to be 90%, and by up to 20% in high accuracy thresholds. Even more, achieving a pattern accuracy comparable to that obtained on the seed problems entails lowering the accuracy threshold by 60% for most models, indicating their generalizability extends only this far. Our results also show variants formed from replacing nouns, are more difficult than those from adjectives or verbs, while NLI models do better when evaluated on variants from the same type of model, effect overturned by a bigger size of the dataset. When it comes to important filtering criteria for suggestions, them being equally plausible in both the premise and the hypothesis seems to be the most important one.

In the future, we plan to expand MERGE to other NLI benchmarks and NLU tasks (e.g. reading comprehension), and other reasoning tasks.

Limitations

One limitation of our methodology is the lack of account for original words that are split into subwords by tokenizers, which we are excluding by default. Additionally, our suggestions are obtained by masking one occurrence of a word at a time. However, when one occurrence is masked, the other ones are still part of the sentence, which might bias the suggestions of models. Future studies could also consider evaluating prompt-based models, to observe if different prompt strategies might help the models perform better on variants.

Acknowledgments

Acknowledgments

References

Lasha Abzianidze, Joost Zwarts, and Yoad Winter. 2023. SpaceNLI: Evaluating the consistency of predicting inferences in space. In *Proceedings of the 4th Natural Logic Meets Machine Learning Workshop*, pages 12–24, Nancy, France. Association for Computational Linguistics.

Erik Arakelyan, Zhaoqi Liu, and Isabelle Augenstein. 2024. Semantic sensitivities and inconsistent predictions: Measuring the fragility of NLI models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 432–444, St. Julian's, Malta. Association for Computational Linguistics.

Jean-Philippe Bernardy and Stergios Chatzikyriakidis. 2019. What kind of natural language inference are nlp systems learning: Is this enough? In *ICAART (2)*, pages 919–931.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. 2019. Abductive commonsense reasoning. *arXiv preprint arXiv:1908.05739*.

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*.

Mikhail Budnikov, Anna Bykova, and Ivan P Yamshchikov. 2025. Generalization potential of large language models. *Neural Computing and Applications*, 37(4):1973–1997.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *Preprint*, arXiv:2003.10555.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Ritam Dutt, Sagnik Ray Choudhury, Varun Venkat Rao, Carolyn Rose, and V.G.Vinod Vydiswaran. 2024. Investigating the generalizability of pretrained language models across multiple dimensions: A case study of NLI and MRC. In *Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP*, pages 165–182, Miami, Florida, USA. Association for Computational Linguistics.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hanna Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. *Preprint*, arXiv:2004.02709.

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.

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. 2021. Deberta: Decoding-enhanced bert with disentangled attention. *Preprint*, arXiv:2006.03654.

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, et al. 2023. A taxonomy and review of generalization research in nlp. *Nature Machine Intelligence*, 5(10):1161–1174.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. *Preprint*, arXiv:1909.12434.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942.

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. *Preprint*, arXiv:1910.13461.

Chuanrong Li, Lin Shengshuo, Zeyu Liu, Xinyi Wu, Xuhui Zhou, and Shane Steinert-Threlkeld. 2020. Linguistically-informed transformations (LIT): A method for automatically generating contrast sets. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 126–135, Online. 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. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

R. Thomas 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.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. 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*.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Daniel Petrov. 2025. From superficial patterns to semantic understanding: Fine-tuning language models on contrast sets. *arXiv preprint arXiv:2501.02683*.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. *Preprint*, arXiv:1805.01042.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Sara Rajaee, Yadollah Yaghoobzadeh, and Mohammad Taher Pilehvar. 2022. Looking at the overlooked: An analysis on the word-overlap bias in natural language inference. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10605–10616, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rachel Rudinger, Vered Shwartz, Jena D Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A Smith, and Yejin Choi. 2020. Thinking like a skeptic: Defeasible inference in natural language. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4661–4675.

