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

Anonymous ACL submission

Abstract

With many benchmarks becoming saturated, it has been paramount to create new datasets that evaluate the generalization capacity of current state-of-the-art models in reasoning. However, designing new quality reasoning datasets is challenging, as their manual construction is costly, while their automatic generation is unreliable or often leads to synthetic data with limited scope. In this paper, we propose the Minimal Expression-Replacement GEneralization (MERGE) test that evaluates the robustness of reasoning models against non-adversarial variants of existing evaluation datasets. We automatically obtain high-quality variants from the original instances with Minimal Expression REplacement (MERE) generation that utilizes Masked Language Models (MLMs) and safeguarding filters. We apply the MERGE test to Natural Language Inference (NLI), a popular task of reasoning. We generate new NLI datasets from two existing common ones with the MERE generation and use them to evaluate multiple strong NLI models. The results indicate that both LLMs and fine-tuned NLI models generalize poorly: they struggle to consistently and correctly classify variants that differ only minimally from the original ones, both at the surface level and in terms of reasoning. Further, we also analyze how certain aspects in variant generation, such as the word class and the source MLMs, affect model performance.

1 Introduction

The challenge in the Natural Language Inference (NLI) task is to predict the inference relation between premise p and hypothesis h. Models' high performance on NLI test sets is usually due to the partial exploitation of heuristics learned from the training set (Gardner et al., 2020b; 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; Hupkes et al., 2023; Yang 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 (Hupkes et al., 2023), models generalize poorly (Nie et al., 2019b; Verma et al., 2023), unlike in traditional in-distribution test sets of SNLI (Bowman et al., 2015) or MNLI (Williams et al., 2018).

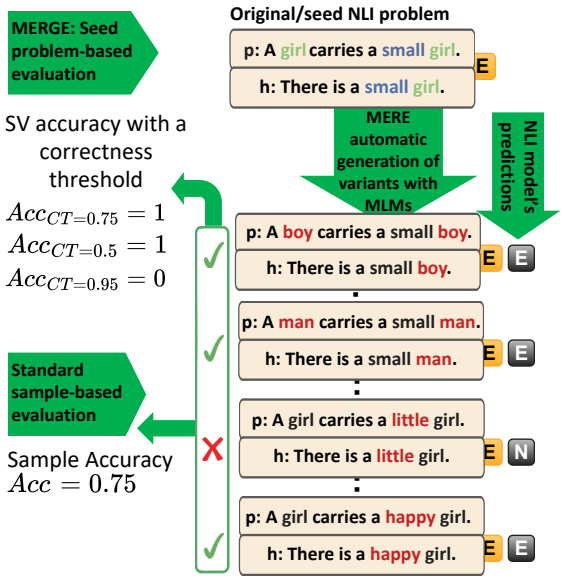

Figure 1: First, the MERE generation creates multiple label-preserving variants from a seed problem. Then MERGE evaluates a model on the seed based on its performance on the variants. Unlike the standard sample-based accuracy metric, MERGE uses a Seed-Variant (SV) accuracy with a correctness threshold,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

OOD NLI datasets are usually challenging when they don't retain common reasoning heuristics, such as (inverse) word overlap (Rajaee et al., 2022) and hypothesis-only artifacts (Gururangan et al., 2018; Poliak et al., 2018; Tsuchiya, 2018) or consist of adversarial problems that try to go against the training data distribution by resembling certain training samples but having different labels (Glockner et al. (2018); Naik et al. (2018); Gardner et al.

(2020a), among others). This prompts the question of whether a generalization challenge set can be designed that doesn't deliberately break common heuristics or rely on adversarial strategies. The construction of such a dataset would open new directions for generalization evaluation and reveal new types of weaknesses in current models.

We propose a method of automatically generating evaluation datasets from existing ones by creating variants of original 'seed' problems through Minimal Expression REplacements, referred to as the MERE generation. Figure 1 shows MERE in action for an NLI problem: to form variants of the seed problem (p, h), MLMs are used to suggest contextually probable replacements for the open-class words shared between p and h. The replacements are such that the obtained variants maintain the underlying reasoning of the seeds, as demonstrated in Figure 1. Moreover, the variants also maintain the seed properties, such as the input/sentence length and the word overlap size, which can be exploited by models as heuristics.

We evaluate NLI models on the generated variants with respect to correctness and consistency. Specifically, we adopt the pattern-based accuracy metric from Abzianidze et al. (2023), hereafter called Seed-Variant accuracy, and evaluated multiple models on Minimal Expression-Replacement GEneralization, shortly MERGE. In a nutshell, a model is evaluated on seed problems based on the correctness and consistency it obtains on the corresponding variants, as shown in Figure 1.

We aim to answer the following questions in the context of NLI: (i) How reliable is our methodology for automatically obtaining variants? (ii) How well NLI & LLM models generalize to minimal variants? (iii) To what extent do factors like the word class or the source MLM influence model performance?

Our contributions are as follows:

1. A fully automated methodology to create a *friendly* test set for generalization, applicable to many reasoning tasks.
2. The results revealing the poor generalization of multiple models against minimal changes.
3. The findings that verbs, followed by nouns and adjectives, are harder to generalize to and that a source MLM does not affect an NLI model's performance.

We review previous works in §2, present our methodology in §3, experiments in 4, and results in §5. The conclusions are in §6.

2 Related Work

Evaluations on NLI OOD datasets (or *contrasts sets* in Li et al., 2020) suggest that models severely lack generalization abilities, with 14–30% decreased performance (Kaushik et al., 2020; Petrov, 2025; Glockner et al., 2018, among others), and inconsistent predictions in 10–16% of variants (Verma et al., 2023; Arakelyan et al., 2024). We review these datasets here, along with several aspects, and classify them in Table 1.

Modification Type Previous studies *replaced* words, *decomposed* or *paraphrased* problems, or used a combination of *multiple* operations for constructing variants, shown in Column *Type*.

Modification Procedure Alterations were made *automatically*, *manually*, or by using a *mix* of automatic and manual methods, shown in Column *Procedure*.

Validation Variants were validated by i) manual validation – focused on a *partial* (HVal_p), a *full* set of variants, or a combination of both (HVal_{pf}); (ii) a *mix* of automatic and manual validation methods (Mix.); or iii) not validated at all (*N/A*), shown in Column *Validation*.

Meaning, Reasoning, Word Overlap and Syntax The meaning (M), underlying reasoning (R), syntax (s), or word overlap (WO) of $p-h$ [1] were *preserved* (*Y*), *changed* (*N*), or a *mix* (Mix.) of both, shown in Columns *M*, *R*, *S* and *WO*.

Modified Unit The modifications can be applied to the *update*[2] (U), p, h, both of them separately (p/h) and together ($p\&h$), shown under Column *Unit*.

Evaluation Variant predictions are evaluated by comparing them (i) individually — a) with the *gold label* (*V-G*), b) the prediction of the original NLI problem (*V-O*); (ii) as a group – a) with the prediction on the original NLI problem (*Vs-O*), b) the gold label (*Vs-G*), or c) with each other (*Vs-Vs*). See column *Evaluation*.

Shortcomings of previous variant datasets Performance drops in previous NLI variant datasets cannot be attributed fully to poor generalization, as

[1] Note that word overlap preservation requires modifying $p\&h$ at the same time.

[2] Some NLI datasets evaluate how new information, i.e., updates, might change the entailment label between $p - h$.

Study	Type	Procedure	Validation	Unit	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: OOD NLI datasets classified considering: the **type** of modifications deployed to obtain variants; the automatic or manual **procedure**; whether they were validated and by whom; the **unit** modified; the preservation of **m**eaning, **r**easoning, **s**yntax or **w**ord **o**verlap of the original problem, and the **datasets** which were used to obtain them. Section 2 details the meaning of each classified aspect. * in the **Study** column indicates the variants were used for fine-tuning models. All studies use labels: entailment, contradiction, neutral.

they were constructed with non-syntax-preserving constructions that might be more challenging for models (Li et al., 2020), or changed lexical overlap of $p - h$ due to paraphrased or decomposed NLI problems. $p - h$ word overlap preservation has been partially explored previously in Glockner et al. (2018), but their replacements were reasoning non-preserving and were favoring variants' plausibility at the expense of lexical diversity. Other shortcomings of previous variants' datasets concern their automatic generation, previously criticized for potentially (i) biasing variants in favor of models deployed (Li et al., 2020; Gardner et al., 2020b); and (ii) constructing implausible variants (Dutt et al., 2024). However, manual variant creation is very time-consuming, and unlikely to be faster than the learning rate of new models, making a reliable automatic generation methodology necessary.

MERGE In contrast, we propose a method that automatically creates plausible variants by replacing shared words of $p - h$ 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). Additionally, in our method, lexical diversity is not fixed but can also be increased, for instance, by considering more replacements from more MLMs, unlike in previous list-restricted replacement studies (Glockner et al., 2018).

3 Methodology

The MERE generation uses an original NLI problem $\langle p, h, l\rangle$, labeled with l, as a seed to obtain variant label-preserving problems $\langle p_i, h_i, l\rangle$ from it. This is done by replacing each open-class word o, shared between p and h, with new replacement words. The replacements for $\langle p, h\rangle$ are collected from a set of MLMs $\mathcal{M} = \{M_1, \ldots, M_n\}$, starting at the sentence level and extending to the problem level.

