

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

Mădălina Zgreabă ¹ Tejaswini Deoskar ¹ Lasha Abzianidze ¹

Institute for Language Sciences, Utrecht University, The Netherlands

{b.m.zgreaban, t.deoskar, l.abzianidze}@uu.nl

Abstract

As many benchmarks have become saturated, it is increasingly important to create new datasets that evaluate the generalization capacity of current state-of-the-art models in reasoning. However, creating high-quality reasoning datasets is challenging: manual construction is costly, and automatic generation is error-prone, with the community therefore relying on synthetic datasets with limited scope. In this paper, we propose the Minimal Expression Replacement GEneralization (MERGE) test, to evaluate the robustness and generalization of reasoning models. We use minimal and non-adversarial variants of existing evaluation datasets. First, we obtain high-quality variants automatically from the original instances using Masked Language Models (MLMs) for generation together with safeguarding filters, which we call the Minimal Expression Replacement (MERE) method. We then apply the MERGE test to Natural Language Inference (NLI), a popular reasoning task, by using MERE on two popular existing NLI datasets. We evaluate multiple strong NLI models and LLMs, and the results indicate they generalize poorly: both struggle to consistently and correctly classify variants minimally different in form, but similar in reasoning, from the original ones. We also analyze how aspects of variant generation, such as word class and source MLMs, affect model performance.

1 Introduction

The challenge in the Natural Language Inference (NLI) task is to predict the inference relation between a premise p and hypothesis h . Current models achieve very high performance on NLI, but this performance has been shown to often depend on partial exploitation of heuristics learned from the training set (Gardner et al., 2020; Hupkes et al., 2023; Dutt et al., 2024). Thus, when tested on *out-of-distribution* (OOD) datasets (Hupkes et al., 2023; Budnikov et al., 2025; Dutt et al., 2024; Yang

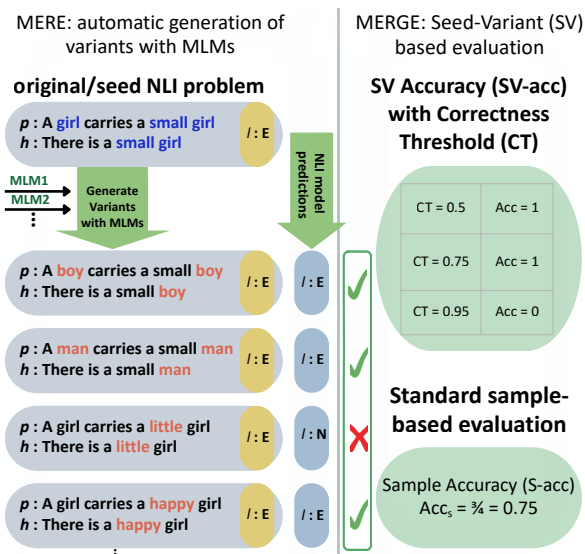


Figure 1: First, the MERE generation creates multiple label-preserving variants from a seed problem. Then MERGE evaluates a model on the seed based on its performance on the variants. Unlike the standard sample-based accuracy metric, MERGE uses a Seed-Variant (SV) accuracy with a correctness threshold, counting the seed as correctly classified if its variants are likewise classified correctly to at least the threshold degree.

et al., 2023), where test items differ from training ones in aspects that are not crucial to solving the task itself, such as text genres, models generalize poorly (Nie et al., 2020; Verma et al., 2023), despite good performance on traditional in-distribution test-sets of datasets like SNLI (Bowman et al., 2015) or MNLI (Williams et al., 2018).

OOD NLI datasets are usually challenging when they don't retain common heuristics, such as word overlap (Rajaei et al., 2022) or hypothesis-only artifacts (Gururangan et al., 2018; Poliak et al., 2018; Tsuchiya, 2018). Additionally, performance also drops on adversarial datasets, designed to go against the training data distribution by resembling certain training samples, but with different labels (Glockner et al., 2018; Naik et al., 2018; Gard-

ner et al., 2020, among others). This raises the question of whether the decreased performance of models is due to a failure to generalize, or because OOD datasets are too disruptive, as argued by Bowman and Dahl (2021). We therefore ask the question: can we create a minimal, non-adversarial generalization-challenge dataset that is “friendly” and deliberately retains such heuristics? Such a challenge set would test the ability of NLI models and LLMs to generalize on the task itself, without conflating it with additional challenges.

In this paper, we propose a method of automatically generating such evaluation datasets from existing ones by creating *minimal variants* of original ‘seed’ problems (Minimal Expression Replacement, MERE). Figure 1 provides an overview of our methodology: the left side shows how MERE operationalizes minimality for an NLI problem. To form variants of the seed problem $\langle p, h \rangle$, MLMs are used to suggest contextually probable replacements for the open-class words shared between p and h . The variants are designed to maintain the underlying reasoning of the seeds, while also preserving common heuristics, such as sentence length, that models can still exploit.

The right side in Figure 1 shows the full methodology for evaluation of generalization on the generated variants (Minimal Expression-Replacement Generalization, MERGE). In MERGE we use a pattern-based accuracy metric (Abzianidze et al., 2023), hereafter called Seed-Variant (SV) accuracy, instead of a usual accuracy measure (i.e., sample accuracy). Thus, a model is evaluated not only on its predictions on seed problems, but also on its consistency, measured by predictions on corresponding variants of the seed problems.

We focus on the following questions: (i) how reliable is the methodology for automatically obtaining variants (MERE)? (ii) how well do NLI models and LLMs generalize to minimal variants (MERGE)? (iii) how do factors of variant generation, like the word class of replacements or the source MLM, influence model performance? Our contributions and findings are as follows:

1. A fully automated methodology for creating high-quality model-friendly test sets for generalization, applied to NLI.
2. Our results reveal poor generalization in NLI, in both fine-tuned NLI models and LLMs, even under minimally changed datasets.
3. We conduct extensive analyses that show verb re-

placements are difficult for models to generalize on, followed by nouns and adjectives. Surprisingly, the source MLM that generates the variant does not affect NLI models’ performance, even those that share architecture.

2 Related Work

We compare various OOD datasets that have been used in prior NLI work to facilitate comparison with our own work, and classify them in Table 1, according to various criteria as below:

Modification Type Modifications in the OOD dataset were made by *decomposing* or *paraphrasing* original problems, by *replacing* their words, or by combining *multiple* operations (Column Type).

Modification Procedure Modifications were made *automatically*, *manually*, or through a *mix* of both (Column Procedure).

Modified Unit Modifications were applied only to p , only to h , or across p and h ($p \wedge h$) (Column Unit).

Validation Modified problems were validated manually by having a partial or full set of variants annotated (H_p , and H_f), through a mix of automatic and manual methods (*Mix*), or not at all (*N/A*) (Column Validation).

Meaning, Reasoning, Word Overlap and Syntax The meaning of the problem (M), underlying reasoning (R), syntax (S), or word overlap of p and h (WO)¹ were *preserved* (+), *changed* (−), or a *mix* of both ($\pm$) (Columns under Preservation).

Evaluation Predictions on modified problems m were evaluated either *individually*, against the gold label (G_m) or the prediction on the initial problem (O_m), or *jointly*, against the gold label (G_{Ms}), the prediction on the original problem (O_{Ms}) or against each other (M_{sMs}) (Column Evaluation).

The performance drops in previous OOD NLI datasets cannot be attributed fully to poor generalization, as they were constructed with non-syntax-preserving constructions that might be more challenging for models (Li et al., 2020), or changed the lexical overlap of p and h due to paraphrased or decomposed NLI problems. Preserving word overlap between p and h has been partially explored

¹WO preservation requires modifying $p \wedge h$ at the same time.

Study	Modification					Preservation				Evaluation
	Type	Procedure	Unit	Dataset	Validation	M	R	S	WO	
Li et al. (2020)	Multiple	Auto.	p	SNLI; MNLI	H_p	$\pm$	$\pm$	$-$	$-$	$G_{Ms}; O_{Ms}$
Glockner et al. (2018)	Replace	Auto.	h	SNLI	H_f	$-$	$\pm$	$+$	$-$	G_m
Verma et al. (2023)	Paraphrase	Auto.	$p; h;$ $p \wedge h$	Pascal RTE1-3	H_f	$+$	$+$	$-$	$-$	O_{ms}
Srikanth et al. (2024)	Paraphrase	Mix	h	α -NLI; δ -NLI	H_f	$+$	$+$	$-$	$-$	$Ms_{Ms}; G_{Ms}$
Arakelyan et al. (2024)	Paraphrase	Auto.	h	SNLI; MNLI; ANLI	H_p	$+$	$+$	$-$	$-$	O_m
Petrov (2025)*	Multiple	Auto.	h	SNLI	N/A	$-$	$-$	$-$	$-$	G_m
Kaushik et al. (2020)	Multiple	Man.	$p; h$	SNLI	H_f	$-$	$-$	$-$	$-$	G_m
Srikanth and Rudinger (2025)	Decompose	Auto.	h	SNLI; δ -NLI	Mix	$+$	$+$	$-$	$-$	$G_m; O_{ms}$
MERGE (our method)	Replace	Auto.	$p \wedge h$	SNLI; MNLI	LM	$-$	$+$	$+$	$+$	$G_m; G_{Ms}$

Table 1: OOD NLI datasets classified by the type of modifications used to obtain variants (Type); the manual or automatic procedure (Procedure); the unit modified (Unit) and its origin dataset (Datasets – Pascal RTE1-3 (Dagan et al., 2005); α -NLI (Bhagavatula et al., 2019); δ -NLI (Rudinger et al., 2020)); and the manual or automatic method of validation (Validation). Other columns show the preservation of the meaning of the original problem (M), its reasoning (R), syntax (S) or word overlap (WO), and how variants were used in evaluation (Evaluation). §2 describes what each element of the columns means. * indicates that the variants were used for fine-tuning models. All studies use labels: entailment, contradiction, and neutral.

previously by Glockner et al. (2018), but their replacements did not preserve reasoning and favored variants’ plausibility at the expense of lexical diversity. Other shortcomings of previous variants’ datasets concern their automatic generation, previously criticized for potentially (i) biasing datasets in favor of models deployed (Li et al., 2020; Gardner et al., 2020); and (ii) constructing implausible datasets (Dutt et al., 2024). However, manual variant dataset creation is very time-consuming and unlikely to keep pace with the rate at which new models are developed, making a reliable automatic generation methodology necessary.

We propose a method that automatically creates high-quality plausible variants by replacing shared words of p and h with felicitous alternatives. Our method is the only one designed to preserve underlying reasoning, lexical overlap of p and h , and their syntax. Our method also limits the implausibility of constructed variants by selecting replacements that language models deem *at least as probable* in context as the words they replace (described in detail in section 3). Additionally, MERE can also vary the level of lexical diversity by including more or fewer MLMs.

3 Methodology

MERE takes an original seed NLI problem $\langle p, h, l \rangle$, labeled with l , and generates a set of replacement

words r from a set of MLMs for each open-class word o shared between p and h . Then, it quality-filters the replacements, unifies them across MLMs, and restricts them to those compatible with both p and h . Finally, MERE generates label-preserving variants $\langle p_i, h_i, l \rangle$ by replacing o with r .

First, the set r is collected from the set of MLMs $\mathcal{M} = \{M_1, \dots, M_n\}$ for a given sentence $S = (s_1, \dots, s_k)$ (either p or h), an MLM M_j , and an open-class word $o = s_i$, indicated by $R_j(S, i)$, as shown in the upper part of Figure 2.

Second, r is filtered with safeguarding constraints (i)-(iii) to create $R_j(S, i^c_>)$:

- (i) r has a higher probability than o in the masked context $S = (s_1, \dots, s_i/\text{MASK}, \dots, s_k)$ under M_j , marked with $i_>$;
- (ii) r and o are of the same word class² in S at position i , marked with i^c ;
- (iii) r does not occur in S

Constraints (i) and (ii) ensure that replacements r are *more* felicitous in S than o , according to MLM probability, while preserving the syntax of S . For example, ungrammatical or implausible variants that could have resulted from replacing the shared noun ‘girl’ in Figure 2 with a lower-probability replacement in the context, such as ‘crane’, or with

²The possible word classes are nouns, verbs, adjectives, or adverbs; note that if s_i is not part of the M_j vocabulary, the set of replacements will be empty.