Neha Srikanth, Marine Carpuat, and Rachel Rudinger. 2024. How often are errors in natural language reasoning due to paraphrastic variability? *Transactions of the Association for Computational Linguistics*, 12:1143–1162.

Neha Srikanth and Rachel Rudinger. 2025. Nli under the microscope: What atomic hypothesis decomposition reveals. *Preprint*, arXiv:2502.08080.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. 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 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Dhruv Verma, Yash Kumar Lal, Shreyashee Sinha, Benjamin Van Durme, and Adam Poliak. 2023. Evaluating paraphrastic robustness in textual entailment models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 880–892, Toronto, Canada. Association for Computational Linguistics.

Leon Weber-Genzel, Siyao Peng, Marie-Catherine De Marneffe, and Barbara Plank. 2024. VariErr NLI: Separating annotation error from human label variation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2256–2269, Bangkok, Thailand. 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.

Linyi Yang, Yaoxian Song, Xuan Ren, Chenyang Lyu, Yidong Wang, Jingming Zhuo, Lingqiao Liu, Jindong Wang, Jennifer Foster, and Yue Zhang. 2023. Out-of-distribution generalization in natural language processing: Past, present, and future. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4533–4559, Singapore. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models. *Preprint*, arXiv:2205.01068.

A Appendix

A.1 Variant Creation

Word Class	T	N(%)	C(%)	E(%)
Nouns	7363	33.4	28.7	37.8
Verbs	3780	31.8	28.2	40.0
Adjectives	1067	33.6	23.7	42.6
Adverbs	76	26.3	13.2	60.5

Table 4: Number of seed problems and label distributions for open-class word categories from the SNLI test set. **T**: seed problems shared between P and H; **N/C/E**: percentage of *neutral*, *contradiction*, and *entailment* labels.

MLMs For the MLMs, we used the models enumerated in Table 5, from which BART was excluded given its suggestions were of lower quality, as shown in Figure 3 for the annotation described in section 4. In addition to the models described in the table, we also tried including DeBERTa, which was too noisy to use for suggestion generation.

Class Filtering Suggestions of different classes than the original word w_i were excluded by tagging each suggestion v_i in the context of its corresponding sentence using the spaCy model `en_core_web_sm`. We calculated the average of excluded suggestions after tagging per model, which is shown in the table in 6.

Probability Filtering We obtained the probability of each v_{ij} of w_i in P or H from the MLM that suggested it, and we excluded those of lower probability. MERGE validates variants by assuming that replacements as likely as the w_i are less likely to be semantically implausible. Original words w_i that were not part of the model's vocabulary were also automatically excluded. Also note that both probability and class filtering are important, as highly probable suggestions might not have the correct class and vice versa. Table 6 shows the percentages of excluded suggestions for having lower probability than the replaced word under Column **Prob. Fil**.

Model	Size	Architecture	Vocabulary
BERT (Devlin et al., 2018) google-bert/bert-base-cased	B	E	28,996
BERT (Devlin et al., 2018) google-bert/bert-large-cased	L	E	28,996
RoBERTa (Liu et al., 2019) FacebookAI/roberta-base	B	E	50,265
RoBERTa (Liu et al., 2019) FacebookAI/roberta-large	L	E	50,265
BART (Lewis et al., 2019) facebook/bart-base	B	E-D	50,265
BART (Lewis et al., 2019) facebook/bart-large	L	E-D	50,265
ALBERT (Lan et al., 2019) albert/albert-base-v2	B	E	30,000
ALBERT (Lan et al., 2019) albert/albert-xxlarge-v2	XXL	E	30,000
ELECTRA (Clark et al., 2020) google/electra-base-generator	B	E	30,522
ELECTRA (Clark et al., 2020) google/electra-large-generator	L	E	30,522

Table 5: Overview of selected pretrained MLMs used for suggestion generation. Column **Size**: Size **B**ase, **L**arge, or **XXL**arge; Column **Architecture**: the model is **E**ncoder, **D**ecorder or **E**ncoder-**D**ecoder. The vocabulary size of the models is shown in the last column.