We define $R_j(S, i^c_>)$, as a set of r replacements, for a sentence $S = (s_1, \ldots, s_k)$, an MLM M_j, and an open-class word $o = s_i$ as:
(i) r is more probable than so under M_j, in the masked context $S[s_i/\text{MASK}]$, marked with $_>$;
(ii) r doesn't occur in S;
(iii) r and o are of the same word class in S at position i, marked with ^{c}.[3]

With constraints (i) and (iii), r is expected to be more felicitous than o in S and to preserve the syntactic structure of S. Constraint (ii) avoids label non-preserving changes, as substituting words like 'boy' with 'poodle' in a problem with p 'two poodles and a boy swim' and h 'only one boy swims' from entailment to contradiction.[4]

The replacement set from M_j for a word o, with multiple occurrences in S, is defined as:

$$R_j(S, o^c_>) = \bigcap_{o=s_i} R_j(S, i^c_>)$$

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

$$R_{\mathcal{M}}(S, o^c_>) = \bigcup_{M_j \in \mathcal{M}} R_j(S, o^c_>)$$

where replacements are validated by the same MLM at each occurrence position of o in S.

[3]The 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.

[4]With having (ii), variants with incorrect inference labels remain possible, although improbable, e.g., the replacement 'dog' results in an incorrect label.

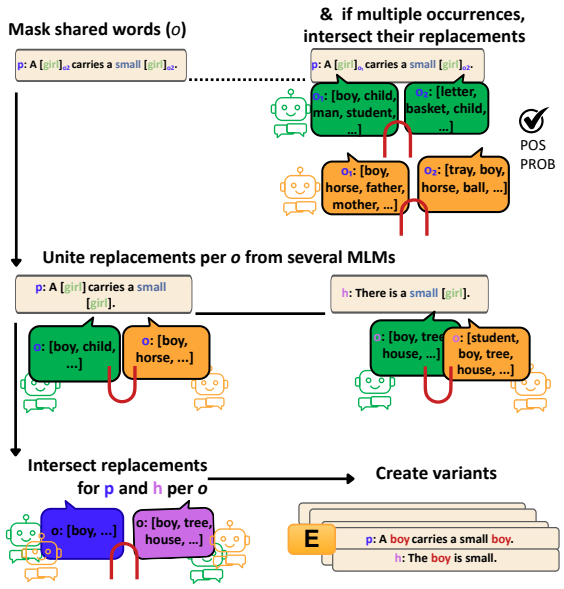

Figure 2: Generating NLI problem variants with MLMs. Checkmarks for POS and PROB indicate that the shown replacements have already been excluded if they had different classes or lower probability than o, or were already part of the problem ('girl').

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:

$$R_{\mathcal{M}}(\langle p, h\rangle, w^c_{>}) = \bigcap_{S \in \{p,h\}} R_{\mathcal{M}}(S, o^c_{>}) \quad (1)$$

and 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 $r_{ij} \in R_{\mathcal{M}}(\langle p, h\rangle, o^c_{i>})$:

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

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 We used seed problems from SNLI test, and MNLI dev-m/-mm, the largest available NLI datasets, which shared nouns, verbs, adverbs, and adjectives across $p - h$, for generating 200 replacements r with 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). Most model sizes were base and large, except for ALBERT (base and xxl), see Appendix Table 4 for a full description.

Quality filtering Replacements were kept if they were: i) same class as o, excluding subwords and punctuation as well; ii) higher probability than o, iii) different from words already in $p - h$, o included; iv) validated by at least one MLM. Most filtering occurred due to (ii), with only 10–30% replacements having equal or higher probability than o, see Appendix Table 5. Problems with fewer than 20[5] replacements across all o_1 after filtering were also excluded, each eligible seed problem finally yielding at least 20 potential variants.

Manual Quality Validation Variants were then annotated considering their F) replacement's **f**luency; and R) preservation of the original infe**r**ence by two (for SNLI), and one author (for MNLI) which assigned a score from 1–5 (1–poor; 5–good) for F, and R with good variants with F+ R ≥ 9. Note that we imposed strict annotation guidelines, since any ungrammaticalities potentially affecting R were penalized, regardless of whether the original sentence was ungrammatical, and as replacements improving variants received no additional positive scores. See the full validation guidelines in Appendix subsection A.3, alongside other specific details such as annotator agreement, or diversity of replacements.

The validation process had two stages. First, we did a fine-grained manual validation of the replacements of 100 random sampled problems per open-class[6] and dataset. Based on the distribution of good and bad examples and their source models shown in the Appendix Figure 9, BART replacements were excluded given their higher contribution to bad variants. Second, we annotated 100 cross-class problems per dataset to check the efficiency of the first stage. The final check revealed that all variants for MNLI-m, 98% for MNLI-mm, and 91% for SNLI had a good F+R score, plotted in Figure 10 of the Appendix.

Final Variant Dataset For each dataset, we randomly sampled 10 times only 20 variants per class per seed problem, forming the $_{\text{Var}}$ dataset. The repeated subsampling and maximum limit of variants

[5] More interpretable, e.g. 95% correct variants of 20 is 19 vs. 9.5 out of 10).

[6] We eventually only replaced nouns, verbs, and adjectives, as few seed problems shared adverbs between $p - h$.

Dataset	Class	Seed	N (%)	C (%)	E(%)	Uni
SNLI	N_{Var}	1808	29	20	50	67
	V_{Var}	506	30	18	50	66
	A_{Var}	259	32	22	44	63
	$_{Var}$	**2222**	**30**	**20**	**49**	**77**
MNLI-m	N_{Var}	3058	25	31	43	86
	V_{Var}	648	22	27	50	69
	A_{Var}	703	26	23	49	72
	$_{Var}$	**3663**	**25**	**30**	**44**	**95**
MNLI-mm	N_{Var}	3452	25	31	43	89
	V_{Var}	683	21	27	51	68
	A_{Var}	779	23	28	48	73
	$_{Var}$	**4016**	**24**	**30**	**45**	**103**

Table 2: Number of eligible **seed** problems per class and dataset for variant generation after filtering, alongside their label percentages (**N**eutral, **C**ontradiction and **E**ntailment), and the number of **Uni**que variants across all 10 random subsamples. Seeds with multiple replaced words of different classes, like the noun 'girl' and adjective 'small' in Figure 2, are counted only once, resulting in non-summative total seed counts.

per seed prevent the final dataset from being overpopulated with the variants of very productive seed problems and from having randomization artifacts. Table 2 shows the seeds of each $_{Var}$, alongside their label distributions.

Evaluation Metrics We evaluate models with two metrics: their standard accuracy on individual seed problems, called seed accuracy (S), and their Seed-Variant accuracy (SV), adapted from Abzianidze et al. (2023). Under SV, a problem is considered correctly labeled only if a minimum correctness threshold (CT) of its variants are also correctly labeled, see Figure 1. In the experiments, we report SV with CT that is averaged across the 10 subsamples of $_{Var}$. If models generalize well, their SV-acc scores should not substantially fall behind their S-acc. To quantify this, we define the matching correctness threshold (MC): the CT required for models to match S scores on SV.

Models We evaluated several NLI fine-tuned models: BERT, RoBERTa, DeBERTa (He et al., 2021), BART, ALBERT, Electra, XLNet (Yang et al., 2019), OPT (Zhang et al., 2022), and GPT-2 (Radford et al., 2019), fine-tuned on SNLI, MNLI, or a combination of SNLI, MNLI, FEVER Nie et al. (2019a), and ANLI–SMFA; and two LLMs: Gemma-2-9b (Team, 2024), and Llama-3.1-8B (Dubey et al., 2024), to observe how NLI fine-tuning affects performance. Full references to the models are in Appendix Table 7.

Evaluation of LLMs For each dataset, LLMs models were evaluated on a subset of 500 randomly chosen seed problems and their variants, which we ensured were preserving the overall Seed distribution and similarity in S scores to the overall dataset. NLI is also formulated as a multiple-choice task where the logit values of the models are used to predict their choice, following the prompting strategy from Madaan et al. (2024). Since five examples were shown to be sufficient for near-maximal performance in the aforementioned study, we evaluated models using six examples exclusively sampled from the same datasets as those of the variants (SNLI-dev or MNLI-train), or with a 3+3 mix from both datasets. Additional details about seed subsampling, prompting, including the few-shot examples, are shown in subsection A.4 of the Appendix.

5 Results

5.1 Do models generalize to variants?

Figure 3 shows results of the two best and one worst model from each training category (SNLI, MNLI, SMFA, or pre-trained models). These are selected by their averaged S test scores for SNLI and MNLI, while the results of all models are shown in Appendix Figure 16. Figure 3 highlights that S scores of SNLI-Test (blue bars) and SNLI-Seed problems (orange bars) are similar, with lower test than seed scores in the case of MNLI-m/-mm. Note that S Test and Seed scores for pre-trained models are generally lower than those of all NLI models by even 20%, as also shown previously by (Madaan et al., 2025). We compared the drawn seed problems with the test set and 100 random subsamples of test problems matched in size with each seed dataset. On average across models, ≈55% of random SNLI subsamples had higher S-acc scores than SNLI seeds. Contrastively, no MNLI-m, or only 0.1% MNLI-mm random subsamples did, indicating MNLI seed problems are consistently easier than MNLI test problems.[7].