Seed NLI problem
p: A **girl₁** carries a small **girl₂**. **h:** There is a small **girl₁**. **I: E**

MLMs replacements for shared words of *p* and *h*

girl ₁ : boy, child, man, crane _{PROB}	girl ₁ : boy, child, student, tree	MLM ₁
girl ₂ : child, student, man, boy, crane _{PROB}		
girl ₁ : boy, horse, tree, young _{ADJ}	girl ₁ : tank, boy, student, man	MLM ₂
girl ₂ : tank, boy, horse		

Quality filtering of replacements

Replacement	Higher or equal PROB	Same POS	S-level Eq. (1)	Problem-level Eq. (3)	PASS
boy	✓	✓	✓	✓	✓
man	✓	✓	✓	✓	✓
young	✓	✗	✗	✗	✗
crane	✗	✓	✓	✗	✗
tank	✓	✓	✗	✗	✗
tree	✓	✓	✗	✗	✗
horse	✓	✓	✓	✗	✗

Variants creation

p: A **man** carries a small **man** **h:** There is a small **man** **I: E**
p: A **boy** carries a small **boy** **h:** There is a small **boy** **I: E**

Figure 2: Generation of NLI variants with MLM replacements. Replacements must be at least as probable as *o* (Higher or equal PROB), preserve the class of *o* (Same POS), and be suggested by the same MLM for all multiple occurrences at the sentence level. At the problem level, replacements must be suggested for both *p* and *h*, although they might come from different MLMs.

a replacement with a different open class, such as ‘beautiful’, are avoided by (i) and (ii). Constraint (iii) also avoids label non-preserving changes. For example, substituting a word like ‘boy’ with ‘poodle’ in a problem with *p* ‘two poodles and a boy swim’ and *h* ‘only one boy swims’ would change the initial entailment to contradiction³.

After filtering, if a word *o* occurs multiple times in *S*, we define the replacement set from *M_j* as:

$$R_j(S, o_{>}^c) = \bigcap_{o=s_i} R_j(S, i_{>}^c) \quad (1)$$

For example, Figure 2 shows the exclusion of the replacement word ‘tank’ as it is not suggested for both sentence-level occurrences of ‘girl’ in *p*.

Afterward, for a set of MLMs $\mathcal{M}$, we define a set of replacements for a word *o* in *S* as:

$$R_{\mathcal{M}}(S, o_{>}^c) = \bigcup_{M_j \in \mathcal{M}} R_j(S, o_{>}^c) \quad (2)$$

where replacements are validated by the same

³While (iii) is a strong filter, variants with incorrect inference labels remain possible, although improbable, e.g., the replacement of ‘boy’ with ‘dog’ in the *p* ‘two poodles and a boy swim’; *h* ‘only one boy swims’.

MLM at each occurrence of *o* in *S*.

Finally, for a sentence pair $\langle p, h \rangle$ and $o \in p \cap h$, we define a set of replacements from $\mathcal{M}$ as:

$$R_{\mathcal{M}}(\langle p, h \rangle, o_{>}^c) = \bigcap_{S \in \{p, h\}} R_{\mathcal{M}}(S, o_{>}^c) \quad (3)$$

Note that intersection occurs when *r* needs to be compatible with two simultaneous positions of *o* (from multiple occurrences or sentences). Union, by contrast, expands the pool of *r* to maximize the potential final set of intersected *r*.⁴

Finally, the variants $\langle p_{ij}, h_{ij}, l \rangle$ of $\langle p, h, l \rangle$ are obtained by replacing the original shared words $o_i \in p \cap h$ with corresponding replacements $r_{ij} \in R_{\mathcal{M}}(\langle p, h \rangle, o_{i>}^c)$:

$$p_{ij} = p[o_i/r_{ij}], h_{ij} = h[o_i/r_{ij}]$$

as also shown in the last part of variant creation in Figure 2. We will use $R^{d=m}$ to denote a subset of size *m* of the replacement set *R*. 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 replacements 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.

Replacement generation Using SNLI test and the matched and mismatched MNLI-m/mm development sets, for each problem sharing at least a noun, verb, or adjective⁵ across *p* and *h*, we generated 200 replacements using each of the following MLMs: BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2019), ELECTRA (Clark et al., 2020), BART (Lewis et al., 2020). We used the base and large sizes for each MLM except ALBERT (details in Appendix Table 4).

Automatic quality filtering Replacements were kept if they were: i) of higher probability than *o* in the masked context⁶, ii) the same word class as *o*,

⁴We tested whether relaxing several criteria could lead to a ‘friendlier’ methodology, such as removing safeguards (i), (ii), applying no filters, or using union for *p* and *h* in Equation 3. However, our current methodology yields the best model results, as shown in subsection A.2, making it the ‘friendliest’.

⁵Adverbs were not included due to their low number of seed problems. SNLI and MNLI were chosen to maximize the pool of possible seed problems, being the largest NLI datasets.

⁶note that *r* rarer than *o* are not excluded by this constraint: we found that *o* are more frequent than corresponding *r* in large corpora, with an ≈ 0.5 higher Zipf frequency in

iii) different from words in p and h , iv) supported by at least one MLM. Most replacements were filtered out by (i), with only 10–30% remaining; see Appendix Table 5 for details. After quality filtering (see below), we retained only seed problems that had at least 20 variants.⁷

Manual Quality Validation The validation process had two stages. First, we manually validated 100 randomly sampled variants per open word class and dataset, for a total of 900 problems. We annotated the variants considering two aspects: the **reasoning** preservation of the original inference relation (R), and the **fluency** of the replacement (F). F also penalizes ungrammatical replacements, regardless of the grammaticality of a seed problem. We scored F and R from 1 to 5,⁸ and kept only variants with $F+R \geq 9$.

Two authors independently annotated 300 SNLI problems first, with a high inter-annotator agreement (Cohen’s $\kappa = .80, .77$, and $.89$ for noun, adjective, and verb variants, respectively). All BART replacements were excluded given they contributed more to forming bad variants, as shown in the distribution of good and bad SNLI variants in Appendix Figure 9. We then repeated this first validation phase for MNLI-m/-mm, where one author annotated all 600 examples. These variants, which already excluded BART replacements, had all good scores (≥ 9) except for one with a lower fluency score.

In the second annotation stage, one author annotated 100 variants per dataset, across classes. All sampled variants for MNLI-m, 98% for MNLI-mm, and 91% for SNLI had a good $F+R$ score. No variant received a rating with an R score below 5 or was judged ungrammatical⁹. MERE thus has a substantially lower error rate than previous work, allowing us to retain more of the automatically created dataset. For comparison, Verma et al. (2023) reported an $\approx 44\%$ error rate. Consult Appendix subsection A.3 for the full validation guidelines, statistics on diversity of replacements, and an additional annotated variant sample with an LLM.

wordfreq (Speer, 2022).

⁷The datasets obtained are publicly available at <https://github.com/mdzgr/MERGE>.

⁸For example, ‘Colorless green ideas sleep furiously’ entailing ‘Green thoughts is angrily sleeping’ should be assessed with 1 for F and 5 for R, in short F1-R5.

⁹The difference in grammaticality between seed problems and their variants is only ≈ 0.03 for all datasets, as revealed by a RoBERTa CoLA-trained model (textattack/roberta-base-CoLA).

Dataset	Class	# Seeds	Gold label (%)			Unique
			N	C	E	
SNLI	Var _N	1808	29	20	50	67
	Var _V	506	30	18	50	66
	Var _A	259	32	22	44	63
	Var	2222	30	20	49	77
MNLI-m	Var _N	3058	25	31	43	86
	Var _V	648	22	27	50	69
	Var _A	703	26	23	49	72
	Var	3663	25	30	44	95
MNLI-mm	Var _N	3452	25	31	43	89
	Var _V	683	21	27	51	68
	Var _A	779	23	28	48	73
	Var	4016	24	30	45	103

Table 2: Number of final seed problems per open word class (#Seeds) and dataset, with gold-label percentages, and the average number of unique variants per problem across all subsamples of Var (Unique). #Seeds is non-summativ, as problems with replacements from multiple word classes are counted only once in Var.

Final variants For each dataset, we randomly sampled only 20 variants per open word class per seed problem, to avoid overpopulating the dataset with variants from ‘productive’ seeds, i.e. seeds with a lot of generated variants. We repeated this subsampling 10 times to avoid randomization artifacts and formed the Var dataset, comprising these 10 subsamples for each open word class. Table 2 shows the number of seeds for each Var, alongside their label distributions.

Evaluation Metrics We evaluate models with two metrics: standard accuracy, where each variant or seed is an individual evaluation point (sample accuracy, S-acc), and SV-acc, borrowed from Abzianidze et al. (2023). Under SV-acc, a seed problem is considered correctly classified only if a minimum proportion of its variants are as well; see Figure 1. We call this proportion the correctness threshold (CT). If models generalize well, their SV-acc¹⁰ over variants should not substantially fall behind their Seed S-acc. To quantify this, we define the matching correctness (MC) threshold as the CT at which a model’s SV-acc over variants matches its S-acc over seed problems.

NLI Models We evaluated several models fine-tuned on NLI to observe how NLI fine-tuning affects performance: BERT, RoBERTa, DeBERTa (He et al., 2021), BART, ALBERT, ELECTRA, XLNet (Yang et al., 2019), OPT (Zhang et al., 2022),

¹⁰Note that for Var the SV-acc for a CT is the averaged SV-acc for that CT across all its 10 subsamples.



Figure 3: S-acc of the two best and the worst models trained either on SNLI, MNLI, SMFA, or, in the case of LLMs, prompted with examples exclusively chosen from the development set of SNLI and the train set of MNLI. Test/Dev is the S-acc on the original test problems for SNLI and dev problems for MNLI, while Seed and Var show the S-acc on seed problems, exactly shown by the orange numbers, and variant problems, respectively.

and GPT-2 (Radford et al., 2019), fine-tuned on SNLI, MNLI, or a combination of SNLI, MNLI, FEVER (Nie et al., 2019), and ANLI-SMFA. Full details of the models used are in Appendix Table 7.

LLMs We also evaluated two LLMs: Gemma-2-9b (Team, 2024), and Llama-3.1-8B (Dubey et al., 2024). For their evaluation, we randomly sampled 500 seed problems and their corresponding variants per dataset, picked to preserve the original seed distribution and Seed S-acc. We evaluated them in a multiple-choice task setting following the prompting strategy from Madaan et al. (2025) with six-shot¹¹ prompting examples either exclusively sampled from the dataset corresponding to the variants (SNLI-dev or MNLI-train), or as a 3+3 mix from both datasets. Lastly, we selected the model’s label prediction based on its highest logit. Section §A.4 of the Appendix shows additional details on seed subsampling, prompting, or few-shot examples.

5 Results

5.1 Do models generalize to variants?

We found that while models achieve high S-acc on individual variants, their performance drops significantly under SV-acc as we increase the Correctness

Threshold (CT).

Figure 3 shows results for the two best-performing and the worst-performing model per training data category¹². Their S-acc on individual variants is close to that on seed problems, as shown by the green and orange bars. In contrast, the matching correctness (MC) heatmap (Figure 5) shows that, across datasets, most models cannot maintain their Seed S-acc when correctness is required on more than $\approx 50\%$ of variants for each seed. Above a CT of $\approx 60\%$, performance drops across models and datasets, especially at higher thresholds. Llama-3.1-8B-exc is the model that generalizes best on variants from MNLI-mm, i.e., with an MC of $\approx 54\%$, while BERT-L-S has the highest MC for MNLI-m and SNLI, i.e., 58% on MNLI-m. Figure 4 further illustrates the models and their performance on different CTs. The differences between models’ SV-acc over variants and Seed S-acc were shown to be statistically significant across models starting from the 60% CT¹³. A particularly steep decline is seen in Figure 4 for LLMs in the green and purple lines, even though Llama-3.1-8B-exc had the highest MC for MNLI-

¹¹Six is the smallest even number above the 5 few-shot setting after which performance gains diminish in Madaan et al. (2025).

¹²Selected by their averaged S-acc on test problems for SNLI and S-acc on dev problems for MNLI. Results for all models are in the Appendix, Figure 12.

¹³Shown via a paired t-test (SNLI with $p = 0.0027$, MNLI-m with $p = 0.0003$, and MNLI-mm with $p = 0.0001$).

mm.¹⁴ In fact, LLMs generally perform worse than all other NLI models, with S-acc lower on test or dev problems by as much as 20%, as also shown previously by Madaan et al. (2025).

