SpaceNLI: Evaluating the Consistency of Predicting Inferences in Space

Lasha Abzianidze **Joost Zwarts** **Yoad Winter**
Institute for Language Sciences, Utrecht University
Utrecht, the Netherlands
{l.abzianidze, j.zwarts, y.winter}@uu.nl

Abstract

While many natural language inference (NLI) datasets target certain semantic phenomena, e.g., negation, tense & aspect, monotonicity, and presupposition, to the best of our knowledge, there is no NLI dataset that involves diverse types of spatial expressions and reasoning. We fill this gap by semi-automatically creating an NLI dataset for spatial reasoning, called SpaceNLI.[1] The data samples are automatically generated from a curated set of reasoning patterns (see Figure 1), where the patterns are annotated with inference labels by experts. We test several SOTA NLI systems on SpaceNLI to gauge the complexity of the dataset and the system's capacity for spatial reasoning. Moreover, we introduce a *Pattern Accuracy* and argue that it is a more reliable and stricter measure than the accuracy for evaluating a system's performance on pattern-based generated data samples. Based on the evaluation results we find that the systems obtain moderate results on the spatial NLI problems but lack consistency per inference pattern. The results also reveal that non-projective spatial inferences (especially due to the "between" preposition) are the most challenging ones.

1 Introduction

Natural language inference (NLI) is a popular task that evaluates NLP systems on text reasoning skills. In the task, a system has to predict an inference relation from a premise text to a hypothesis sentence/phrase. Usually, the task is three- or two-way classification, depending on whether in the inference labels of *entailment*, *neutral*, and *contradiction*, the latter two are merged into *non-entailment*. The task is intended for evaluation of NLP systems on reasoning, however, the systems with competitive results on NLI benchmarks are often exploiting dataset biases (Tsuchiya 2018; Poliak et al. 2018; Gururangan et al. 2018; McCoy et al. 2019, *inter alia*) and their performance suffers from out-of-distribution NLI sample problems (Glockner et al., 2018).

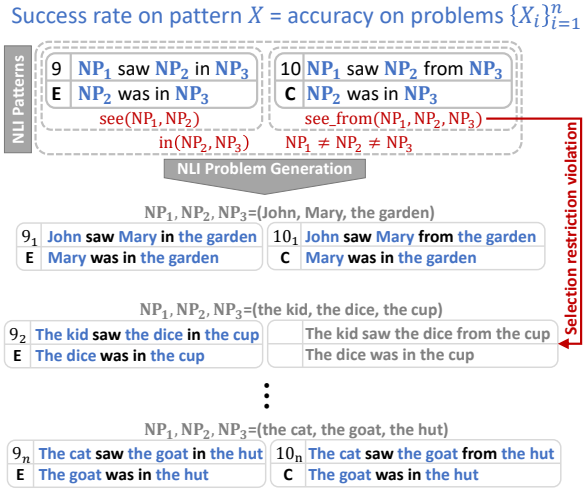

Figure 1: Sampling NLI problem from NLI patterns (with IDs 9 and 10, **E**ntailment and **C**ontradiction, respectively). The problems are generated by replacing NP placeholders with definite NPs that satisfy pattern-specific selection restrictions. A system's success rate on a pattern is defined as the accuracy on its corresponding NLI problems.

To better evaluate the reasoning skills of NLI systems, a series of works have been (semi-)automatically or manually creating NLI datasets that specialize in certain semantic phenomena. While some of these datasets come with a training part, most of them are intended solely for evaluation. For example, several datasets have been dedicated to monotonicity reasoning (Yanaka et al., 2019b,a, 2020), negation was targeted by Hossain et al. (2020), the dataset by Kober et al. (2019) focuses on temporal and aspectual inferences, Jeretic et al. (2020) semi-automatically generated NLI problems for implicatures and presuppositions. There are also NLI datasets that cover several semantic phenomena, having a separate section for each of the phenomena (Cooper et al. 1996;

[1] https://github.com/kovvalsky/SpaceNLI

Proceedings of the 4th Natural Logic Meets Machine Learning Workshop (NALOMA23), pages 12–24

Richardson et al. 2020, *inter alia*).