S scores on variants (green bars) are close to the S seed scores in Figure 3, suggesting models generalize well when we consider variants individually. However, the results show poor generalization when correctness and consistency across variants are enforced with SV-acc. Figure 5 shows that most models match their S-acc scores when the SV

[7]Consequently, MNLI seeds subsampled for pre-trained models are also easier, given they were sampled to match the main seed dataset distribution

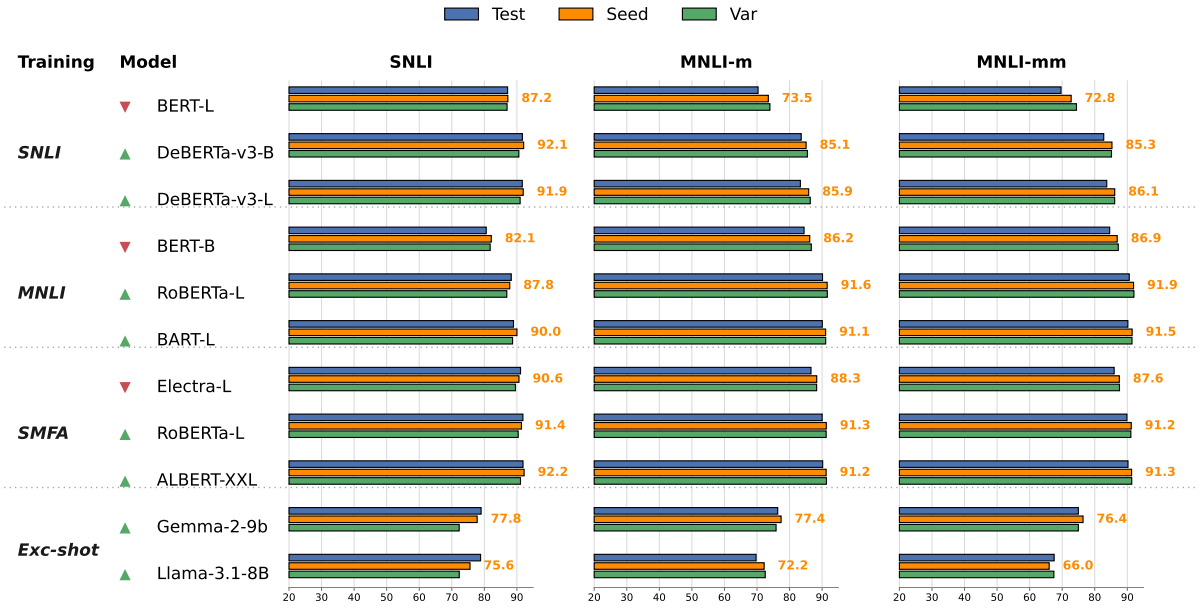

Figure 3: Results of the two best and the worst models trained either on SNLI, MNLI, SMFA, or prompted with examples exclusively chosen from SNLI-dev (for SNLI-var) or MNLI-train (for MNLI-m/-mm). **Test** scores are S scores on test problems, while **Seed** and **Var** are S scores on seed and variants.

threshold is set to at most ≈50%. Thus, generalization over only half of the variants per seed should be satisfactory to maintain the initial performance, with better generalization on SNLI than MNLI variants. Beyond the ≈60% threshold, performance drops across models and datasets, especially in higher thresholds, as further illustrated in Figure 4, with a particularly steep decline for pre-trained LLMs. Overall, Gemma-2-9b-exc is the model that generalizes best across variants from both SNLI and MNLI, i.e. ≈60%, though still showing a drastic drop in generalization, with DeBERTa-v3-B-S having the highest MC, i.e., 61% on SNLI.

Overall, our results are in line with previous studies that showed models fail to generalize to variant datasets (Verma et al., 2023; Arakelyan et al., 2024; Glockner et al., 2018, etc), with our work additionally demonstrating this on easier datasets[8] where correctness and consistency are equally enforced. Our SV scores especially show how considering variants per seed might reveal shortcomings in generalization, while individual S scores blur this effect by displaying high values, due to the models compensating failed variants with almost perfect predictions on variants of easy seeds.

[8]We evaluated the hypothesis-only baselines on the generated variants for SNLI and MNLI and observed that the hypothesis-heuristics are still substantially present in the variants, with 57% and 51%, respectively, in contrast to 33% of a random baseline.

5.2 Which classes are more difficult to generalize?

To investigate whether replacing nouns, verbs, or adjectives affects models' performance, we compared the scores of variants grouped by the open-class category that was replaced to form them. To directly compare how replacing one of the classes in a problem affects the scores of models, we only used seed problems that share at least two of the three open-classes. With such problems, like 'A girl carries a small girl', we can directly compare the effect of replacing the noun or adjective, while also controlling for the possible interfering effect of the dataset size of variants, given that seed problems are not balanced across open-classes, as shown in Table 2.

We show the model with the highest averaged S scores for SNLI and MNLI Test in Figure 6. In the legend, the larger acronym shows which class was replaced to form variants when the seed shared both classes in the acronyms. Across datasets, adjectives are easier than verbs on higher thresholds (in green lines), though they had similar scores in lower thresholds. In lower CT, verbs are easier than nouns in MNLI-m and SNLI, and more difficult than them in MNLI-mm (red lines). In high CT, verbs are more difficult than nouns across datasets. Lastly, adjectives are easier than nouns in high CT, with similar scores in low ones, as shown by the blue lines.

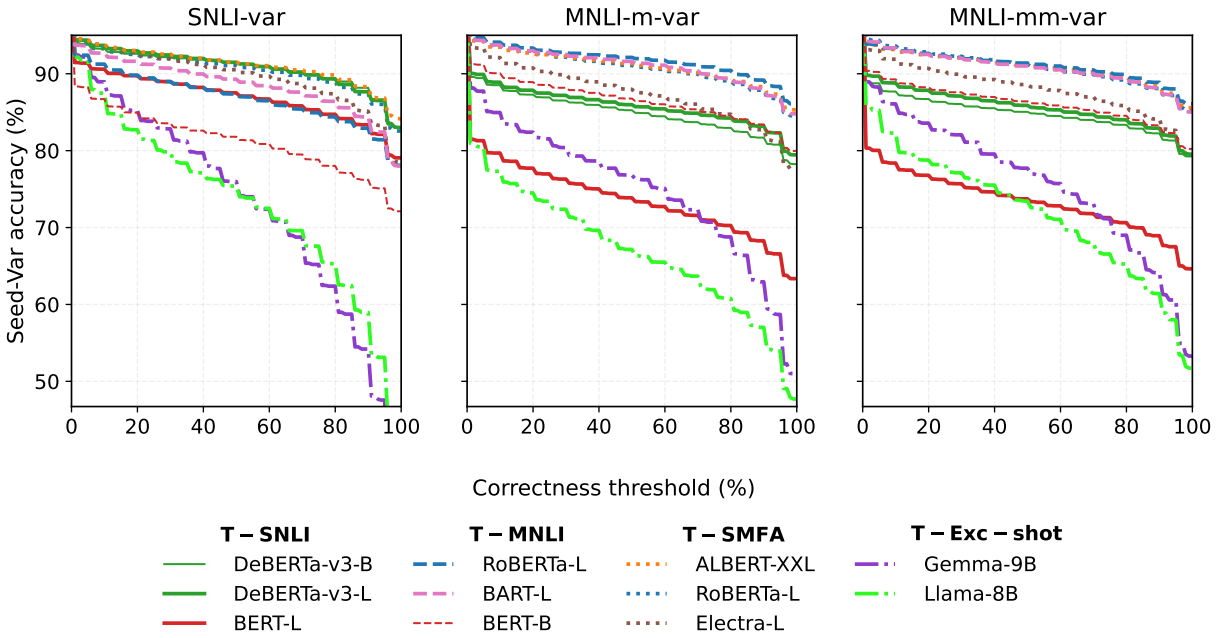

Figure 4: SV scores on variants from SNLI-var, and MNLI-m/-mm-var. The legend headers show which datasets NLI models were fine-tuned on, i.e. SNLI, MNLI or a combination of NLI datasets, e.g., SMFA, or, when applicable, that the few-shot prompts for the models were exclusively from SNLI-dev (for SNLI-var), or MNLI-train (for MNLI-m/-mm var).

5.3 Do source MLMs affect generalization?

We further looked into the scores of NLI models that had a MLM counterpart, e.g. BERT, RoBERTA, Electra and ALBERT models, to observe whether they generalize better on variants generated by their corresponding MLM, as previously it was suggested that using models for data generation might bias the data (Li et al., 2020; Gardner et al., 2020b).

We plot the SV curves of the best performing model in Figure 7, where dataset acronyms stand for variants formed with replacements from the same MLM similar in size (Equiv), a MLM similar to the NLI model but of different size (Size), e.g. BERT-B-S evaluated on BERT-L-generated variants; more than one MLMs, potentially including the evaluated model (Multi), or a single model of a different architecture (One)[9]. To ensure that we isolate the effect of the origin from the imbalanced seed problems across origins (see Appendix Table 10 for statistics), we considered only seed problems with variants across all four categories.

While scores are consitently high on Equiv variants (green lines) across datasets, the overall scores of other datasets suggest origin does not influence model predictions. If it did, we would expect Size (orange lines) to always have the second-highest scores across datasets, which is contradicted by its highest scores for SNLI and MNLI-m. Similarly, we would expect One variants (black line) to have the lowest scores, followed by Multi (blue line), since the latter includes replacements suggested by the evaluated model which could favor performance. However, the graphs show that Multi has the lowest scores in higher CT.