Models generalize better on MNLI-m/-mm variants than on those from SNLI, as shown by the higher MCs in Figure 5. This may be due to differences in seed difficulty: MNLI seeds are easier than their respective full MNLI dev sets, while SNLI seeds are comparable in difficulty to their SNLI test problems. This difference is statistically significant for both MNLI-m/-mm variant datasets¹⁵. A random-subset analysis further supports this result: MNLI-m/-mm seeds were consistently easier than 100 equally sized random subsets drawn from their corresponding Dev sets.¹⁶

Overall, our results confirm but are stronger than previous studies that showed models fail to generalize to datasets containing minimal changes (Verma et al., 2023; Arakelyan et al., 2024; Glockner et al., 2018), with our work additionally demonstrating

¹⁴See description of selecting models for presentation, full model results and statistical analysis in the Appendix.

¹⁵A paired t-test comparing model S-acc on Seed vs. Test/Dev was not significant for SNLI ($p = 0.6222$), but was significant for both MNLI-m/-mm ($p < 0.0001$).

¹⁶In comparison, models achieved higher S-acc in 55% of the randomly drawn SNLI subsamples. The MNLI-m/-mm variants subselected for LLM evaluation are also less difficult, as they were chosen to match the main Seeds in difficulty.

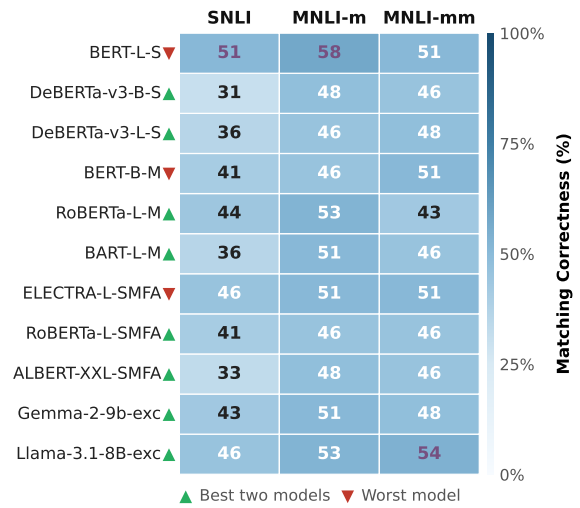


Figure 5: Matching Threshold, MC (the CT at which a model’s SV-acc on variants matches its Seed S-acc), across models and datasets. 100% MC would indicate perfect generalization, i.e., accuracy on all variants is the same as accuracy on seed. The last acronym in the model names specifies the datasets the models were fine-tuned on, or the type of prompting used. Best MCs are marked in purple, while MCs under 45 are in black.

that this is true on even easier datasets¹⁷, obtained

¹⁷We evaluated the hypothesis-only baselines on the generated variants for SNLI and MNLI and observed that the hypothesis-only artifacts are still substantially present in the variants, with 57% and 51% accuracy scores, respectively, in

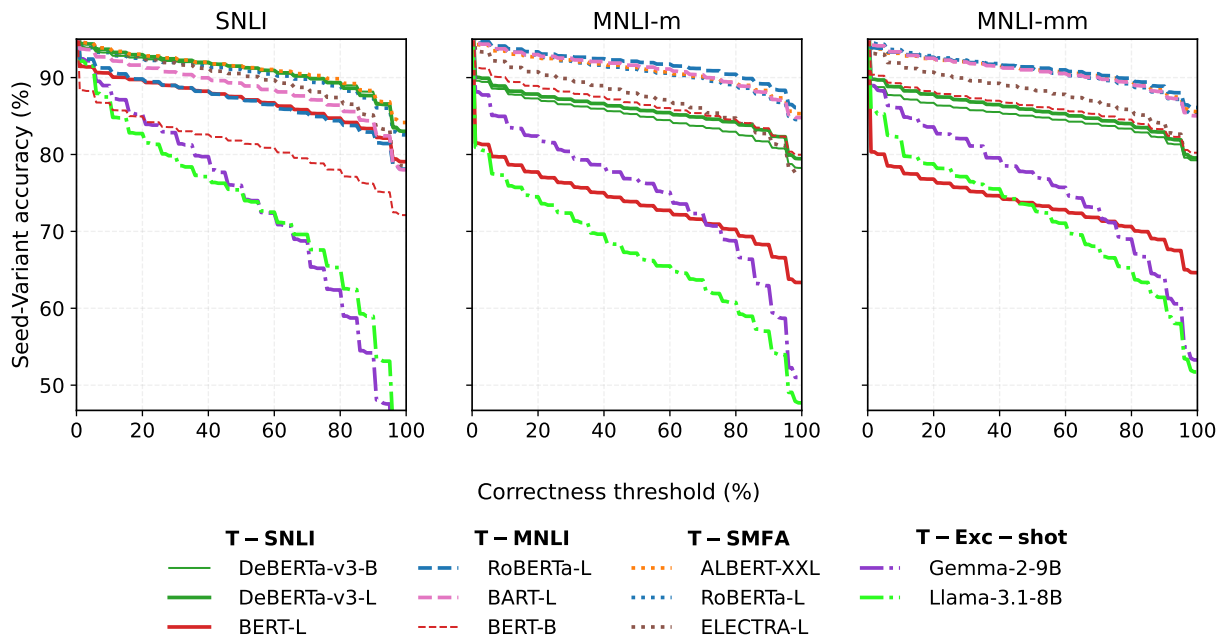


Figure 4: SV-acc vs. CT on variants from SNLI and MNLI-m/-mm. The legend headers show the datasets the NLI models were fine-tuned on, i.e., SNLI, MNLI, or a combination of NLI datasets e.g., SMFA, or, for LLMs, the few-shot source: exclusively drawn from the source dataset of the variants (SNLI-dev for SNLI; MNLI-train for MNLI-m/-mm).

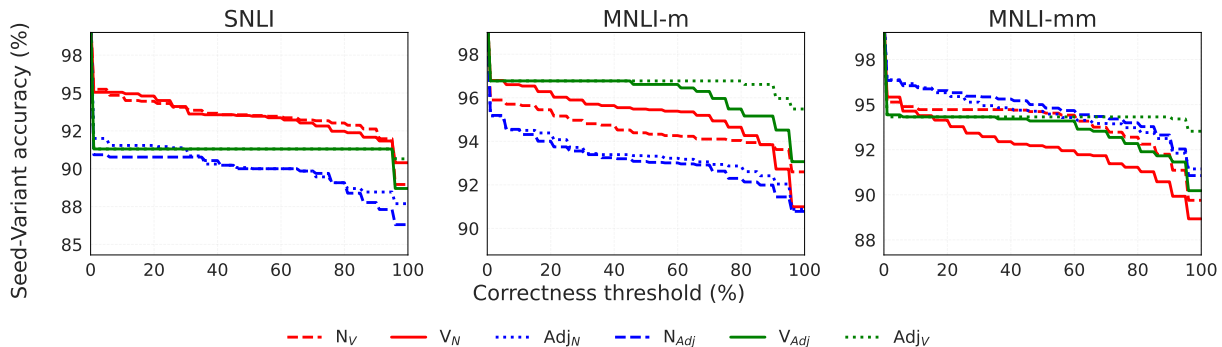


Figure 6: SV curves of ALBERT-XXL on seed problems sharing at least two out of the three open classes. The baseline letters in the dataset names (N, V, Adj) show which open-class was replaced to form variants.

from MERE. Our use of SV-acc especially shows how considering variant proportions per seed can reveal generalization shortcomings, while depending on just sample accuracy S-acc on variants, as is common, often blurs this effect.

5.2 Which word classes are harder to generalize?

To investigate whether replacing nouns, verbs, or adjectives affects model performance, we compared scores on variants grouped by the open class replaced. We restrict this test to seed problems that share at least two of the three open classes replaced¹⁸, such as ‘A girl carries a small girl’, to directly observe the effect of replacing one of the classes as opposed to the other, in this case the noun ‘girl’ opposed to the adjective ‘small’, and separate that from the differences in problem difficulty and imbalanced number of variants across open-classes, as shown in Table 2.

We found that variants generated by replacing verbs are more difficult than those generated by replacing nouns, followed by adjectives, as shown in the results of the model with the highest averaged S-acc for SNLI and MNLI, ALBERT-XXL, from Figure 6. In the legend, baseline letters indicate which open class was replaced in seed problems sharing the two open classes in each dataset name. Across datasets, adjectives are initially similar to verbs in lower CTs but become easier in higher ones (green lines). At lower CTs, verbs are easier than nouns in MNLI-m and SNLI, and more difficult than nouns in MNLI-mm (red lines). At high CTs, verbs are more difficult than nouns in MNLI, with nouns being more difficult in SNLI. Lastly,

contrast to 33% of a random baseline.

¹⁸For SNLI (NV=202; NAdj=130; VAdj=46); MNLI-m (NV=422; NAdj=312; VAdj=62); and MNLI-mm (NV=516; NAdj=360; VAdj=88).

adjectives are easier than nouns in high and low CT, as shown by the blue lines.

5.3 Do source MLMs affect generalization?

We further investigated whether NLI models generalize better on variants generated by an MLM with a similar backbone, as it has been suggested in the case of synthetic evaluation datasets that they can favor evaluated models that share the same architecture as the models used to generate the data (Li et al., 2020; Gardner et al., 2020).

We restricted the set of seed problems to those for which variants could be generated using replacements from each of the following MLMs: BERT, RoBERTa, ELECTRA, and ALBERT.¹⁹ We also consider only the results of the corresponding NLI models.

Figure 7 shows the SV curves of the best-performing NLI model. Dataset names indicate the MLM used to generate replacements and its relation to the evaluated NLI model: an MLM with the same backbone and size (Equiv); an MLM with the same backbone, but a different size (Size), e.g. BERT-B-S evaluated on BERT-L-generated variants; multiple MLMs, potentially including one with the same backbone as the evaluated NLI model (Multi); or one MLM model with a different architecture (One).²⁰

While scores are consistently high on Equiv variants across datasets (green lines), overall scores on other datasets suggest that the MLMs do not favor NLI models that share their architecture. If they did,

¹⁹Amounting to 150 problems in SNLI, 530 in MNLI-m, and 424 in MNLI-mm, subsampled to isolate the effect of the origin of replacements from that of imbalanced seed problems (see Appendix Table 6 for statistics).

²⁰Note that the same variants, e.g. from BERT-B, can be the equivalent size variants for BERT-B-S, and of different size for BERT-L-S.

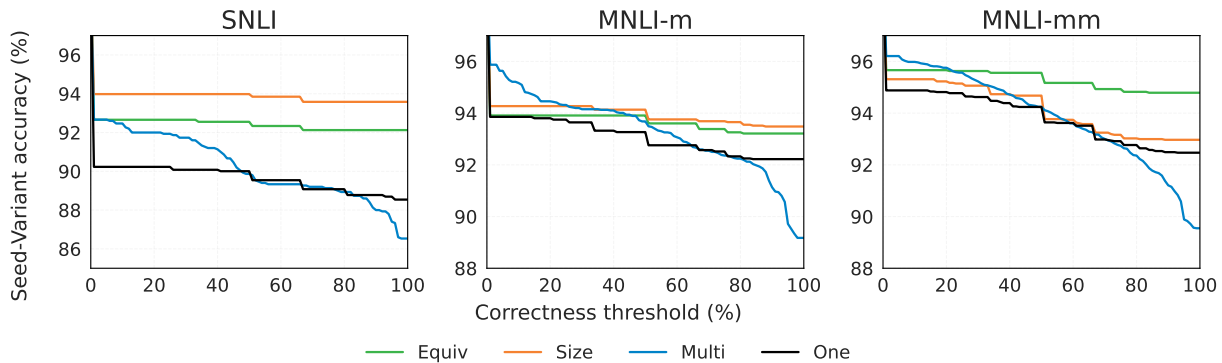


Figure 7: SV curves for ALBERT-XXL on the seed problems that had variants formed with replacements from BERT, RoBERTa, ELECTRA and ALBERT, all sizes tested, grouped by their origin MLM.

we would expect Size to always have the second-highest scores across datasets (orange lines), which is contradicted by its highest scores for SNLI and MNLI-m. Similarly, we would expect One variants to be classified most poorly (black line), followed by Multi variants (blue line), since it also includes replacements from MLMs that share the architecture with the evaluated NLI model, which could favor performance. However, the graphs show that models score lowest at higher CTs on Multi.