A.2 Annotation

We annotated variants in terms of how much they change the fluency and the original logical label of the NLI problems, which we considered to be correct. The scores used were: 1 – **poor**, 2– **mostly poor**, 3– **uncertain**, 4–**mostly good**, 5–**good**. Table 7 shows how we annotated certain variants, and their attributed fluency and reasoning scores, alongside an explanation for them. Variants were classified as **poor** if they were: 1) ungrammatical; 2) had missing arguments; 3) nonsense; or 4) logic non-preserving. Aspects 1) and 2) were chosen given they directly hinder the evaluation of fluency and reasoning.

The instructions in the annotation guidelines are provided below with the demo examples in Table 7:

> The NLI problems are assessed on the fluency (grammaticality and sensibility) and reasoning. The reasoning com-

Model	POS	POS Fil.		Prob. Fil.	
		P	H	P	H
ALBERT-B	A	45.7	48.4	87.6	80.4
	N	7.2	7.8	88.4	79.9
	V	26.5	29.3	91.1	87.1
ALBERT-XXL	A	45.3	48.4	87.8	78.1
	N	6.1	6.7	89.6	75.9
	V	24.5	24.5	92.1	84.5
BART-B	A	63.9	65.8	90.2	84.2
	N	28.3	26.6	91.3	84.8
	V	38.7	37.2	90.0	88.5
BART-L	A	64.0	65.9	89.4	83.4
	N	30.6	30.8	91.5	83.3
	V	37.2	38.1	88.7	87.6
BERT-B	A	45.1	48.7	93.5	89.6
	N	10.6	11.6	92.2	85.8
	V	26.4	28.6	92.5	90.4
BERT-L	A	44.4	47.4	94.5	91.0
	N	10.4	12.3	94.0	87.9
	V	28.9	29.7	94.5	92.6
Electra-B	A	45.4	48.4	89.6	84.9
	N	11.0	11.3	88.2	82.4
	V	27.6	27.5	88.3	88.9
Electra-L	A	45.6	49.0	91.1	86.2
	N	10.7	10.7	90.9	85.0
	V	25.5	26.5	90.5	89.5
RoBERTa-B	A	49.2	52.3	93.9	89.1
	N	7.7	8.2	94.2	88.3
	V	31.1	33.8	95.0	91.8
RoBERTa-L	A	47.7	50.9	95.2	91.3
	N	7.7	8.1	95.7	90.2
	V	33.2	32.3	96.6	93.6

Table 6: Average percentages of suggestions with other POS Categories (Pos Fil.) and lower probabilities (Prob Fil.). The percentages under *Premise* and *Hypothesis* show how many suggestions, on average, out of 200 had a different class or lower probability than the original masked word.

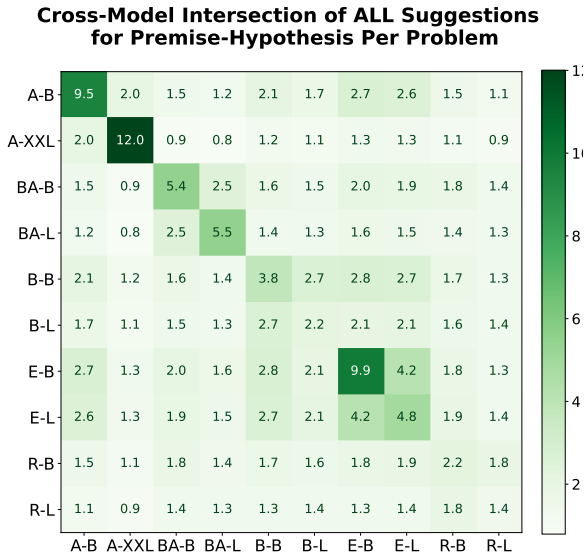

Figure 9: Confusion matrix showing how models' suggestions intersect on average, across problems, with the diagonal being the averaged number of original or model-specific suggestions a model generates. For instance, in the problem 'A small girl carries a girl', a suggestion like 'bear' counts on the diagonal if unique to one model, or in the other cells if shared between models.