5.4 What 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 0.7% SNLI (16), 0.05% MNLI-m (2) and 0.29% MNLI-mm (12) problems were incorrectly classified by all NLI models[10], with 66% problems across datasets having high label variation (see the problems in Table 11), adding evidence to the well-known issue of annotation variation in NLI (Pavlick and Kwiatkowski, 2019; Weber-Genzel et al., 2024), and bigger model error on such problems (Madaan et al., 2025). Models are also better on seed problems initially correctly classified, as also shown by Ohmer et al. (2024), i.e. 97% of their variants are correctly classified, com-

[9]Note 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.

[10]These problems were not part of the sub-sample for in-context models.

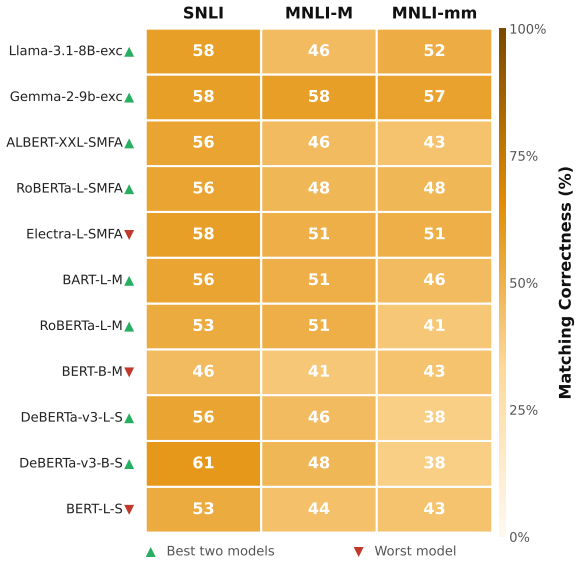

Figure 5: MC across models and datasets: the CT on which models get similar SV scores to their S ones. 100% would indicate perfect generalization, i.e. all variants can be considered with similar S scores. The last acronym in model names indicates the dataset they were fine-tuned on, or the prompting used.

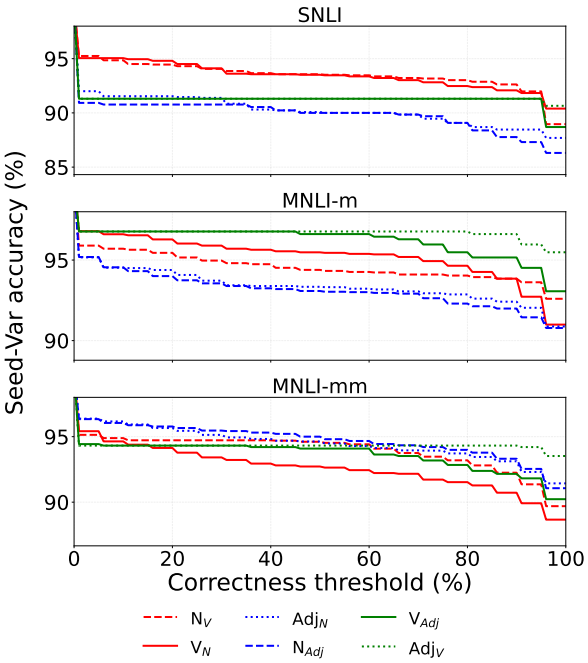

Figure 6: SV curves of ALBERT-XXL on seed problems sharing at least two out of the three open-class words for **SNLI** (NV=202; NAdj=130; VAdj=46); **MNLI-m** (NV=422; NAdj=312; VA=62); and **MNLI-mm** (NV=516; NAdj=360; VAdj=88). The first acronym shows which class was replaced to form variants.

pared to ≈ 80% for incorrect seeds. Occasionally, models correctly predicted variants of incorrect seeds (avg. across models of SNLI=98, MNLI-m=207, and MNLI-mm=210).

6 Conclusions

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

We show that models, across sizes and architectures, fail to generalize to more than around half of the variants of two datasets. More specifically, surpassing the ≈50% CT yields lower scores across models than their original seed-obtained scores. Our results also indicate that verbs are more difficult to generalize to, followed by nouns and adjectives, and that replacements from different MLMs do not influence models' scores, indicating a steady loss in generalizability sources for replacements.

Although MERGE is currently applied to NLI, we plan to extend this test to other NLU tasks (e.g. reading comprehension) in the future.

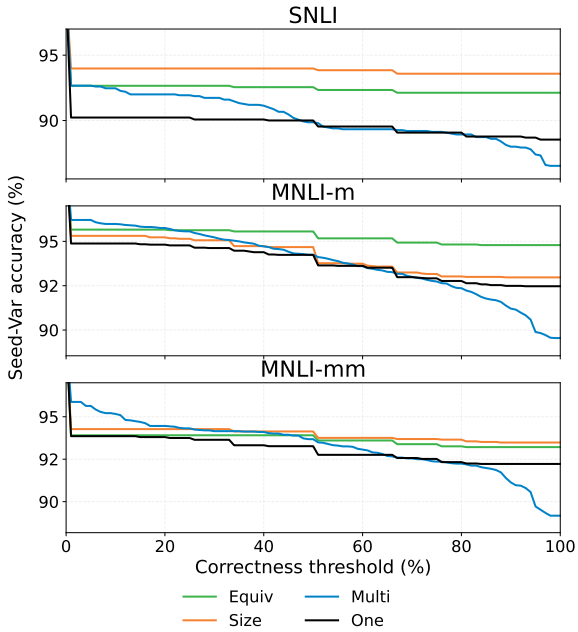

Figure 7: SV curves for ALBERT on the seed problems (SNLI=150, MNLI-m=530, MNLI-mm=424) that had variants formed with replacements from BERT, RoBERTa, Electra and ALBERT, all sizes tested, grouped by their origin MLM.

Limitations

When it comes to our methodology, we do not replace words that are tokenized into subwords, as aggregating their probabilities could introduce artifacts, a limitation that could be explored in future inquiries. Additionally, when original words have several occurrences in the problem, we mask only one instance at a time, leaving the other unmasked, which could bias the replacements of models. However, problems with such cases are a small proportion of the seed datasets (SNLI=≈2%, MNLI-m=≈6%, and MNLI-mm=≈9%). Another methodological limitation is that we do not replace morphological derivations of original words to avoid incorrectly replacing them when part of 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=≈5%; MNLI-m=≈7%, and MNLI-mm=≈6%). Lastly, another potential limitation could be that larger pre-trained models could be added to the analysis.

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.

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.

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. 2020a. 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.

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. 2020b. 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.

Lovish Madaan, David Esiobu, Pontus Stenetorp, Barbara Plank, and Dieuwke Hupkes. 2024. Lost in inference: Rediscovering the role of natural language inference for large language models. *Preprint*, arXiv:2411.14103.

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. 2019a. 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. 2019b. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*.

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, 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, Online. Association for Computational Linguistics.

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.

Gemma Team. 2024. Gemma.

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-dev	N	7363	33	29	38
	V	3780	32	28	40
	Adj	1067	34	24	42
	Adv	76	26	13	61
MNLI-m	N	5,979	29	31	40
	V	4,438	27	32	41
	Adj	1,923	26	25	48
MNLI-mm	N	6,640	29	31	40
	V	4,614	27	32	42
	Adj	2,197	26	28	47

Table 3: Number and the **n**eutral, **c**ontradiction, and **e**ntailment labels of **p**otential **s**eed problems of SNLI-test and MNLI-m/mm-dev sharing at least one replacement class, i.e. at least one **n**oun, **v**erb, **adj**ective, or **adv**erb. Note that adverbs were excluded due to their low count in SNLI, and therefore were not computed for MNLI.

MLMs The used MLMs for replacement generation are shown Table 4. Besides the aforementioned models, DeBERTa was considered as an option, but its replacements were too noisy for inclusion. From the models shown in the aforementioned table, BART was excluded, given its bigger contribution to bad variants in SNLI, as shown in Figure 9, read subsection A.3 for more details about the exclusion itself. For MNLI replacements, BART was not considered by default to not create plausability differences across variants of different datasets/

Class & Probability Filtering We tagged replacements r in the context of their corresponding problems using the spaCy model `en_core_web_sm`, and excluded those whose classes differed from the original word o. We also stored the probability of each r of o in p or h from the MLM that suggested it, and we excluded those of lower probability than the original replaced word. MERE validates variants by assuming that replacements

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 4: Overview of MLMs used for replacement generation. The columns show the **Size** – **B**ase, **L**arge, or **XXL**arge; **Architecture**: **E**ncoder, **D**ecorder or **E**ncoder-**D**ecoder, or the **vocabulary** size of the models.

as likely as the 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 class filtering are important, as highly probable replacements might not have the correct class and vice versa. Table 5 shows the percentages of replacements with higher probability than o under Column **Prob.** $\triangledown$, and the average of correct replacements after excluding different classes in **POS.** $\triangledown$.

A.2 Which filtering criteria matter?