5.4 What do the hard NLI problems look like?

We analyzed how successfully, on average, the models classified the original seed problems for all three datasets. We found that all NLI models incorrectly classified 0.7% of SNLI (16), 0.05% of MNLI-m (2), and 0.29% of MNLI-mm (12) problems²¹. Out of these problems, 66% had high label variation (see the problems in Table 12). For example, a problem with p ‘A man in a colorful shirt and a lady in a white blouse sign copies of books for people’ and h ‘Two people sign copies of their latest novel’ has been classified as entailment by 3 out of 5 annotators. The world knowledge that people most often sign books they wrote explains the entailment label; however, omitting the adjective ‘latest’ from h also supports the neutral label. This adds evidence to the well-known issue of annotation variation in NLI (Pavlick and Kwiatkowski, 2019; Weber-Genzel et al., 2024), and higher model error rates on such problems (Madaan et al., 2025). Models also have higher accuracy on variants of seeds they initially classified correctly, as shown by Ohmer et al. (2024): 97% of their variants are correctly classified, compared to $\approx 80\%$ for seeds which were incorrectly classified.

Occasionally, models correctly predicted variants of incorrect seeds (avg. seeds across models of SNLI=98, MNLI-m=207, and MNLI-mm=210).

6 Conclusions

We introduced MERE, a methodology for automatically generating minimal generalization variants of NLI problems designed to preserve reasoning and word overlap. Its strength lies in minimal, controlled replacements with quality filters for fluency and consistency, with MERGE being the full methodology of testing models under the simplest generalization conditions.

We show that while NLI models, across sizes and architectures, perform well on variants considered individually, they perform worse when correctness is required for more than about half of the variants of SNLI and MNLI-m/-mm. More specifically, MC values are overall around $\approx 50\%$, indicating seed accuracy holds only until that CT. Our results also indicate that variants formed from verbs are generally harder to generalize to than those formed from nouns and adjectives, and that replacements from different MLMs do not affect model scores, suggesting a loss in generalizability regardless of the (mis)match between the origins of MLMs and NLI models.

Although we currently applied MERGE to NLI, we believe that the method is general and robust enough to be extended to other natural language understanding tasks that allow (consistent) replacements of words without affecting the ground truth answers, such as, extractive QA, multiple-choice commonsense/reasoning QA, and semantic textual similarity.

²¹These problems are not in the sub-sample used for LLMs.

Limitations

In MERE we did not replace words tokenized into subwords, as aggregating their probabilities could introduce artifacts, which could be explored in future work. Additionally, when original words appear multiple times in a problem, we mask only one instance at a time, leaving the others unmasked, which could bias MLM replacements. However, problems with such cases are a small proportion of the seed datasets (SNLI $\approx$ 2%, MNLI-m $\approx$ 6%, and MNLI-mm $\approx$ 9%). Another methodological limitation is that we do not replace morphological derivations of original words to avoid incorrectly replacing them when they appear in nominal compounds. However, this can result in low-fluency variants. For example, not replacing *sleep* from ‘sleeping’ in the hypothesis ‘She sleeps, and she will be forever sleeping’, could result in implausible variants such as ‘She runs, and she will be forever sleeping’. Such cases are rare in the seed datasets as well (SNLI $\approx$ 5%; MNLI-m $\approx$ 7%, and MNLI-mm $\approx$ 6%). Lastly, our analysis could be extended to include larger LLMs.

Acknowledgments

We thank the anonymous reviewers for their useful comments. We also thank the NLP Reading Group at Utrecht University and the NALOMA audience for their useful feedback on this work.

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.
- 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](#). *International Conference on Learning Representations (ICLR)*.
- Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. [A large annotated corpus for learning natural language inference](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.
- Samuel R. Bowman and George Dahl. 2021. [What will it take to fix benchmarking in natural language understanding?](#) In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online. Association for Computational Linguistics.
- 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.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.
- 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 Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh 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](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

- 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](#). *International Conference on Learning Representations (ICLR)*.
- 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](#). *International Conference on Learning Representations (ICLR)*.
- Andrew K Lampinen, Ishita Dasgupta, Stephanie C Y Chan, Hannah R Sheahan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. 2024. [Language models, like humans, show content effects on reasoning tasks](#). *PNAS Nexus*, 3(7):pgae233.
- 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, Veselin Stoyanov, and Luke Zettlemoyer. 2020. [BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- 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.
- Lovish Madaan, David Esiobu, Pontus Stenetorp, Barbara Plank, and Dieuwke Hupkes. 2025. [Lost in inference: Rediscovering the role of natural language inference for large language models](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9229–9242, Albuquerque, New Mexico. 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, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Association for the Advancement of Artificial Intelligence (AAAI)*.
- Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. [Adversarial NLI: A new benchmark for natural language understanding](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.
- Xenia Ohmer, Elia Bruni, and Dieuwke Hupke. 2024. [From form\(s\) to meaning: Probing the semantic depths of language models using multisense consistency](#). *Computational Linguistics*, 50(3):1507–1556.
- 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](#). In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

- Sara Rajaei, 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, Online. Association for Computational Linguistics.
- Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and et al. 2026. [Openai gpt-5 system card](#). *Preprint*, arXiv:2601.03267.
- Robyn Speer. 2022. [rspeer/wordfreq: v3.0](#).
- 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](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2574–2589, Albuquerque, New Mexico. Association for Computational Linguistics.
- Gemma Team. 2024. [Gemma 2: Improving open language models at a practical size](#). *Preprint*, arXiv:2408.00118.
- Masatoshi Tsuchiya. 2018. [Performance impact caused by hidden bias of training data for recognizing textual entailment](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- 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

Dataset	R	#PS	N(%)	C(%)	E(%)
SNLI	N	7363	33	29	38
	V	3780	32	28	40
	Adj	1067	34	24	42
	Adv	76	26	13	61
MNLI-m	N	5979	29	31	40
	V	4438	27	32	41
	Adj	1923	26	25	48
MNLI-mm	N	6640	29	31	40
	V	4614	27	32	42
	Adj	2197	26	28	47

Table 3: Number of potential seed problems from SNLI-test and the development set of MNLI-m/-mm (#PS). Columns R, N, C, and E show the divisions by replaced open-class and gold label (% neutral, contradiction, entailment). We ultimately excluded adverbs because of their final low seed count in SNLI.

MLMs The MLMs used for replacement generation are shown in Table 4. In addition, we

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

Table 4: Overview of MLMs used for replacement generation. The Size, Architecture, and Vocabulary columns show the model size (Base – B, Large – L, or XXL – XXL), architecture (Encoder, Decoder, or Encoder-Decoder), and vocabulary size.

considered DeBERTa, but its replacements were frequently too implausible to include. From the models shown in Table 4, we excluded BART because it contributed more to bad variants in SNLI, as explained in subsection A.3. We did not consider BART by default for MNLI-m/-mm replacements, to avoid creating plausibility differences across variants from different datasets.

Class & Probability Filtering We POS-tagged replacements r in the context of their corresponding problems using the spaCy model `en_core_web_sm`, and excluded those whose word classes differed from the original word o . We also recorded the probability the source MLM assigned to each r of o in p or h , and excluded those with a lower probability than the original word. MERE uses probability as a safeguarding criterion, assuming that replacements as likely as o are less likely to be semantically implausible. Original words o_i that were not part of the model’s vocabulary were also automatically excluded. Note that both probability and open-class filtering are important, as highly probable replacements might not have the correct open-class and vice versa. Table 5 shows the percentages of replacements with higher probability than o under rows Prob, and with the correct open-class in POS.

A.2 Which filtering criteria matter?

To test if different quality filtering criteria for r affect NLI models’ scores, we formed new variants for the seed problems of SNLI Var with r i) from the union of p and h , instead of their intersection in Equation 3 – $p \cup h$; ii) only having o^c – POS; iii) only having $o_{i>}$ – Prob; iv) of any open-class or probability – None; v) with their letters randomly scrambled – Scr. We still excluded punctuation signs and r already part of p and h . Three out of the five datasets had more unique variants given the less strict filtering criteria, i.e. $\approx 380k$, while Scr and Var have $\approx 190k$. We plot the averaged SV-acc of all models in Figure 8.

While Scr starts as the lowest curve (in green) in Figure 8, at very high CTs, models end up having the highest robustness on it, surprisingly. This was also shown with the nonce words in Lampinen et al. (2024).

The other variant datasets seem to follow a constant downward trend, namely, the highest score is achieved on Var, followed by Prob, POS, None, and $p \cup h$, where the more unique variants are considered, the lower the scores get. Out of these, $p \cup h$, POS, and None have the lowest scores, indicating that not controlling for probability or plausibility might disadvantage the models. Thus, the higher scores achieved on Var show that our enforced quality criteria successfully create ‘friendly’ tests.

A.3 Annotation

Two authors annotated variants for SNLI, and one for MNLI. We evaluated replacements based on how much they changed fluency and the original logical label of the NLI problems, which we considered correct. The scores used were: 1 – poor, 2 – mostly poor, 3 – uncertain, 4 – mostly good, 5 – good. Table 10 shows the annotation guidelines, alongside an explanation for them. Variants were classified as poor if they: 1) had ungrammatical replacements making manual validation difficult; 2) had missing arguments; 3) were nonsense; or 4) were logic-non-preserving. Aspects 1) and 2) were chosen given that they directly hinder the evaluation of fluency and reasoning.

The instructions in the annotation guidelines are provided below with the demo examples in Table 10:

The NLI problems are assessed on fluency (grammaticality and sensibility) and reasoning. The reasoning component focuses on the relation

D	C	M	A-B	A-XXL	BA-B	BA-L	BE-B	BE-L	E-B	E-L	R-B	R-L
SNLI	A	POS	52	52	36	35	52	53	52	52	49	50
		Prob	17	18	15	16	10	8	14	12	9	8
	N	POS	93	94	78	75	91	91	90	90	92	92
		Prob	18	20	14	16	13	11	16	14	10	8
	V	POS	72	76	63	63	72	71	72	74	68	68
		Prob	12	13	12	13	9	7	12	11	7	6
MNLI-m	A	POS	53	54	37	34	53	53	53	53	53	53
		Prob	34	22	28	27	19	15	22	20	18	15
	N	POS	88	90	73	72	86	86	86	86	88	88
		Prob	34	23	28	26	22	18	26	22	19	16
	V	POS	75	76	58	59	73	73	71	71	71	71
		Prob	11	8	11	11	7	6	8	7	6	5
MNLI-mm	A	POS	55	56	38	34	56	56	54	55	54	55
		Prob	33	22	27	28	18	16	22	19	17	14
	N	POS	89	90	73	72	87	88	86	87	89	89
		Prob	32	21	27	26	21	18	25	21	19	16
	V	POS	74	75	58	60	72	72	70	70	71	71
		Prob	11	7	11	10	6	5	8	6	6	5

Table 5: The average percentage of r with the same open class (POS) and equal or higher probability than o (Prob) for sentences (p and h) across the three datasets. Column names indicate the origin model of r , e.g., ALBERT (A), BART (BA), BERT (BE), ELECTRA (E), and RoBERTa (R), for different model sizes.

between the meaning of p and h rather than their fluency.

The original NLI 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, which helps to reconstruct the original NLI 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.

In the first stage of annotation, only a few SNLI (5%) variants received scores lower than 3 on R,

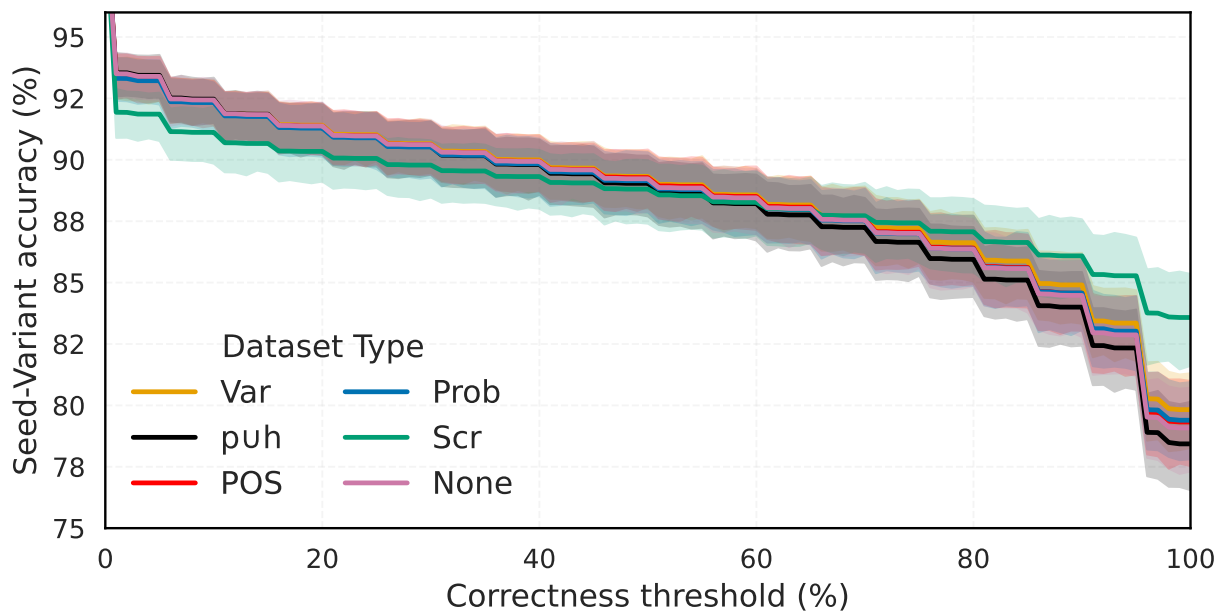


Figure 8: Averaged SV curves of all models on SNLI datasets formed with different quality filtering criteria, for all CTs.