While spatial reasoning has been included in several multi-modal QA datasets (Antol et al., 2015; Suhr et al., 2017; Johnson et al., 2017; Hudson and Manning, 2019) and in a couple of text-based QA datasets (Weston et al., 2016; Mirzaee et al., 2021), to the best of our knowledge, no NLI dataset has specifically covered it.[2] This paper fills the gap by semi-automatically creating an NLI dataset for spatial inferences. First, we collected a diverse set of NLI problems inspired by the inference examples found in the literature on spatial semantics. Second, the NLI problems were manually converted into NLI patterns (see Figure 1), and finally, we automatically generated a large number of NLI problems from the patterns.

The paper makes two main contributions:

C1. SpaceNLI: the spatial NLI dataset with diverse types of spatial inferences; The inference labels of the generated problems are highly faithful (97%) to the labels of the corresponding original patterns.

C2. Pattern accuracy and its curve: they measure systems' performance on patterns and the consistency of predictions on samples from the same patterns.

The conducted experiments answer the following research questions:

Q1. How much spatial reasoning current SOTA NLI systems are capable of?

A1. We found out that the SOTA NLI systems have problems with fine-grained spatial inferences. Their performance drops at least by 24% compared to their results on common NLI datasets. Moreover, their consistency in predictions is sensitive to irrelevant lexical substitutions.

Q2. What types of spatial inference problems are easy or challenging for the SOTA NLI systems?

A2. The results showed that the non-projective spatial relations are most challenging for the models. This was mainly due to difficulty associated with "between" and its frequent occurrence in the evaluation dataset.

[2]Even the FraCaS dataset (Cooper et al., 1996; MacCartney, 2009), which was curated by linguists and semanticists, doesn't cover spatial semantics within its nine sections.

2 Spatial expressions and inferences

2.1 Types of spatial expressions

Spatial expressions consist of spatial prepositions and other expressions with spatial information (e.g., *far*, *the left of*, and *in front of*). They usually describe a relation between two entities, the *figure* and the *ground*. The site or path of the figure is the focus of the discussion and is characterized with respect to the ground. For example, in (9_1) and (10_1) from Figure 1, *Mary* is a figure and *garden* a ground. *John* is also a figure in the premise of (10_1).

Spatial expressions are roughly divided into *locative* and *directional* expressions, where locatives can be further classified into *projective* and *non-projective* (Herskovits, 1986). The locative expressions describe static, locative relations between the figure and the ground while directional ones describe a more *dynamic* relation involving a movement and/or path. An example with a directional preposition is *Cindi walked into the market*. The spatial expressions in Figure 1 are all locative except for *from*, which is directional. These locative expressions are non-projective since they require only the spatial location of the figure and the ground. In contrast, projective locatives additionally require further information from the ground in terms of a deictic frame of reference (i.e., an orientation structure). For example, the site of the house is not sufficient to interpret Mary's location in *Mary is behind the house*, it requires knowledge about the frame of reference of the house, in particular, what counts as a back side of the house.

2.2 Types of spatial inferences

We characterize spatial inferences depending on the type of spatial expressions licensing them. An inference might depend on several spatial expressions of a different type, which makes partitioning the inferences challenging, if not impossible. We define the following classes that represent a coarse-grained partition of spatial inferences. The classes will be later referred to in §3.[3]

Argument orientation In spatial literature, an argument orientation entailment identifies which

[3]Licensing contradiction and neutral problems will be assumed from the perspective of a related entailment problem. For example, we assume that the neutral problem (16) in Table 1 is licensed in the same way as its related entailment (15). Put differently, one can see (16) as an adversary to (15) and assume that solving (15) requires competence comparable to the one required for solving (16).