To test if different filtering criteria for r_{ij} affect NLI models trained on SNLI and SMFA, we selected new variants for the seed problems of SNLI $_{\text{Var}}$ (ALL_{Var}) forming datasets with: i) r_{ij} of the union of $p - h$, instead of their intersection in Equation 1 – $_{p \cup h}$; ii) r_{ij} only having o^c – $_{\text{POS}}$; iii) r_{ij} only having $o_{i>}$ – $_{\text{Prob}}$; iv) r_{ij} of any class or probability – $_{\text{None}}$; v) r_{ij} with their letters randomly scrambled – $_{\text{Scr}}$. Except for $_{\text{Scr}}$, the datasets are more diverse given the less strict filtering, while still excluding punctuation signs and v_{ij} already part of $P\&H$. We plot the averaged SV scores of all models in Figure 8, highest performance being achieved on $_{\text{Scr}}$, followed by $_{\text{Var}}$, $_{\text{Prob}}$, $_{\text{Pos}}$, $_{\text{None}}$, and $_{P \cup H}$. While $_{\text{Scr}}$ starts as the lowest curve (in green), it ends up as the one with the highest scores on very high thresholds. The other variant datasets seem to follow a constant downward trend where the more variants are considered, the lower the scores get. Out of these, $_{P \cup H}$, $_{\text{Pos}}$, and $_{\text{None}}$ have the lowest scores, which might be caused by the higher number of unique variants[11] per dataset, which enlarges its overall lexical diversity. Thus, the number of variants seems more important than controlling for factors such as probability, or plausibility.

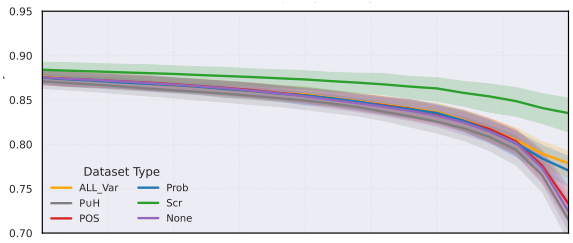

Figure 8: Averaged SV curves of all models on SNLI datasets formed with r_{ij} having any class or probability – $_{\text{None}}$, being scrambled – $_{\text{Scr}}$, with higher probability than the original word – $_{\text{Prob}}$, with the same class – $_{\text{Pos}}$, suggested in either p and h –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

A.3 Annotation

Two authors annotated variants for SNLI, and one for MNLI. Replacements were evaluated considering how much they changed the fluency and the original logical label of the NLI problems, which were considered to be correct. The scores used were: 1 – poor, 2– mostly poor, 3– uncertain, 4– mostly good, 5–good. Table 9 shows the annotation guidelines, 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 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 9:

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

[11] All three datasets have around $\approx$ 380k unique variants each, while each of the others have $\approx$ 190k.

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

Table 5: Average percentages of replacements per datasets with the same class (Pos ▽) and equal or higher probability than the original replaced word (Prob ▽) out of 200 replacements, for both the **P**remise and **H**ypothesis

.

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.

The Cohen's kappa for annotators was 80% for nouns, 77% for adjectives, and 89% for verb-formed variants. In SNLI, only a few (5%) variants received scores lower than 3 on R, due to a lack of label preservation or ungrammaticalities. After annotation, 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. Other models were not excluded as a confusion matrix, shown in Figure 11, indicated there were more unique than overlapping replacements across models, with BERT and Electra models showing nearly equal proportions of both. Replacements for MNLI variants already excluded BART ones, and the annotation for both MNLI-m and MNLI-mm, i.e. 100 problems for each open-class category, had all problems with good R scores, and only one problem with a lower fluency score, indicating the aforementioned exclusion of the model was sufficient in overall creating plausible variants.

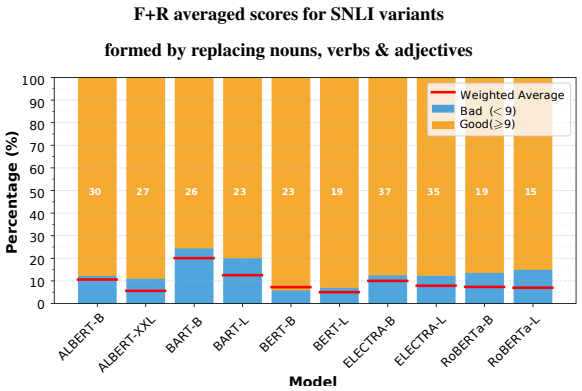

Figure 9: Averaged **F**luency and **R**easoning scores for the normalized counts of 100 random SNLI variants for nouns, verbs, and adjectives. The red lines show bar plots weighted considering the distribution of open classes of SNLI seed problems (N=67%, V=23%, ADJ=10%). Good variants have a score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

A.4 Evaluated models

Evaluated Models The models we evaluated alongside with their model cards are in Table 7.

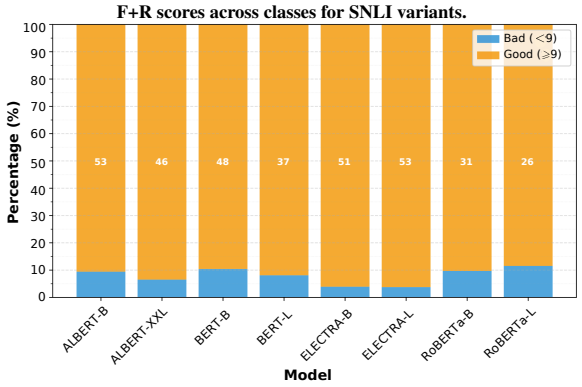

Figure 10: **F**luency and **R**easoning scores for 100 randomly sampled SNLI variants across classes, their normalized counts, and the models that validated them, after exclusion of replacements from BART. Note that 91 out of the 100 examples evaluated had $F + R \geq 9$.

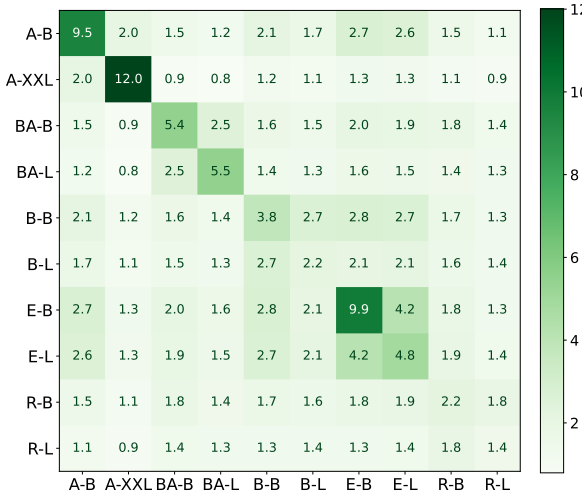

Figure 11: Confusion matrix showing how models' replacements intersect on average, across problems, with the averaged number of model-specific replacements on the diagonal. For instance, a replacement like 'bear' counts on the diagonal if unique to one model, or in the other cells if shared between models.

A.4.1 Pre-trained models

Seed problems From each dataset, we randomly selected 100 random 500 seed problems and compared models' NLI S scores on them. Model scores on subselected seeds varied around 3% from their scores on the bigger seed datasets, indicating that random subsampling preserves original seed distribution. From these sub-samples, we selected a random sample from each dataset to test pre-trained models 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 6, while the few-shot examples per dataset appear in Table 8. Note that the few-shot examples are balanced across labels and of 2 types, i.e. randomly drawn once only from the variants' source dataset, or a mix from both datasets. With our evaluation method, we replicated the scores of Madaan et al. (2025) of Llama-3.1-8B on SNLI dev twice: with the current batch of exclusively SNLI-dev shots from Table 8, and 6 other randomly chosen ones. 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, both models were only evaluated once over the set of unique variants to halve computing time. We investigated whether running Llama directly on the whole variant dataset, which contains duplicates, changes its prediction, and found that they do so in 0.1% of the cases.

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 6: Prompt template used for pre-trained model evaluation from Madaan et al. (2025).

Model	Model Card
SNLI-only models	
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
SMFA models	
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
MNLI-only models	
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
Pre-trained LLMs	
Gemma-2-9b-exc/mix	google/gemma-2-9b-it
Llama-3.1-8B-exc/mix	meta-llama/Meta-Llama-3-8B

Table 7: Evaluated models and their Hugging Face references.

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 8: Few-shot examples used for pre-trained model evaluation. Note that few-shot for SNLI and MNLI contain random examples exclusively from SNLI-dev and MNLI-train, while the mixed row corresponds to few-shots comprised of examples from both datasets.

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 9: 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.

	SNLI		MNLI-m		MNLI-mm	
Model	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 10: Variants across datasets grouped by the source model for their replacement, with raw and % counts shown. The **Multi** row shows replacements that were validated by at least two models.