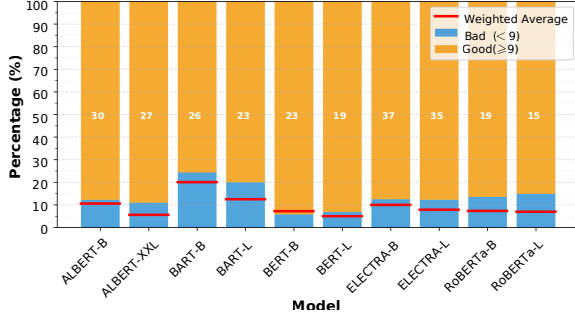


Figure 9: Averaged fluency and reasoning scores for the normalized counts of 100 random SNLI variants for nouns, verbs, and adjectives. The red lines show bar plots weighted by the distribution of open classes in the SNLI seed problems ($N=67\%$, $V=23\%$, $ADJ=10\%$). Good variants have a score of $F + R \geq 9$.

due to a lack of label preservation or ungrammaticalities. After this first stage, BART replacements were excluded as they contributed more to bad variants, as shown in Figure 9, resulting in a minor reduction of 318 SNLI seed problems. We did not exclude other models, since the pairwise overlap matrix in Figure 11 showed more unique than overlapping replacements across models, with only BERT and ELECTRA showing nearly equal proportions of both.

Given that only one author annotated all MNLI-m/mm variants (800), we asked GPT-5.3 (Singh et al., 2026) to flag implausible or non-preserving reasoning variants. The model flagged 18 examples (15 for plausibility and 3 for reasoning), out of which only 1 was implausible, also identified by the previous author annotator: e.g. p : ‘Well, it

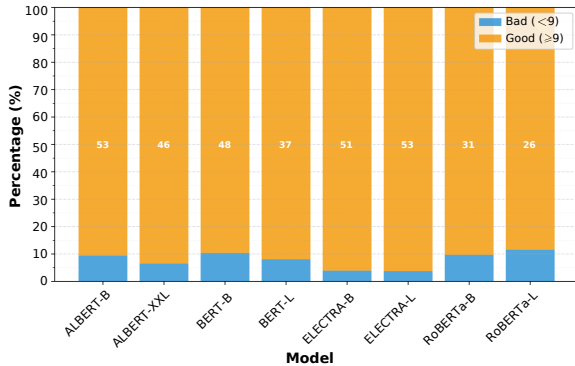


Figure 10: Fluency and reasoning scores for 100 randomly sampled SNLI variants across classes, their normalized counts, and the models that validated them, after exclusion of replacements from BART. Note that 91 out of the 100 examples evaluated had $F + R \geq 9$.

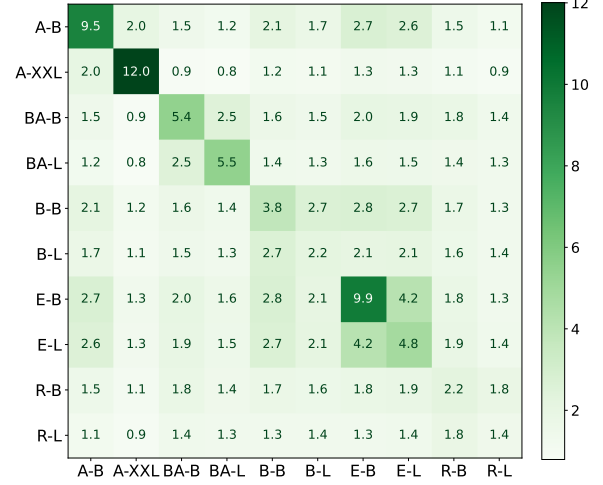


Figure 11: Pairwise overlap matrix showing how models’ replacements intersect on average, across problems, with the averaged number of model-specific replacements on the diagonal. For instance, a replacement like ‘bear’ counts on the diagonal if it is unique to one model, or in the other cells if it is shared between models.

would probably been things like living back in the hills of TN and having to ride a mule to wau, things like that’, h : ‘While living in the hills of TN, I rode a mule to wau’, where the word ‘school’ was replaced by ‘wau’.

A.4 Evaluated models

Evaluated Models The evaluated models and their model cards are shown in Table 7.

A.4.1 LLMs

Seed problems For each seed dataset, we randomly formed 100 subsets of 500 randomly sampled seed problems and compared models’ NLI S-acc on them. Model scores on subselected seeds varied by around 3% from their scores on the big-

Model	SNLI		MNLI-m		MNLI-mm	
	Raw	%	Raw	%	Raw	%
ALBERT-B	7,814	5	37,568	10	40,802	10
ALBERT-XXL	12,820	7	13,657	4	15,784	4
BERT-B	1,979	1	4,308	1	5,209	1
BERT-L	1,043	1	2,405	1	2,890	1
ELECTRA-B	7,750	5	10,156	3	11,117	3
ELECTRA-L	2,827	2	4,749	1	5,476	1
RoBERTa-B	1,441	1	4,114	1	4,658	1
RoBERTa-L	775	0	2,825	1	3,242	1
Multi	135,615	79	281,866	78	325,113	78

Table 6: Variants across datasets grouped by the MLM that suggested their replacement, with raw and % counts shown. The Multi-row shows replacements suggested by at least two MLMs.

Evaluated Model	Model Card
<i>SNLI-only models</i>	
BART-B-S	varun-v-rao/bart-base-snli-model1
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
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
<i>MNLI-only models</i>	
BERT-B-M	textattack/bert-base-uncased-mnli
RoBERTa-B-M	roberta-base-mnli
RoBERTa-L-M	roberta-large-mnli
BART-L-M	facebook/bart-large-mnli
<i>SMFA models</i>	
ALBERT-XXL-SMFA	ynie/albert-xxlarge-v2-snli_mnli_fever_anli_R1_R2_R3-nli
BART-L-SMFA	ynie/bart-large-snli_mnli_fever_anli_R1_R2_R3-nli
ELECTRA-L-SMFA	ynie/electra-large-discriminator-snli_mnli_fever_anli_R1_R2_R3-nli
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
<i>LLMs</i>	
Gemma-2-9b-exc/mix	google/gemma-2-9b-it
Llama-3.1-8B-exc/mix	meta-llama/Meta-Llama-3.1-8B

Table 7: Evaluated models and their Hugging Face references. For LLMs, the suffix *-mix/-exc* indicates which type of few-shot examples were used in the prompts, i.e. exclusively from the dataset the variants are based on, or a mix of both SNLI and MNLI datasets.

ger seed datasets, indicating that random subsampling roughly preserves the original seed distribution. From these subsets, we selected a random sample from each dataset to test LLMs on.

Instructing and evaluation We obtained logits for google/gemma-2-9b and meta-llama/Llama-3.1-8B with: temperature=0, max. no. of tokens=20, torch dtype=torch.bfloat16, and attention = sdpa. The prompt structure is shown in Table 8, while the few-shot examples per dataset appear in Table 11. Note that the few-shot examples are balanced across labels and are of

Prompt Template
{% for x in few_shot -%} Premise: {{ x["premise"] }} Hypothesis: {{ x["hypothesis"] }} A. Entailment B. Neutral C. Contradiction Answer: {{ x["answer"] }}
{% endfor -%} Premise: {{ premise }} Hypothesis: {{ hypothesis }} A. Entailment B. Neutral C. Contradiction Answer: {{ choice_text, e.g. A }}

Table 8: Prompt template used for LLM evaluation from Madaan et al. (2025).

Comparison	Selected	All
<i>SNLI</i>		
S-SV 50	0.0118	0.0002
S-SV 60	0.0027	< 0.001
S-SV 70	0.0020	< 0.001
S-SV 85	0.0037	< 0.001
S-Test/Dev	0.6222	0.9867
<i>MNLI-m</i>		
S-SV 50	0.0776	0.0087
S-SV 60	0.0003	< 0.001
S-SV 70	0.0002	< 0.001
S-SV 85	0.0005	< 0.001
S-Test/Dev	0.0018	< 0.001
<i>MNLI-mm</i>		
S-SV 50	0.1738	0.0285
S-SV 60	0.0001	< 0.001
S-SV 70	0.0009	< 0.001
S-SV 85	0.0017	< 0.001
S-Test/Dev	< 0.001	< 0.001

Table 9: P-values from paired *t*-tests comparing Seed S-acc with SV-acc at different correctness thresholds and with Test/Dev S-acc. Selected refers to the two best-performing and one worst-performing model shown in Figure 3; All includes all evaluated models.

2 types, i.e. randomly drawn from the variants’ source dataset, or a mix from both datasets. With our evaluation method, we replicated the scores of Madaan et al. (2025) for Llama-3.1-8B on SNLI-dev twice: with the current batch of exclusively SNLI-dev shots from Table 11, and 6 other randomly chosen ones from SNLI-test. In both cases, our scores were around 74-76% accuracy, which is slightly higher than their

reported 70% for 5-shot examples. Finally, for our variant evaluation, we evaluated both LLMs only once over the set of unique variants in v_{ar} to halve computation time. We investigated whether running Llama directly on the full variant dataset, which contains duplicates, changes its predictions, and found that it does so in 0.1% of cases.

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 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 being metaphorical, and it is still likely to preserve the initial inference label, despite the hypothesis's ungrammaticality.
This child is took a pedicure.	Child took a manicure.	C	getting	took	1	3	Uncertain: the ungrammaticality of the premise makes it difficult to assess its fluency; however, the main source of the contrast giving the contradiction (i.e. <i>pedicure</i> vs. <i>manicure</i>) 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 <i>with</i> a broom, running <i>on</i> something does not entail running <i>with</i> 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 assess.

Table 10: 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 neutral (N), entailment (E), and contradiction (C).