ponent focuses on the relation between the meaning of P and H rather than their fluency. However, poor fluency can negatively affect the reasoning part. For example, "Colorless green ideas sleep furiously" entailing "Green thoughts is angrily sleeping" should be assessed with fluency 1 and reasoning 5, in short F1-R5.

The original SNLI problems could suffer from fluency and reasoning; however, the obtained NLI problem variants should be assessed with respect to the original NLI problems. Annotation should assess the fluency and reasoning of the variants while assuming that the fluency and reasoning of the original NLI problems are fine. That's why variant NLI problems are provided with the origin word that helps to reconstruct the original SNLI problem.

Below are several examples to demonstrate different combinations of fluency and reasoning. The scale 1-5 should be interpreted as: poor, mostly poor, uncertain, mostly good, good.

Evaluated NLI Models The NLI models we evaluated alongside with their model cards are in Table 9.

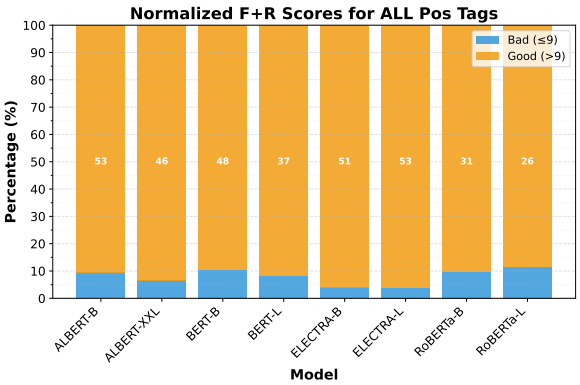

Figure 10: **F**luency and **R**easoning scores for 100 randomly sampled variants and their normalized counts, from all pos tags, and the models that validated them, after exclusion of suggestions from BART. Note that 91 out of the 100 examples evaluated had $F + R => 9$.