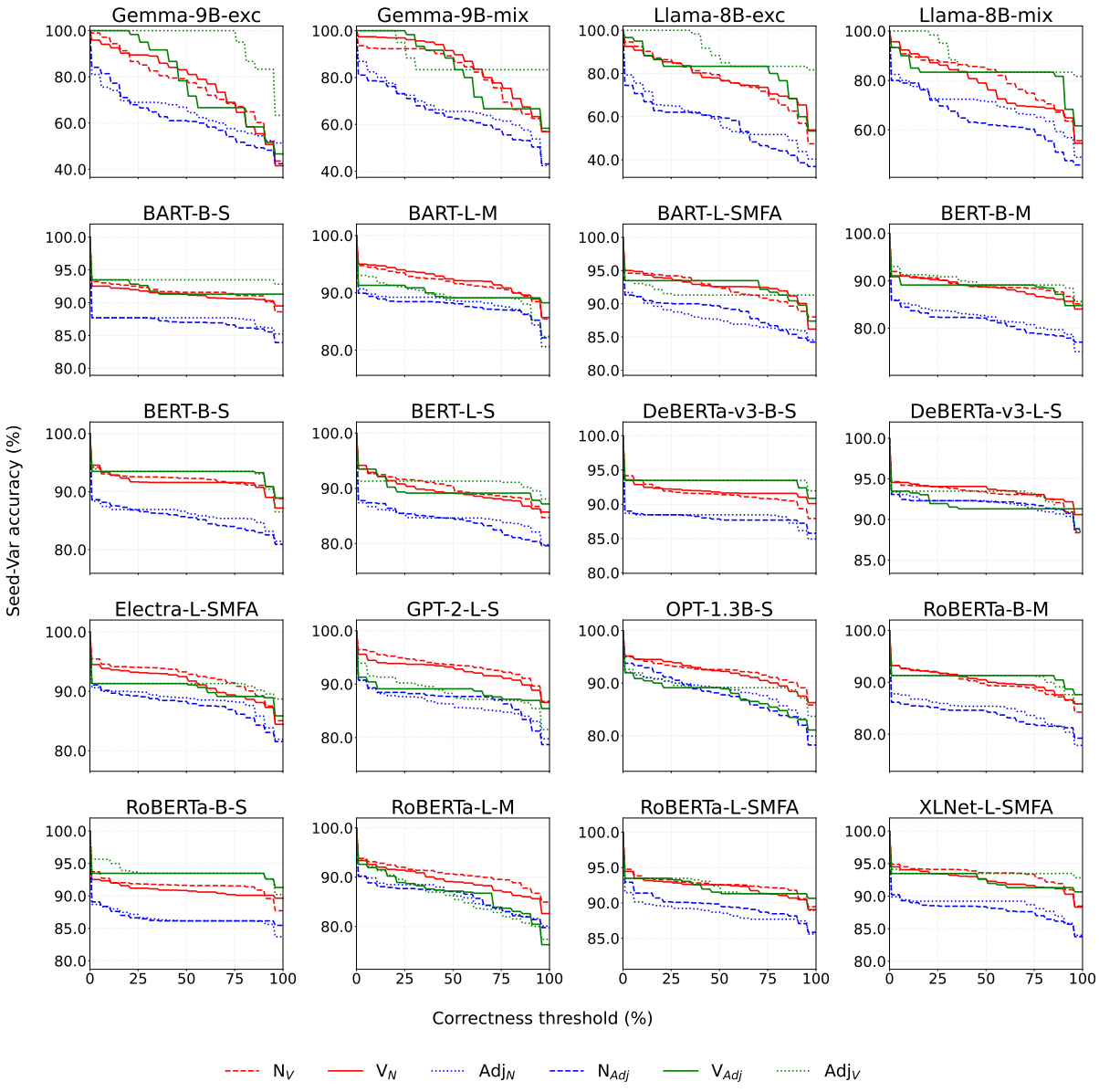

Figure 12: Seed-Var accuracy scores for SNLI seed problems with at least two out of the three open replaced classes in each seed. The legend shows, by the bigger letter of each dataset, which class was replaced in the seed problems that had both of them.

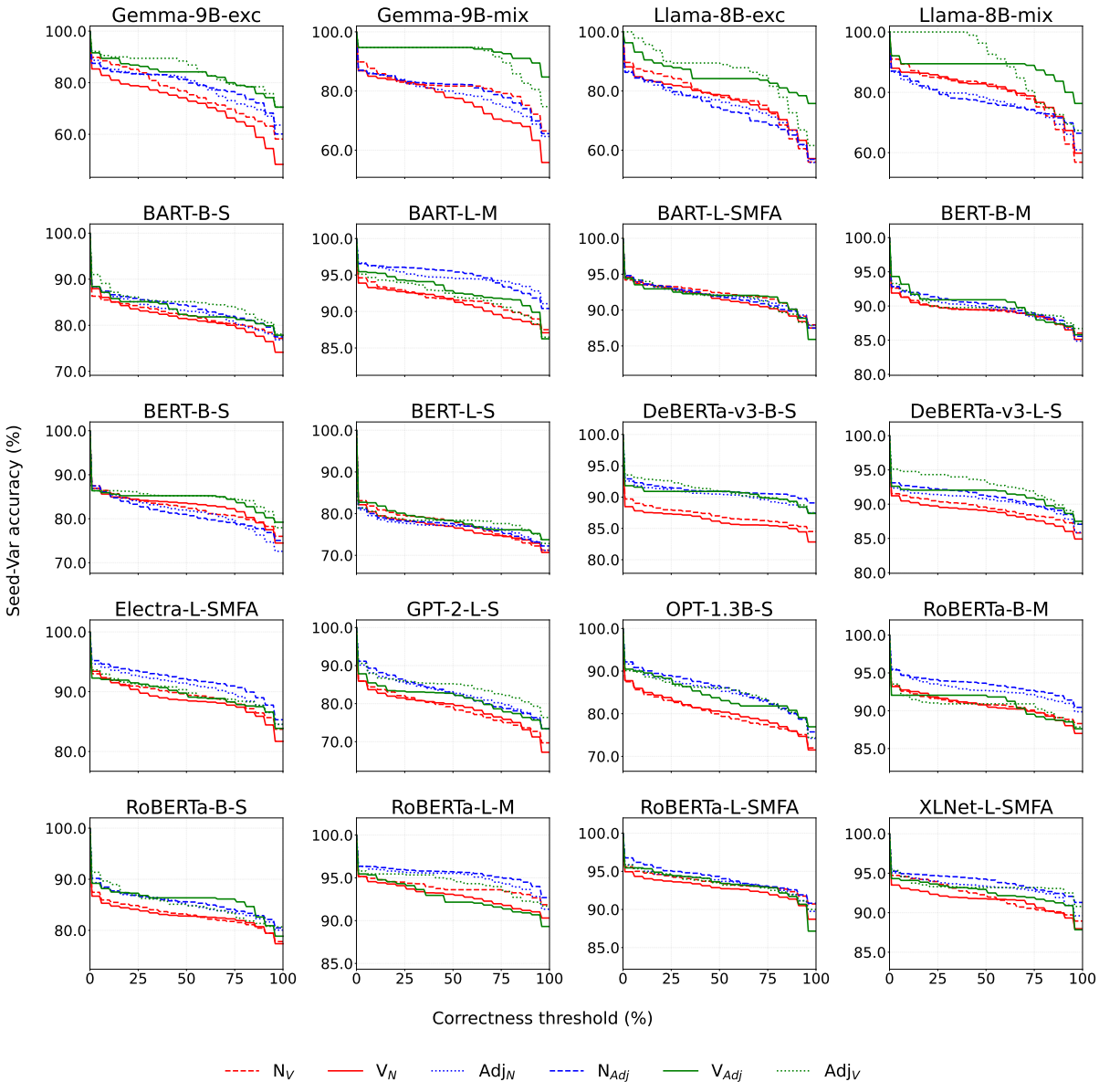

Figure 13: Seed-Var accuracy scores for Mnli-mm seed problems with at least two out of the three open replaced classes in the seed problems.