Dataset	Few-shot examples
SNLI	Premise: A man with a beard skateboarding and a boy with a blue and black backpack riding a green bike in the background. Hypothesis: There is a man and a boy outside. A. Entailment B. Neutral C. Contradiction Answer: A Premise: There is a room full of pictures all in the wall and a woman in a coat is looking back over her shoulders strangely. Hypothesis: the room is large A. Entailment B. Neutral C. Contradiction Answer: B Premise: A Black woman on the street is talking on her cellphone. Hypothesis: A black woman is on the payphone, ordering pizza A. Entailment B. Neutral C. Contradiction Answer: C Premise: The side of a building next to a church is painted with a brightly colored Coca-Cola sign. Hypothesis: There is a Coca-Cola sign on a church next to a building. A. Entailment B. Neutral C. Contradiction Answer: C Premise: A butterfly costumed girl waves at the crowd. Hypothesis: A butterfly costumed girl is standing on the stage. A. Entailment B. Neutral C. Contradiction Answer: B Premise: A man holding his two adorable babies. Hypothesis: A man has more than one child. A. Entailment B. Neutral C. Contradiction Answer: A
MNLI	Premise: Ras Mohammed National Park has over 1,500 species of fish and 150 types of coral along with an offshore vertical sea wall. Hypothesis: The national park contains many types of fish and coral. A. Entailment B. Neutral C. Contradiction Answer: A Premise: Celebrating Trajan's campaigns against the Dacians in what is now Romania, the minutely detailed friezes spiraling around the column constitute a veritable textbook of Roman warfare utilizing some 2,500 figures. Hypothesis: The friezes were added to the column four centuries ago. A. Entailment B. Neutral C. Contradiction Answer: B Premise: And fell. Hypothesis: Did not fall. A. Entailment B. Neutral C. Contradiction Answer: C Premise: On the radio, they are still talking about the Brooklyn Museum's controversial art exhibit. Hypothesis: The Brooklyn Museum has an exhibit that is making people talk. A. Entailment B. Neutral C. Contradiction Answer: A Premise: and for i don't know it must have been two or three weeks there they were doing this expanded nightly news Hypothesis: The nightly news wanted to provide more extensive coverage of what was happening. A. Entailment B. Neutral C. Contradiction Answer: B Premise: The sheets are creamy white and the tissue lining in the envelope a bluer white. Hypothesis: The sheets and the lining are black. A. Entailment B. Neutral C. Contradiction Answer: C
Mixed	Premise: A man with a beard skateboarding and a boy with a blue and black backpack riding a green bike in the background. Hypothesis: There is a man and a boy outside. A. Entailment B. Neutral C. Contradiction Answer: A Premise: There is a room full of pictures all in the wall and a woman in a coat is looking back over her shoulders strangely. Hypothesis: the room is large A. Entailment B. Neutral C. Contradiction Answer: B Premise: A Black woman on the street is talking on her cellphone. Hypothesis: A black woman is on the payphone, ordering pizza A. Entailment B. Neutral C. Contradiction Answer: C Premise: This was enlarged by Herod, sacked in a.d. 70, and totally flattened by Emperor Hadrian in a.d. 135. Hypothesis: 65 year passed between being sacked and flattened. A. Entailment B. Neutral C. Contradiction Answer: A Premise: it's slow it's uh there are many better machines on the market right now for Hypothesis: This machine is too old, that's why it's slow. A. Entailment B. Neutral C. Contradiction Answer: B Premise: well i think it made parts of it a lot easier i i is this your first that you're having Hypothesis: I think it did not help at all. A. Entailment B. Neutral C. Contradiction Answer: C

Table 11: Few-shot examples used for LLM evaluation, containing, when datasets are mentioned separately, random examples exclusively drawn from SNLI-dev for SNLI variants or MNLI-train for MNLI variants, with the mixed row contains 3 examples from each SNLI-dev and MNLI-train.

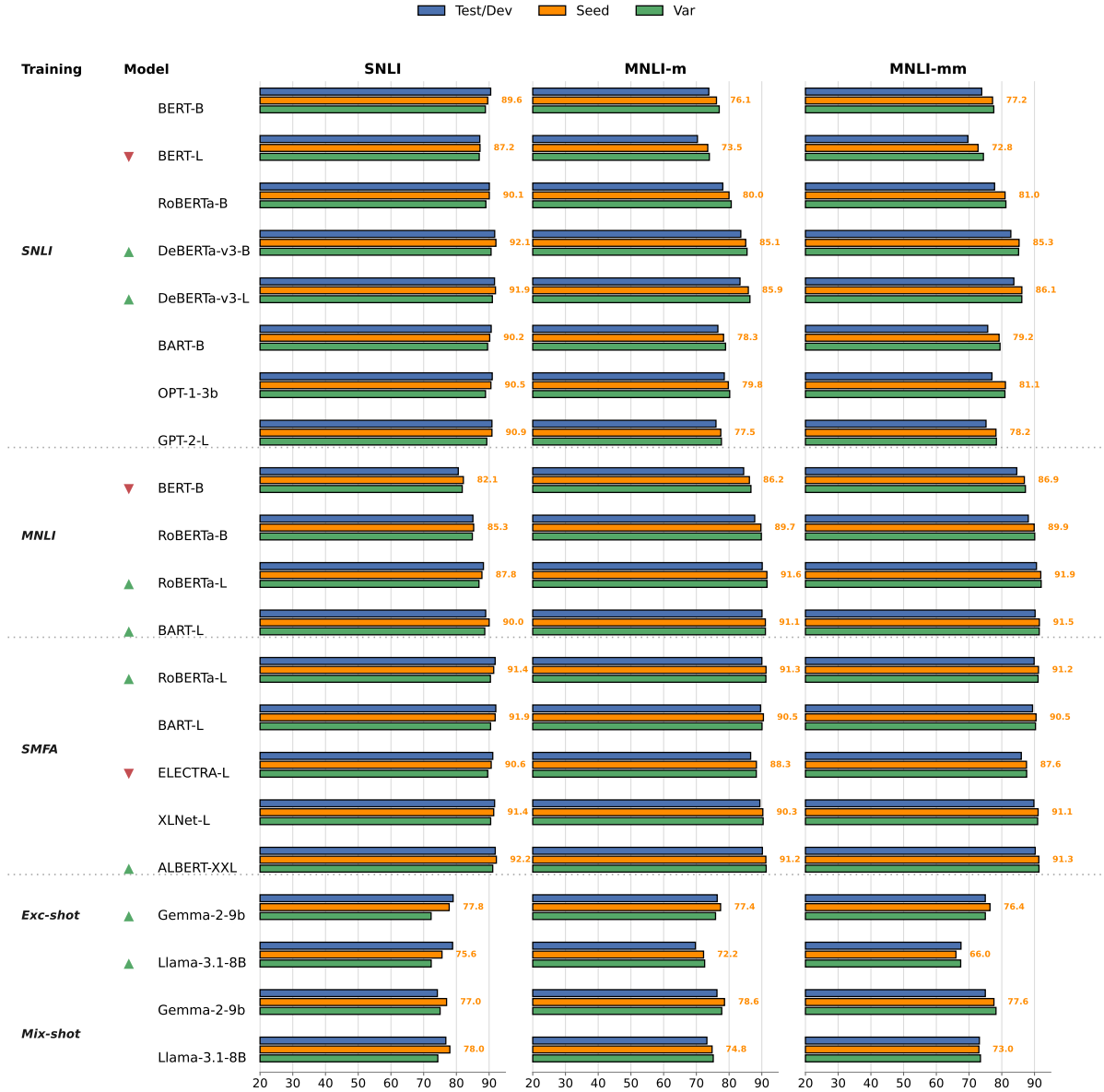


Figure 12: Results of all models, grouped by their training either fine-tuned on SNLI, MNLI, SMFA, or LLMs prompted with examples exclusively chosen from SNLI-dev (for SNLI-var) or MNLI-train (for MNLI-m/-mm). Test/Dev scores are S-acc on the test problems for SNLI and dev problems for MNLI, while Seed and Var are S-acc scores on seed and variant problems. The green triangles in front of the models in each training category show the two best models, while the red ones show the worst.

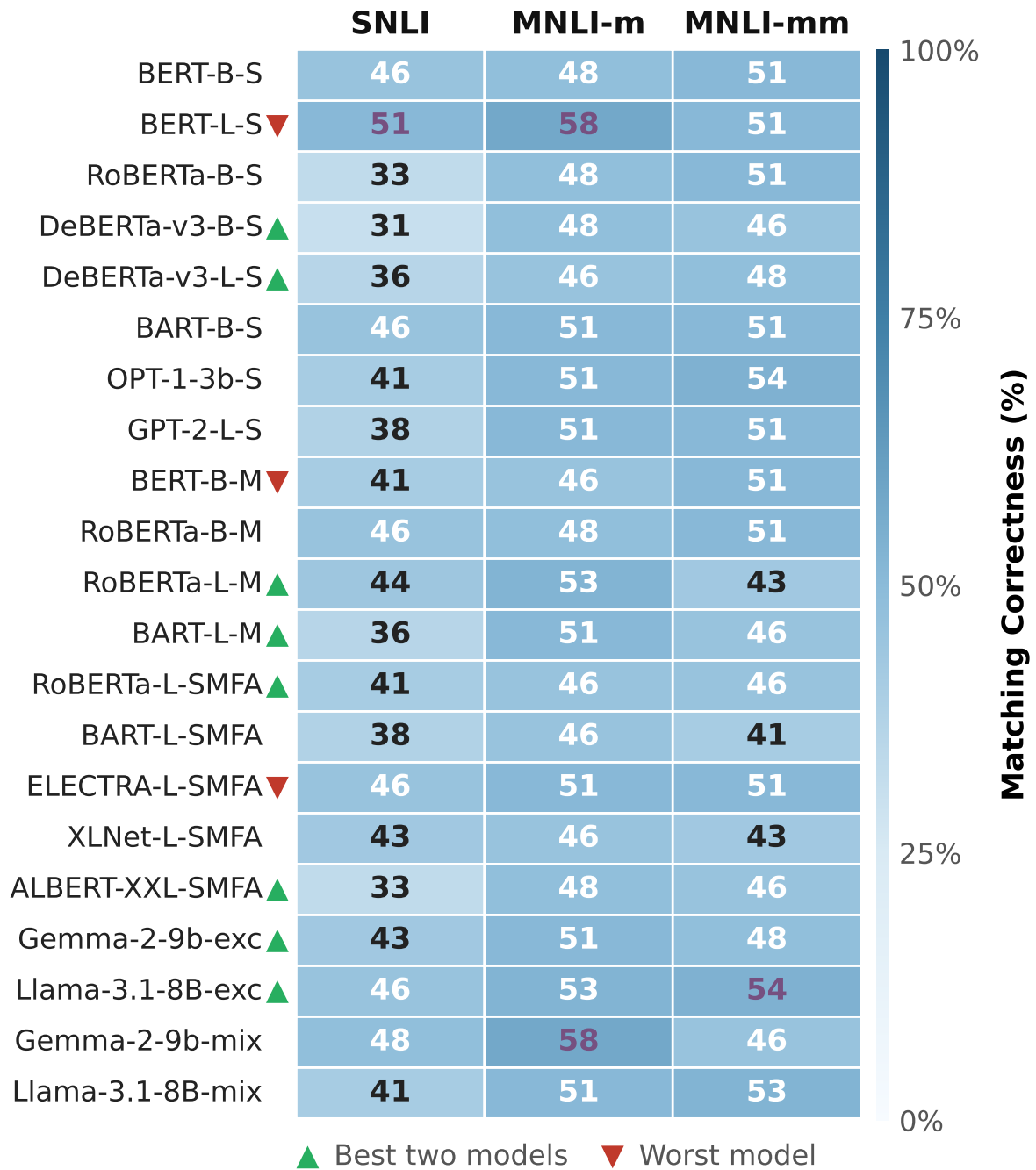


Figure 13: MCs across all models and datasets: the numbers show on which CT models get the closest SV-acc on variants to their initial S-acc on seed problems. As most numbers are $\approx 50\%$, they indicate that models do not generalize to more variants than that. Purple numbers mark the highest MCs per dataset, while black ones indicate MCs under 45.