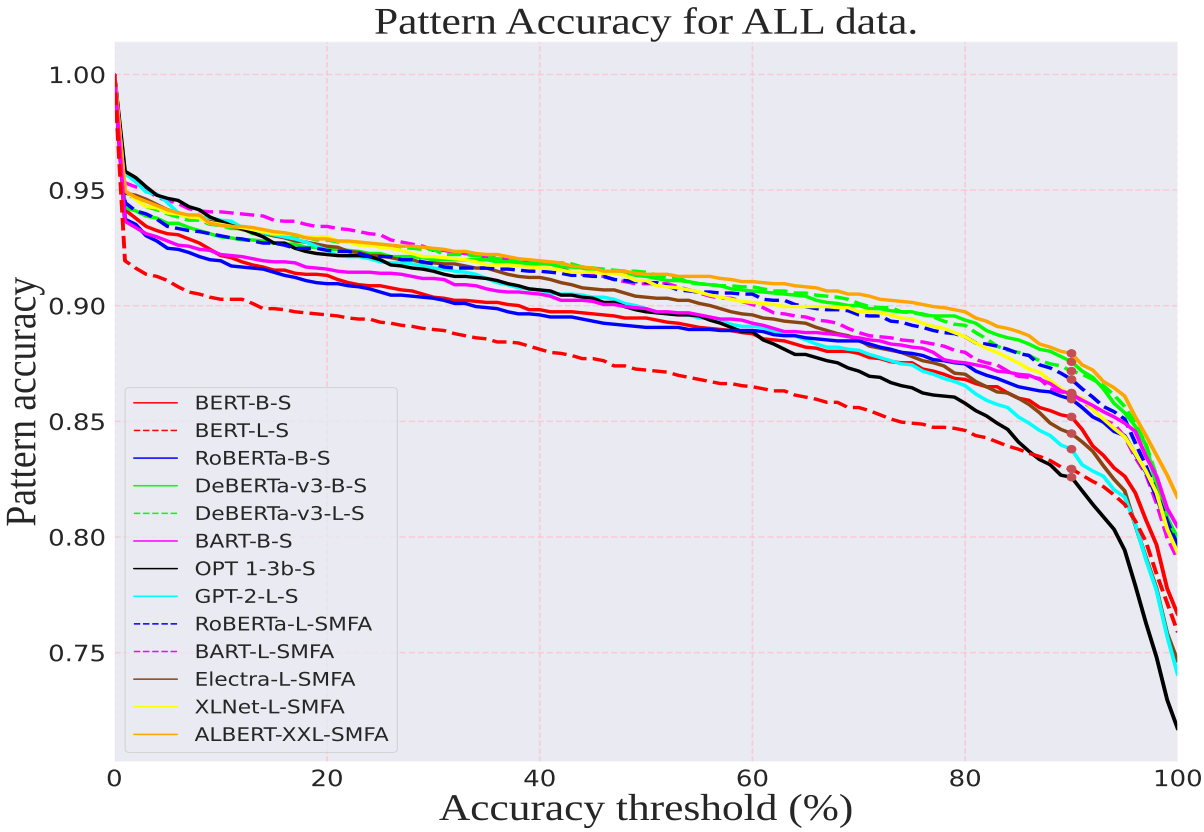

Figure 11: PA scores from threshold 0% for all models on ALL_{Var}. The red dots on each line indicate the scores of models at 90% – the quality threshold we set for ALL_{Var}.

Premise	Hypothesis	Label	Original	Suggested	F	R	Explanation
A man in a black shirt, in a detached kitchen, holding up meat he took out of a bag.	A woman in a black shirt, in a detached kitchen, holding up meat he took out of a bag.	C	commercial	detached	5	5	**Good**: the variant is fluent and preserves reasoning.
A teenage dog is running in a field near a mountain.	A teenage dog is running outdoors.	E	yellow	teenage	5	5	**Good**: the variant is fluent and preserves reasoning.
A man driving while at a restaurant.	A man driving while at a restaurant eating.	N	laughing	driving	4	5	**Mostly good**: the variant might be likely in a certain scenario (e.g. there might be a restaurant where tables are like cars).
A man is celebrating his victory while smiling and wearing champagne in the air with his teammate.	A man is happily celebrating his victory while smiling and wearing champagne in the air	N	shooting	wearing	3	5	**Uncertain**: the variant might be likely in a certain scenario (e.g. the man is wearing a champagne-shaped costume), but a less likely one.
A man slung into the ear wearing a striped shirt in a small boat filled with many people.	A man is slung into the ear and wearing a light striped shirt.	N	pointing	slung	2	5	**Mostly poor**: the scenario is very unlikely, e.g. a man being forced to hear something.
clutched to her ear, a woman bends forward at the side of a busy street.	clutched to her ear, a woman bends forward	E	Phone	clutched	1	5	**Poor**: the variant is ungrammatical, thus making it difficult to verify its fluency, i.e. missing the theme of "clutched".
A hole is on a cherry picker in a palm tree.	A hole falls out of a tree.	C	worker	hole	1	5	**Poor**: the variant is not fluent, until we consider a very specific metaphorical context, i.e. a hole cannot be have agency.
A shirt booth with a man got a shirt.	The man is got some pants.	C	printing	got	1	4	**Mostly Good**: the variant is still fluent, despite it being metaphorical, and it is still likely to preserve the initial inference label, despite the ungrammaticality of the hypothesis.
This child is took a pedicure.	Child took a manicure.	C	getting	took	1	3	**Uncertain**: the ungrammticality of the premise makes it difficult to asses its fluency, however the main source of the contrast giving the contradiction (i.e. *pedicure* vs. *manicure*) is kept.
A brown dog wearing a collar is chasing and running on a red broom.	There is an animal running a broom.	E	biting	running	4	2	**Very Poor**: Even though "running a broom" might be running *with* a broom, running *on* something does not entail running *with* X.
Two construction workers produced the steel ribbed exterior of a new building at their work site.	Two workers are produced a building	N	climbing	produced	1	1	**Poor**: the hypothesis is hard to understand given its ungrammaticality, while the subject of 'produced' is unclear, which makes reasoning impossible to asses.