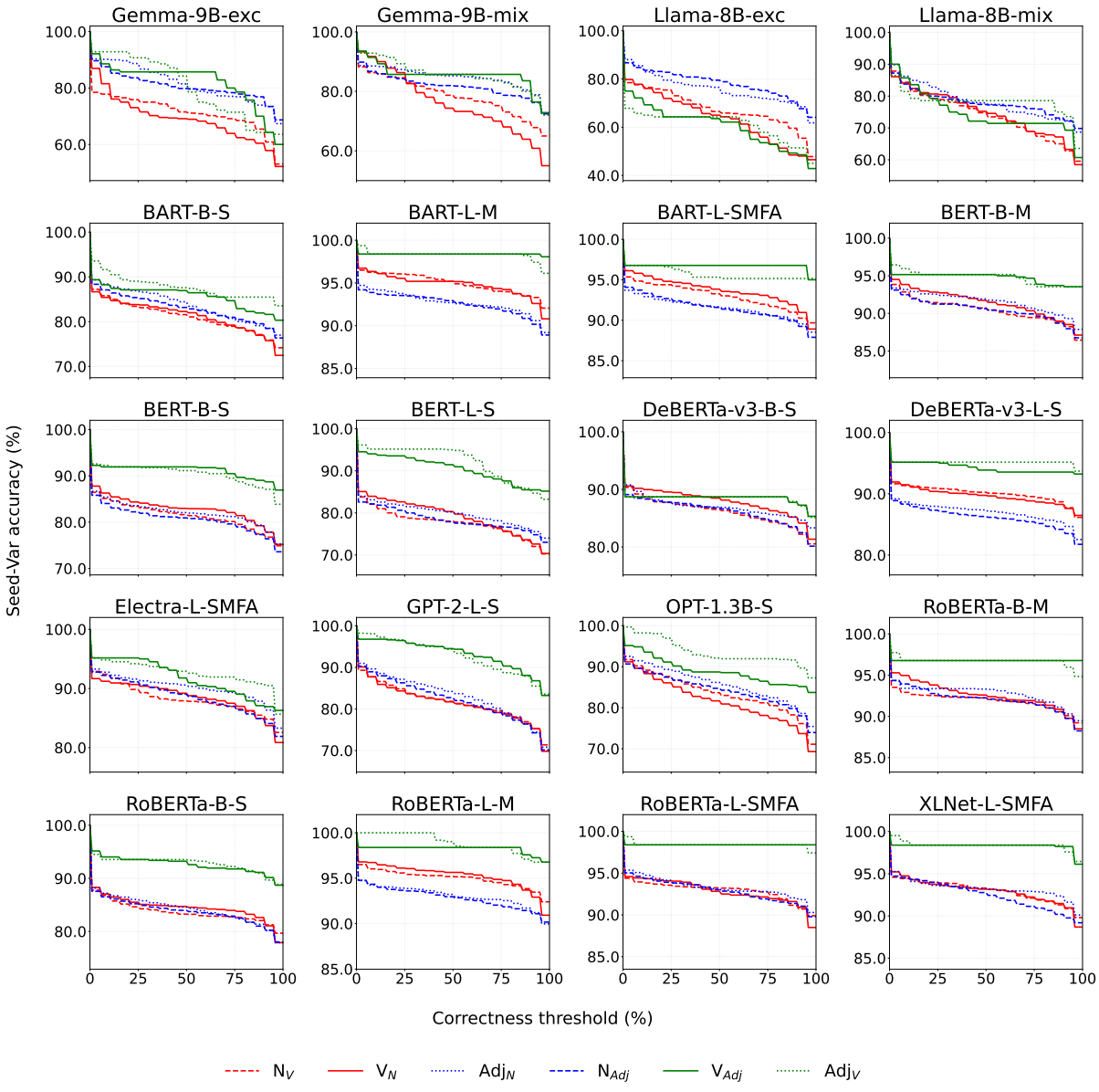

Figure 14: Seed-Var accuracy scores for Mnli-m seed problems that have at least two out of the three open replaced classes in them.

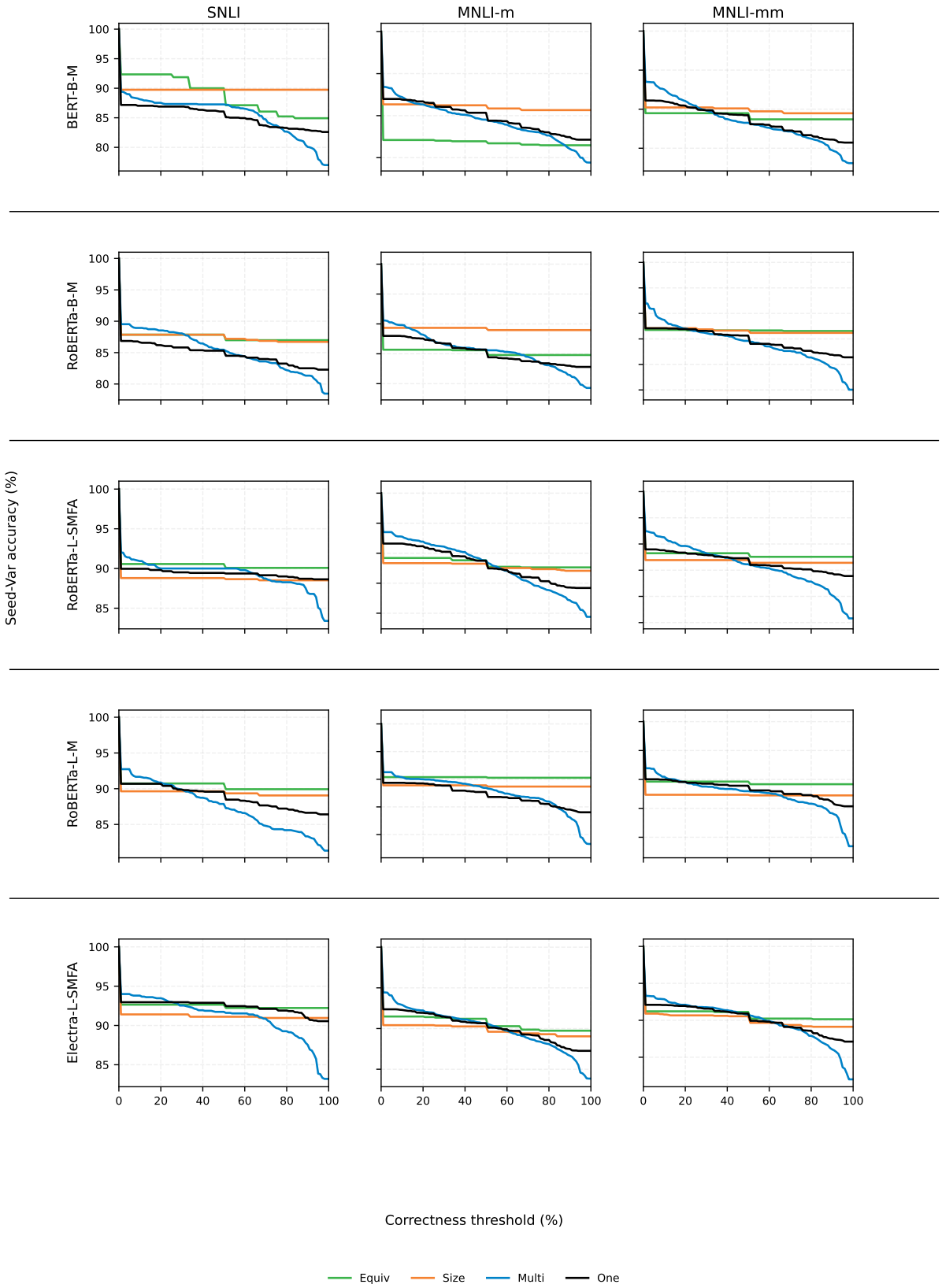

Figure 15: Seed-var accuracy curves across datasets for variants grouped by the origin model that suggested them. We only show the models fine-tuned on SMFA and MNLI, as SNLI fine-tuned models had lower scores and were less informative.

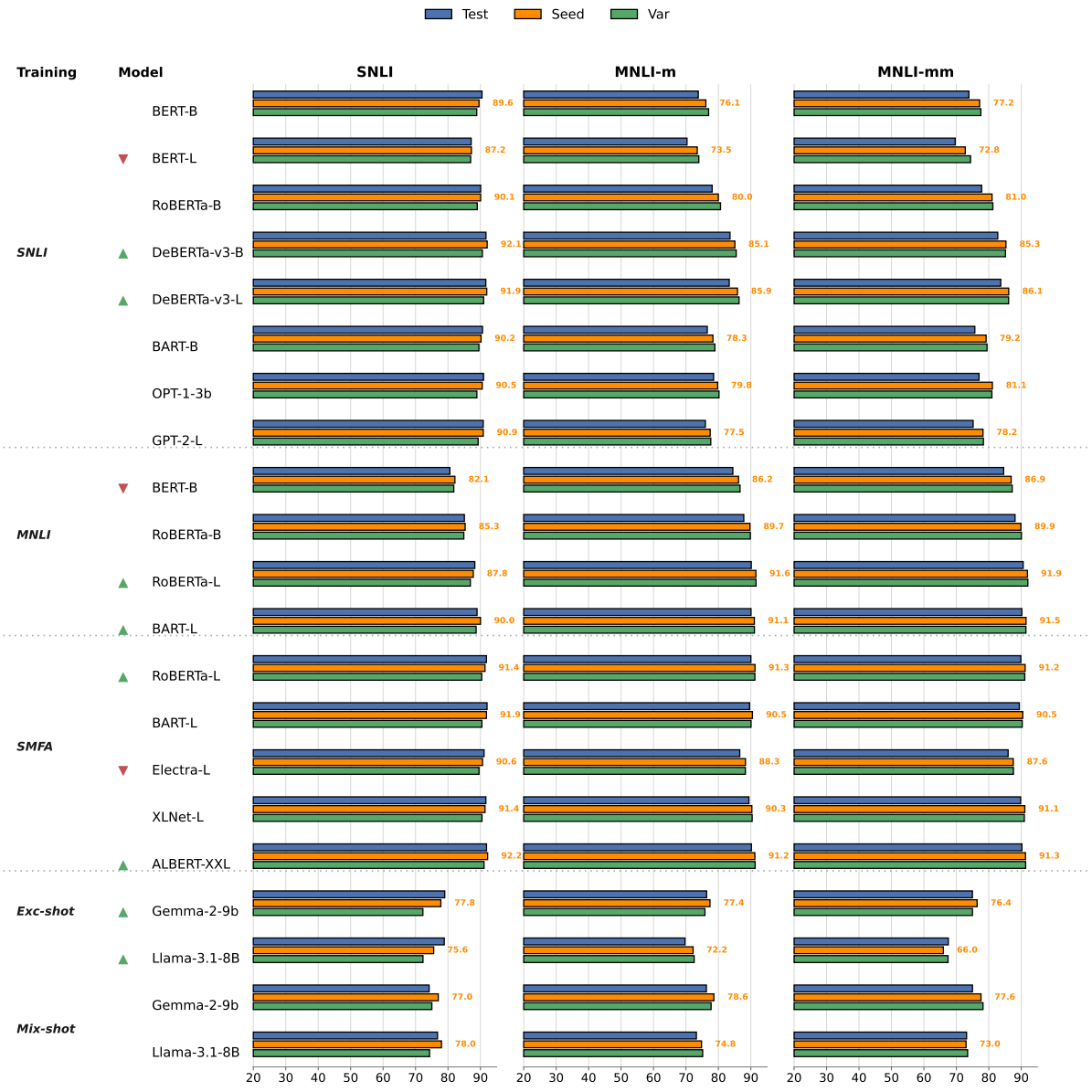

Figure 16: Results of all models, grouped by their training either fine-tuned on SNLI, MNLI, SMFA, or pre-trained models prompted with examples exclusively chosen from SNLI-dev (for SNLI-var) or MNLI-train (for MNLI-m/-mm). **Test** scores are S scores on the test problems of each dataset, while **Seed** and **Var** are S-acc scores on seed and variant problems. The green triangles in front of the models from each category show which two models were best, while the red ones show the worst.