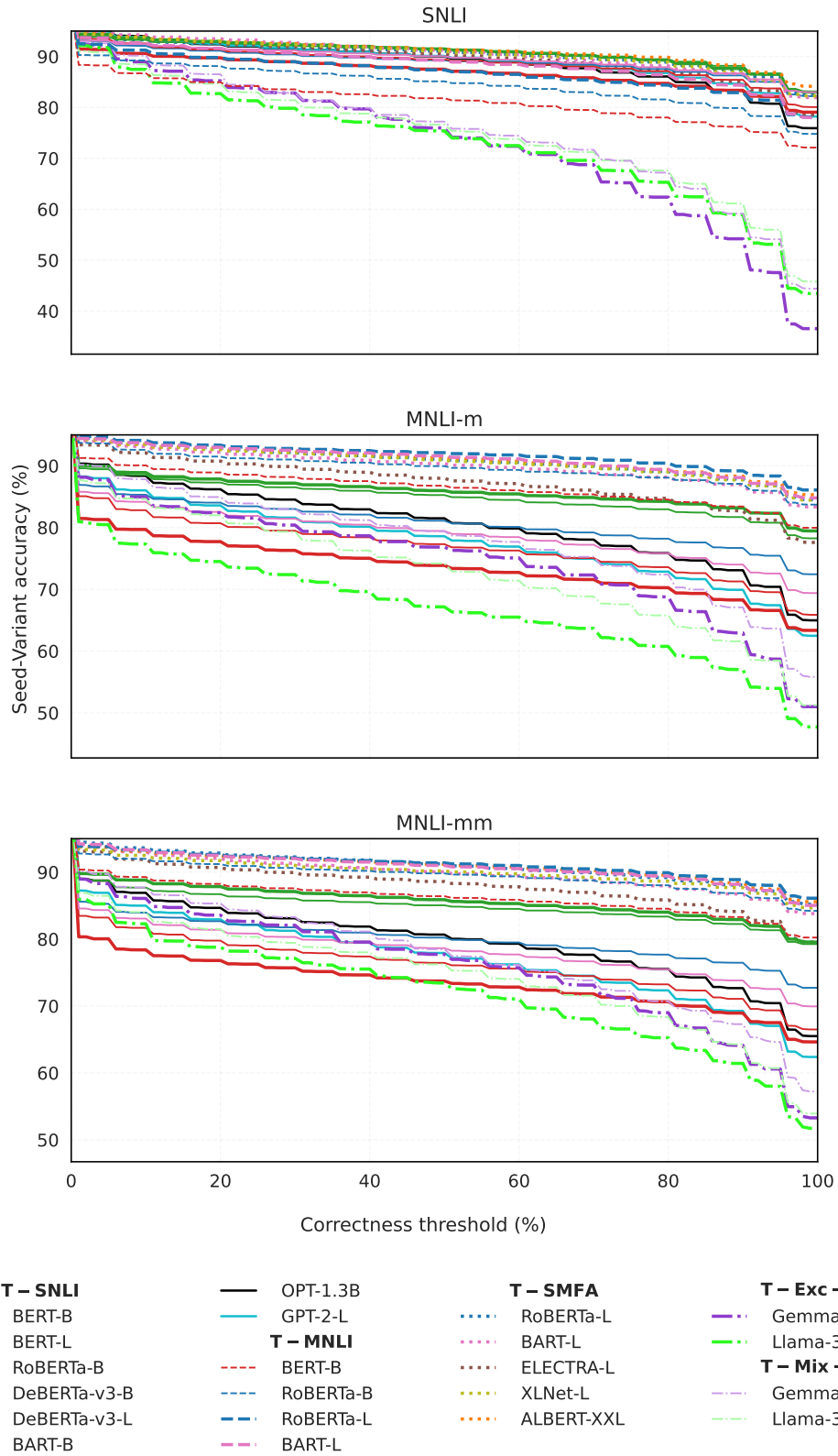


Figure 14: SV-acc curves on variants from SNLI and MNLI-m/-mm. The legend shows NLI models and the datasets they were fine-tuned on, i.e., SNLI, MNLI, or a combination of NLI datasets e.g., SMFA, or the type of few-shot prompts we used for LLMs, i.e., exclusively taken from SNLI-dev (for SNLI variants) or MNLI-train (for MNLI-m/-mm variants) or a mix of both MNLI and SNLI.

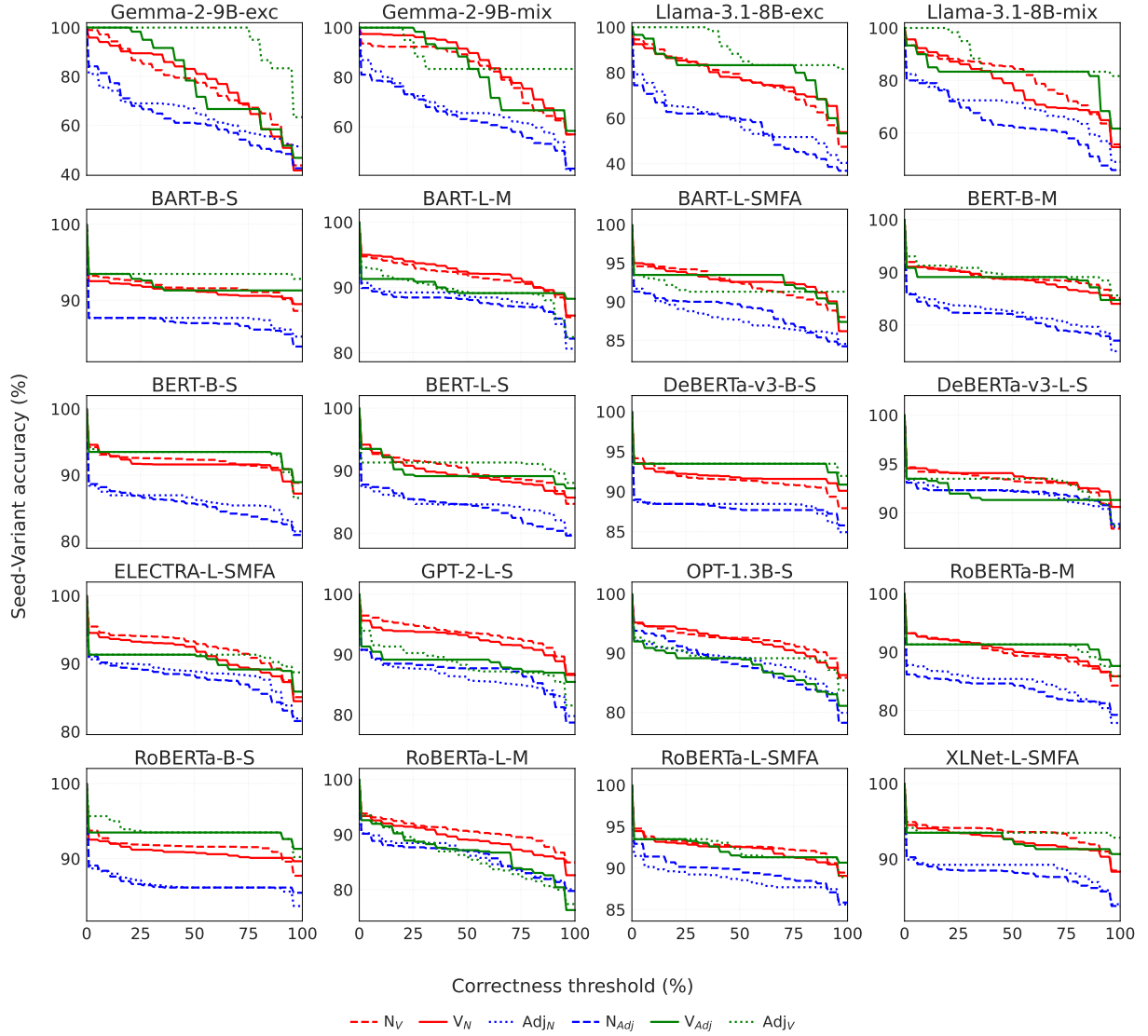


Figure 15: SV-acc for SNLI seed problems with at least two out of the three replaced open classes in each seed ($N_V=202$; $N_{Adj}=130$; $V_{Adj}=46$). The legend shows, by the larger letter of each dataset, which open class was replaced in seed problems that shared at least two.

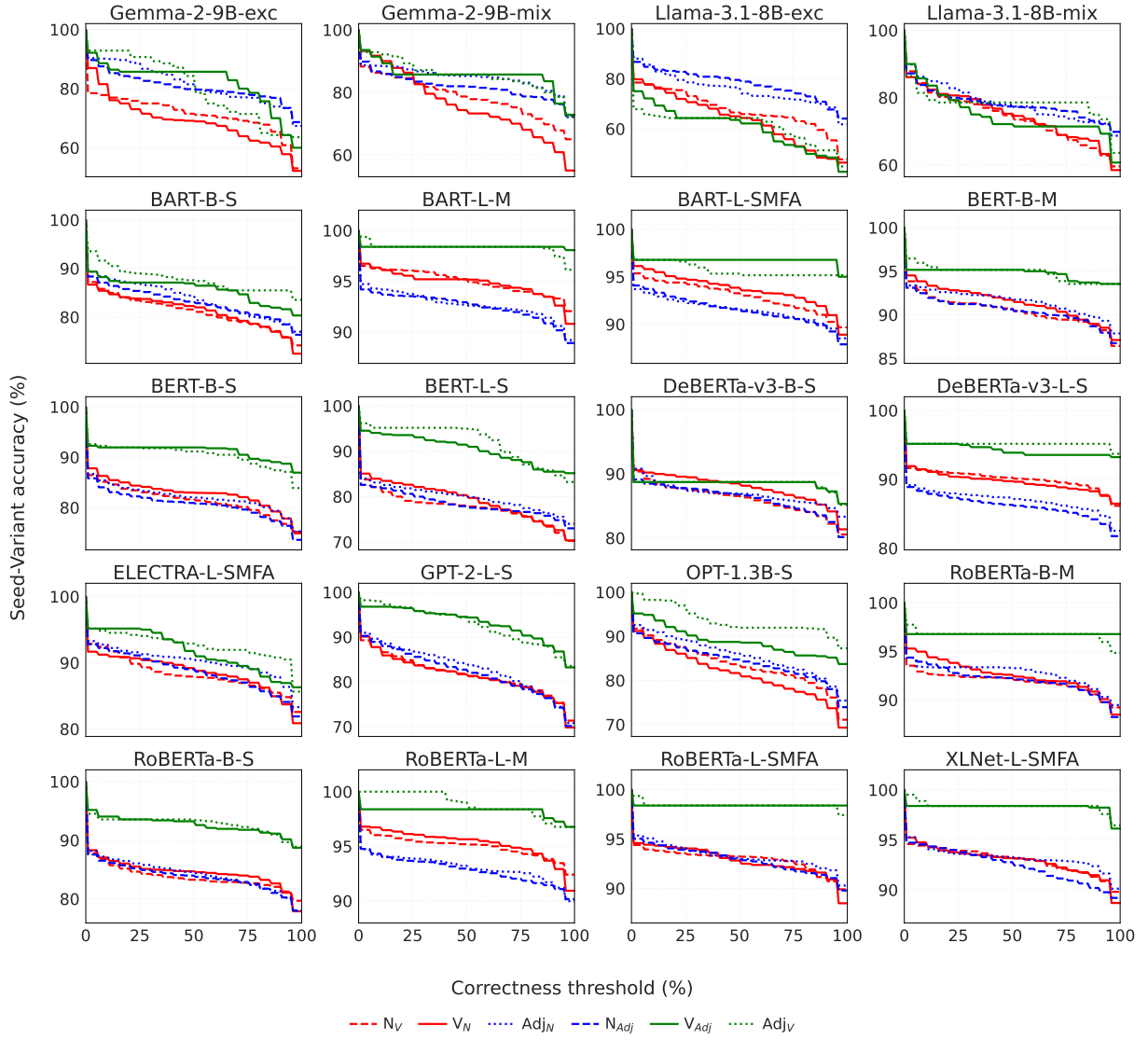


Figure 16: SV-acc for MNLI-mm seed problems with at least two out of the three replaced open classes in the seed problems ($N_V=516$; $N_{Adj}=360$; $V_{Adj}=88$).