Table 7: Examples of annotation scores for NLI variants, considering their fluency (**F**) and the preservation of the original NLI label (**R**), alongside explanations for their scores. Note that the labels under Column *Label* are **N**eutral, **E**ntaliment, and **C**ontradiction.

Model	Avg	Avg Other
BERT-B	473.7	9572.0
ALBERT-B	2085.8	4073.9
Electra-L	700.1	7445.4
BERT-L	228.0	9572.0
ALBERT-XXL	4114.0	4073.9
Electra-B	2128.2	7445.4
RoBERTa-B	355.9	9729.8
RoBERTa-L	188.0	9729.8
Both	41186.3	10273.7

Table 8: Average entries per subsample when testing the effect of MLM origin of variants and scores of NLI models. The **Av**era**g**e of **oth**er numbers are the averaged entries per minidatset of all other MLMs models, excluding different sizes of the model. (excluding model family and *both*).

Model	Model Card
ALBERT-XXL-SMFA	ynie/albert-xxlarge-v2-snli_mnli_fever_anli_R1_R2_R3-nli
BART-B-S	varun-v-rao/bart-base-snli-model1
BART-L-SMFA	ynie/bart-large-snli_mnli_fever_anli_R1_R2_R3-nli
BERT-B-S	textattack/bert-base-uncased-snli
BERT-L-S	varun-v-rao/bert-large-cased-lora-1.58M-snli
DeBERTa-v3-B-S	pepa/deberta-v3-base-snli
DeBERTa-v3-L-S	pepa/deberta-v3-large-snli
Electra-L-SMFA	ynie/electra-large-discriminator-snli_mnli_fever_anli_R1_R2_R3-nli
GPT-2-L-S	varun-v-rao/gpt2-large-snli-model3
OPT-1-3b-S	utahnlp/snli_facebook_opt-1.3b_seed-3
RoBERTa-B-S	pepa/roberta-base-snli
RoBERTa-L-SMFA	ynie/roberta-large-snli_mnli_fever_anli_R1_R2_R3-nli
XLNet-L-SMFA	ynie/xlnet-large-cased-snli_mnli_fever_anli_R1_R2_R3-nli

Table 9: NLI evaluated models and their Hugging Face model cards.