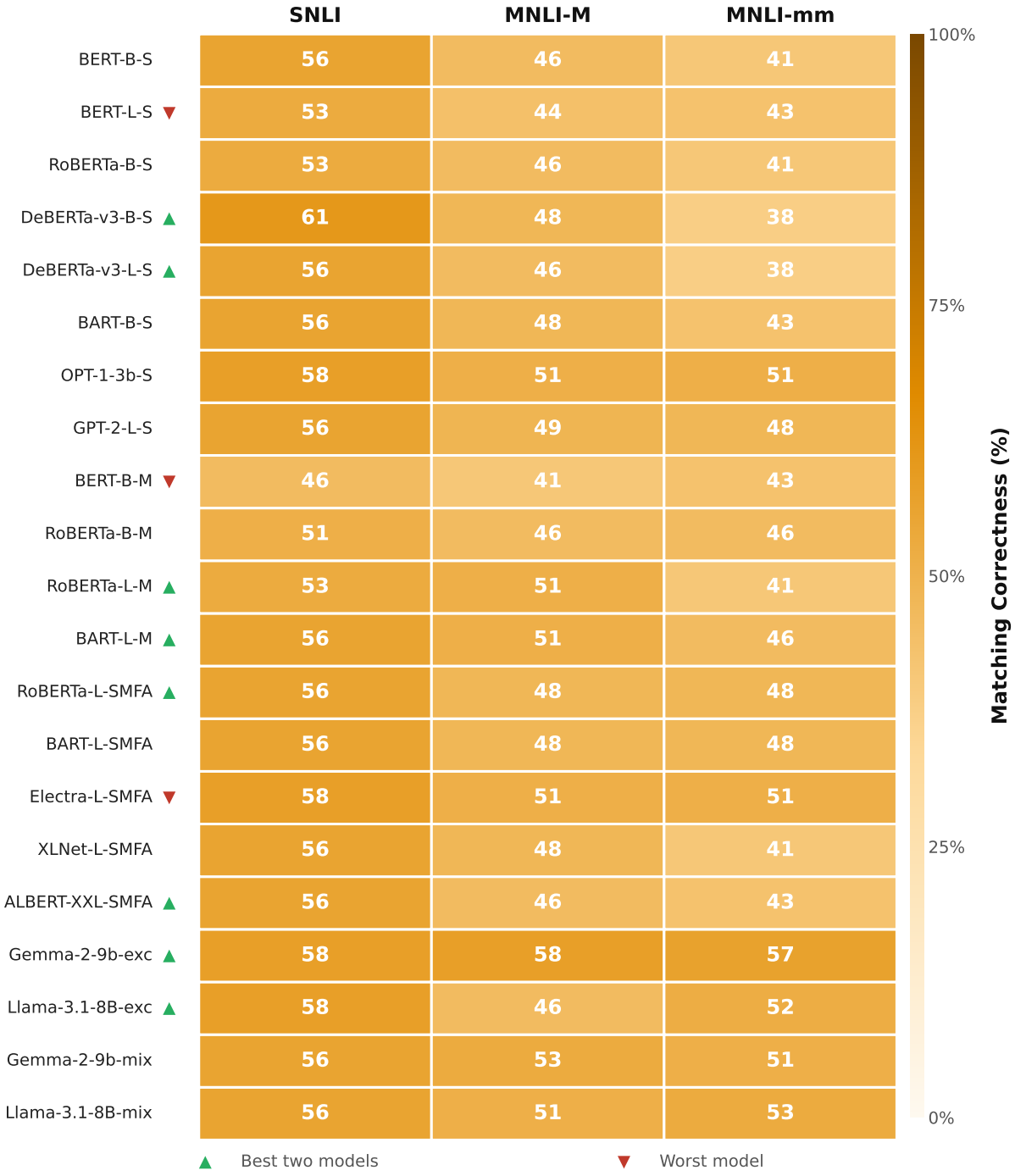

Figure 17: Matching correctness thresholds across all models and datasets: the numbers show on which CT models get the closest SV scores to their initial S seed scores. As most numbers are $\approx 50\%$, they indicate models do not generalize to more variants than that.

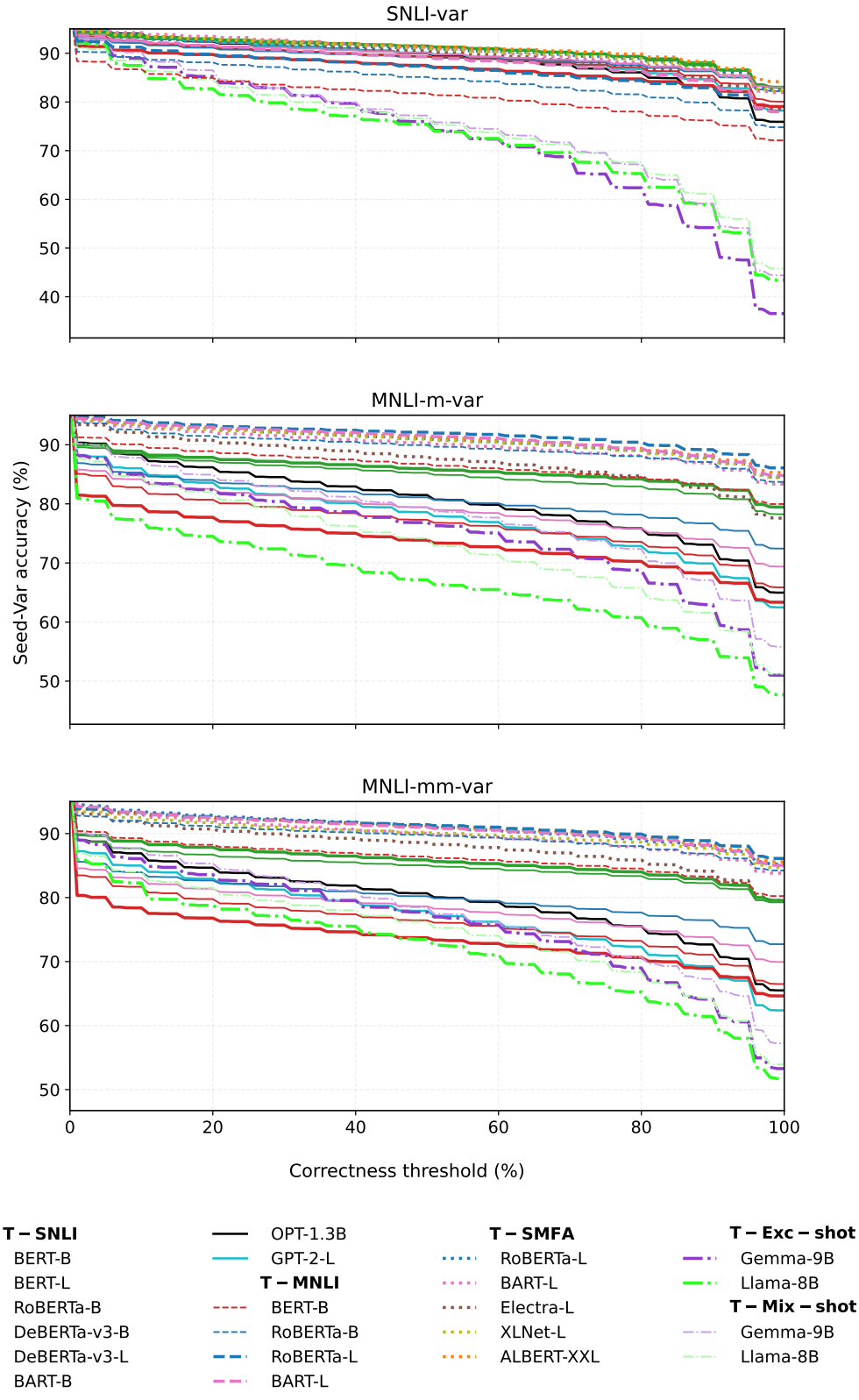

Figure 18: SV accuracy curves on variants from SNLI-var, and MNLI-m/-mm-var. The legend shows NLI models and the datasets they were fine-tuned on, i.e. SNLI, MNLI or a combination of NLI datasets, e.g., SMFA, or, if applicable, the type of few-shot prompts we used for pre-trained models, i.e. exclusively taken from SNLI-dev (for SNLI-var), or MNLI-train (for MNLI-m/-mm var).

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 cliche; at any rate, Chatterbox hadn't previously been aware of its overuse.
H: Robust is a legitimately new cliche 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 11: The original SNLI, and MNLI-m/-mm problems whose variants were most poorly classified by all NLI models, in-context models were also considered, but their sub-selected seed did not contain 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 data, where the gold label is selected based on the majority voting. **Correct label** is an inference we thought 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. 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.