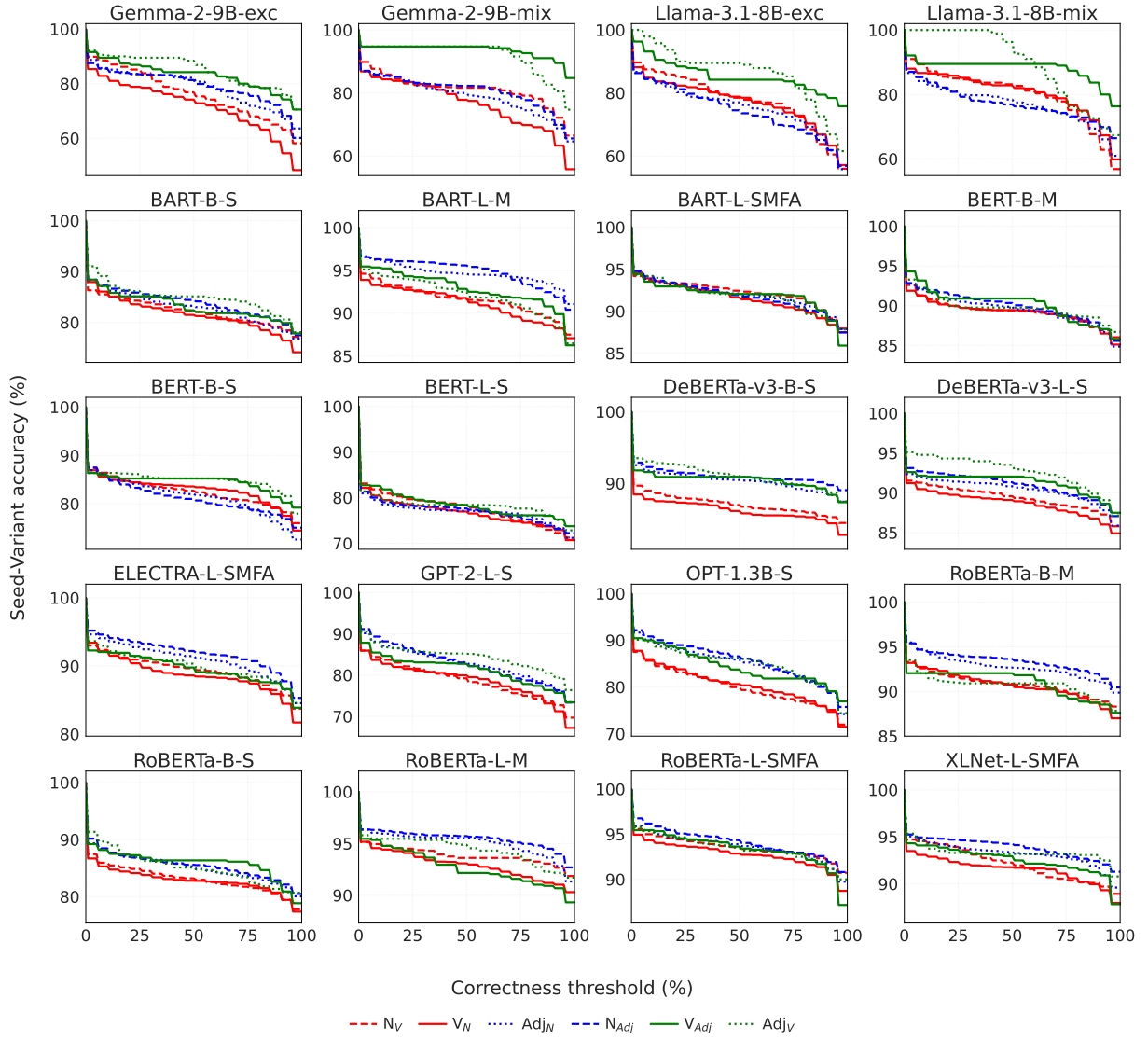


Figure 17: SV-acc for MNLI-m seed problems that have at least two out of the three replaced open classes in them ($N_V=422$; $N_{Adj}=312$; $V_{Adj}=62$).

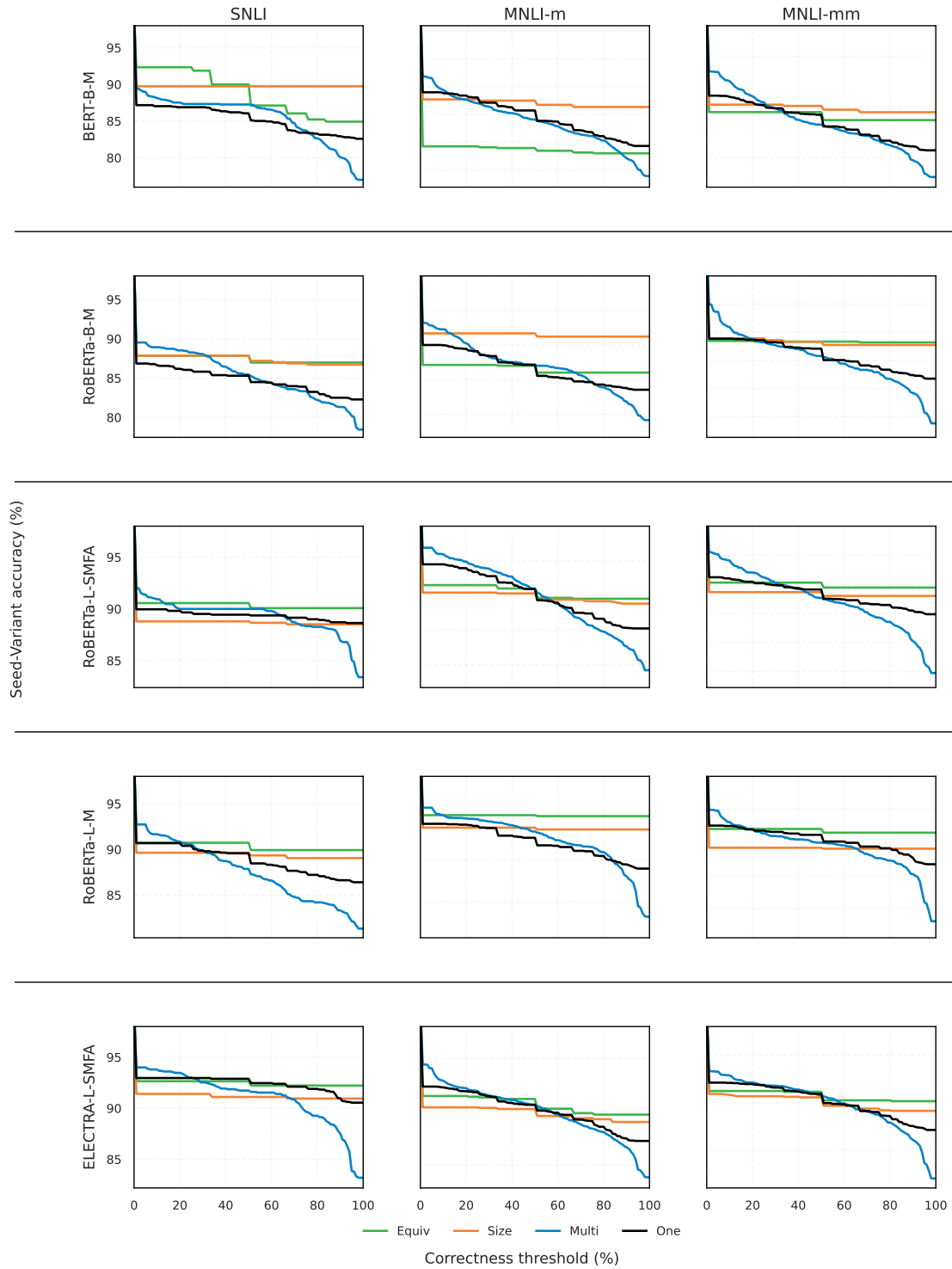


Figure 18: SV-acc curves across datasets for variants grouped by the origin model that suggested them (SNLI=150, MNLI-m=530, MNLI-mm=424). We only show the models fine-tuned on SMFA and MNLI, as SNLI fine-tuned models had lower scores and were less informative.

SNLI

P: A man pushing a hand-truck of boxes is bending over to pick up a pear.	
H: A happy man is picking up a pear.	
Gold label: E Correct label: N Annotations: E(3) N(2) C(0) Avg rate: 0.00	
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: 0.00	
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: 0.00	
P: A wet child stands in chest deep ocean water.	
H: The child s playing on the beach.	
Gold label: E Correct label: N Annotations: E(3) N(1) C(1) Avg rate: 0.00	
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: 0.00	
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: 0.00	
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: 0.00	
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: 0.00	
P: people standing at a beach with Cameras.	
H: A group of people standing at a beach filled with cameras.	
Gold label: N Correct label: N Annotations: E(2) N(3) C(0) Avg rate: 0.00	
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: 0.00	
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: 0.00	
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: 0.00	
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: 0.00	
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: N Annotations: E(3) N(0) C(2) Avg rate: 0.00	
P: A group of people are in a rowboat in the ocean surrounded by seagulls.	
H: A bunch of people are in a wooden object on the water.	
Gold label: N Correct label: N Annotations: E(2) N(3) C(0) Avg rate: 0.00	
P: A man in a hard hat looks intimidated.	
H: He is working in a potentially dangerous field that requires a hard hat.	
Gold label: E Correct label: N Annotations: E(4) N(1) C(0) Avg rate: 0.00	

MNLI-m

P: Robust came in third among words and phrases submitted (220 citations in the CR), and unlike the previous two, it seems to be a genuinely new cliché; at any rate, Chatterbox hadn't previously been aware of its overuse.	
H: Robust is a legitimately new cliché unlike its predecessors.	
Gold label: N Correct label: E Annotations: E(2) N(3) C(0) Avg rate: 0.00	
P: 3) The gap between the productivity of women and the productivity of men.	
H: The gap of genders.	
Gold label: N Correct label: N Annotations: E(2) N(3) C(0) Avg rate: 0.00	
P: A 1994 Roper Poll concluded that the NewsHour is perceived by the public as the most credible newscast in the country.	
H: A 1984 Poll concluded NewsHour is seen as the most credible newscast by the public.	
Gold label: E Correct label: E Annotations: E(3) N(0) C(2) Avg rate: 0.00	
P: Once the pious devotions are over, however, wine flows, fireworks explode, espetada (kebab) stalls flourish, and Monte regains normality for another 363 days.	
H: Monte is a normal location, outside of a period of pious devotion.	
Gold label: N Correct label: E Annotations: E(2) N(3) C(0) Avg rate: 0.00	
P: Christ on a crutch, what does he have to do to lose your support, stab David Geffen with a kitchen knife?	
H: Your support is unwavering.	
Gold label: E Correct label: E Annotations: E(3) N(1) C(1) Avg rate: 0.00	

MNLi-mm

P: Hani Hanjour, assigned to seat 1B (first class), soon followed.
H: Bob was assigned to seat 2a
Gold label: N Correct label: E Annotations: E(0) N(4) C(1) Avg rate: 0.00
P: For example, even at age 2, they waited patiently to open a small gift until a guest had departed' proper etiquette in Chinese culture.
H: They waited to open a small gift when they were very small because they'd been taught to do so.
Gold label: E Correct label: E Annotations: E(3) N(2) C(0) Avg rate: 0.00
P: . . . You got a conflict on that direction?
H: No one brought up a varying opinion on that direction.
Gold label: N Correct label: N Annotations: E(0) N(3) C(2) Avg rate: 0.00
P: This information infrastructure proved most beneficial when it gave Penney's major vendors access to sales data via direct broadcast satellite.
H: This infrastructure of information was used in order to provide customers with information regarding sales.
Gold label: C Correct label: E Annotations: E(0) N(1) C(4) Avg rate: 0.00
P: The operator must first select the work to be done, put aside the tickets that indicate she performed the sewing appropriate for those bundles and should be paid at the specified rate for the job, open the appropriate bundles, and position the pieces to be joined on the sewing table in preparation for sewing.
H: Some operators prefer to set aside the tickets first.
Gold label: C Correct label: N Annotations: E(0) N(2) C(3) Avg rate: 0.00
P: However, we recognize that contributions at the Maennerchor Society level are not possible for all.
H: Maennerchor Society level contributions are impossible for most everyone.
Gold label: C Correct label: N Annotations: E(0) N(1) C(4) Avg rate: 0.00
P: They're actually both teachers.
H: The are teachers together.
Gold label: N Correct label: N Annotations: E(2) N(3) C(0) Avg rate: 0.00
P: Even though we receive operating funds from the state, there are a myriad of additional expenses to be met, such as welding equipment for sculpture, pottery wheels for ceramics, and computers for graphics.
H: The state won't fund welding equipment, pottery wheels or computers.
Gold label: E Correct label: N Annotations: E(3) N(1) C(1) Avg rate: 0.00
P: Business units adopting both bar codes and EDI are therefore able to reduce the transaction costs for processing information about sales and orders.
H: Business units use either bar codes or EDI.
Gold label: N Correct label: N Annotations: E(0) N(4) C(1) Avg rate: 0.00
P: Did your dad read to you too?
H: Did someone read to you?
Gold label: N Correct label: E Annotations: E(1) N(4) C(0) Avg rate: 0.00
P: But he places most blame on how contemporary children are reared.
H: People put the most blame on how contemporary children are reared.
Gold label: N Correct label: N Annotations: E(1) N(4) C(0) Avg rate: 0.00
P: I forgot about comic books!
H: I did not mention comic books yet.
Gold label: N Correct label: N Annotations: E(1) N(3) C(1) Avg rate: 0.00

Table 12: The original SNLI and MNLi-m/-mm problems whose variants were most poorly classified by all NLI models. We also considered LLMs, but their sub-selected seed did not include these problems. **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 and MNLi data, where the gold label is selected based on majority voting. **Correct label** is an inference we thought was correct. We are aware of the inherent disagreements in NLI labeling (Pavlick and Kwiatkowski, 2019), but we refer to our corrected labels as the most likely labels. Note that we also further looked into the problems that were classified with < 5% average accuracy across all models (SNLI=27, MNLi-m=29, MNLi-mm= 24), out of which only 13 for SNLI, and MNLi-m, and 10 for MNLi-mm had a correct gold label.