P: A man in a colorful shirt and a lady in a white blouse sign copies of books for people. **H:** Two people sign copies of their latest novel. **Gold label:** E | **Correct label:** N | **Annotations:** E(3) N(2) C(0) | **Avg rate:** 0
P: A black dog is swimming with a ball in his mouth. **H:** A black dog found a ball in the water and is bring it back to its owner. **Gold label:** E | **Correct label:** N | **Annotations:** E(3) N(2) C(0) | **Avg rate:** 0
P: A young boy runs across a road in front of a sky blue building with barred windows. **H:** A boy runs across a road in front of an abandoned building. **Gold label:** E | **Correct label:** N | **Annotations:** E(3) N(2) C(0) | **Avg rate:** 0
P: The man in the brown shirt is holding the hand of the long-haired child in front of a painting. **H:** A male has clothes on with his hand holding another young male in front of a painting. **Gold label:** N | **Correct label:** N | **Annotations:** E(2) N(3) C(0) | **Avg rate:** 0
P: A mom and her boy are riding in a bumper car. **H:** The mom and boy are at an amusement park. **Gold label:** E | **Correct label:** N | **Annotations:** E(3) N(2) C(0) | **Avg rate:** 0
P: A man wearing a blue apron and long rubber boots is dragging a flotation device from a long row of flotation devices. **H:** A man in rubber boots and a work apron is rubbing his face in between pulling floating objects. **Gold label:** E | **Correct label:** N | **Annotations:** E(3) N(1) C(1) | **Avg rate:** 0
P: Two women stand in the street both wearing matching outfits with police like jackets that say Polotie. **H:** Two women are dancing in the street wearing dresses. **Gold label:** N | **Correct label:** C | **Annotations:** E(0) N(3) C(2) | **Avg rate:** 0
P: One girl sips a soda while another looks on, standing on a street in front of a bunch of bicycles. **H:** A girl drinks a soda on the street in front of people **Gold label:** N | **Correct label:** N | **Annotations:** E(2) N(3) C(0) | **Avg rate:** 0
P: A woman holds a newspaper that says "Real change". **H:** a woman on a street holding a newspaper that says "real change" **Gold label:** E | **Correct label:** N | **Annotations:** E(3) N(2) C(0) | **Avg rate:** 0
P: A happy woman quite stands in front of a business that displays a closed sign and looks very animated. **H:** A woman stands by a business. **Gold label:** N | **Correct label:** E | **Annotations:** E(2) N(3) C(0) | **Avg rate:** 0
P: A small child dressed for winter stares across a lake while leaning on a mesh fence in a scenic park setting. **H:** A small child dressed for winter stares across a large lake while leaning on a mesh fence **Gold label:** E | **Correct label:** N | **Annotations:** E(4) N(1) C(0) | **Avg rate:** 0
P: Two women competing in a roller derby with several teammates and referee in the background. **H:** Two roller derby competitors skate quickly ahead of teammates. **Gold label:** E | **Correct label:** E | **Annotations:** E(3) N(2) C(0) | **Avg rate:** 0
P: A young man blew up balloons to craft into animals for the seven excited children that looked on. **H:** The children watch the man make dogs and giraffes out of balloons **Gold label:** E | **Correct label:** N | **Annotations:** E(3) N(2) C(0) | **Avg rate:** 0
P: A train conductor in coveralls is standing in the door of the trail. **H:** The conductor is walking in a field. **Gold label:** N | **Correct label:** C | **Annotations:** E(0) N(3) C(2) | **Avg rate:** 0
P: An older gentleman looks at the camera while he is building a deck. **H:** An older gentleman in overalls looks at the camera while he is building a stained red deck in front of a house. **Gold label:** E | **Correct label:** N | **Annotations:** E(3) N(2) C(0) | **Avg rate:** 0
P: A group of people stand near and on a large black square on the ground with some yellow writing on it. **H:** a group of people stand **Gold label:** N | **Correct label:** E | **Annotations:** E(2) N(3) C(0) | **Avg rate:** 0
P: Children, including one with a painted face, pet tiny turtles that are crawling in the green grass. **H:** Turtles are crawling in the white grass. **Gold label:** E | **Correct label:** C | **Annotations:** E(3) N(0) C(2) | **Avg rate:** 0

Table 10: The original SNLI Problems whose variants were most poorly classified by the NLI models. **Avg rate** represents the average accuracy of NLI models across all variants per seed/original problem. **Annotations** reports the annotation labels of 5 annotators from the SNLI data, where the gold label is selected based on the majority voting. **Corrected label** is an inference label the authors think is correct. We are aware of the inherent disfigurements in NLI labeling (Pavlick and Kwiatkowski, 2019), especially in SNLI, but we dub our corrected labels as the most likely label.