ID	Class	Premise(s)	L	Hypothesis
15	Dir	John threw the ball into the box.	E	The ball went into the box.
16	Dir	John threw the ball at the box.	N	The ball went into the box.
31a	Dir	Los Angeles is in California. John came from California.	N	John came from Los Angeles.
38	NonP	John is in the garden. The garden is in the church.	E	John is in the church.
41	Dir	John drove through the tunnel.	E	John was in the tunnel.
47a	Dir	Cindi walked into the market.	E	Cindi was outside the market.
56c	Proj	The trash can is to the right of the tree from John.	C	The tree is to the right of the trash can from John.
70	Proj	Mary is between the tree and the house. The tree is behind the house.	E	Mary is behind the house.
80	NonP	The cat is between the house and the fence. The cat is between the fence and the tree.	C	The cat is between the house and the tree.
99*d	Proj	The bucket is above the bowl. The pencil is above the bowl.	N	The bucket is below the pencil.
96b	ArgO	Mary met John at the party.	N	Cindi was not at the party.
100	NonP	The house is far from the school.	E	The school is far from the house.
102a	ArgO	Mary has taken the cup out of the cabinet.	C	The cup is in the cabinet.
102f	ArgO	Mary has hidden the cup behind the cabinet.	E	The cup is not in the cabinet.

Table 1: Examples of the seed NLI problems annotated with spatial inference classes: **Dir**ectional, **Proj**ective, **Non-P**rojective, and **Arg**ument **O**rientation. Initial letters abbreviate the corresponding inference labels.

argument of the verb is the figure of the spatial expression. For instance, (9_1) in Figure 1 show that *Mary* is the figure of the locative PP *in the garden*. In its original interpretation, the argument orientation entailment is not restricted to spatial expressions of a particular type. Here, we restrict the class of argument orientation to the entailment problems (and their neutral and contradiction counterparts) that come close to resolving a PP attachment. For example, correctly resolving the PP attachment in (9_1) boils down to the hypothesis. The problems in this class contain a hypothesis with a copula and a predicative spatial PP, where the PP is contrasted to a tightly related PP in the premise(s). For more examples of the NLI problems in the argument orientation class, see Table 1.

Directional The directional class contains spatial inferences where directional spatial expressions play the key role. Examples of such inferences are given in Table 1. Some of these NLI problems pertain to a path-place relation: (47a) shows that *walking into* infers *being outside*;[4] (41) entails *being in the tunnel* from the premise that states that the driving path was through the tunnel. (31a) combines a part-whole relation with the movement path.

Projective This class contains inferences that hinge on a frame of reference introduced by projective spatial expressions. In principle, the frame of reference can introduce six directions that can be referred to using the expressions like *front, behind, left, right, above, below, under, on top of*, etc. (see the examples of NLI problems in Table 1). The NLI problems that contain *on top of* as only projective spatial expression, and when its projective interpretation is not crucial for the inference, are put in a different class.

Non-projective We classify a problem as having non-projective inference if the inference is driven only by non-projective spatial expressions. Therefore, an occurrence of non-projective spatial expressions in a problem is necessary but not sufficient for assigning the problem to this class, e.g., see directional problems (31a) and (41). NLI problems that depend on spatial expressions with the semantics of order and proximity are also in this class, see *between* (80) and *far* (100) in Table 1.

3 Dataset construction

3.1 Pattern construction

Patterns are labeled NLI problems with NPs replaced by variables as illustrated in Figure 1. The NLI patterns are obtained from the seed NLI problems. To collect the latter, we extracted the initial 56 problems from Zwarts and Winter (2000) and Nam (1995), where a majority of the problems were labeled as entailment due to obvious biases in the semantic literature towards licensing entail-

[4] Since moving along the path is related to the change of the location, sometimes spatial entailments interfere with tense and aspect.

Class (#patterns)	Spatial expression counts
Directional (95)	in (20), from (17), into (9), to (8), on (8), away from (7), towards (7), out of (4), back (3), through (3), across (2), at (2), outside (2), opposite (1), part of (1), by (1)
Argument orientation (67)	in (21), at (10), from (9), away from (4), out of (4), near (3), with (3), inside (3), on (2), under (2), through (1), opposite (1), towards (1), far from (1), on top of (1), behind (1)
Projective (70)	behind (16), between (11), in front of (10), below (6), above (6), under (6), on top of (5), front of (3), opposite (2), to the right of (2), on (2), to the left of (1)
Non-projective (48)	between (22), in (9), far from (5), close to (4), outside (3), on top of (2), on (2), opposite (1)

Table 2: The spatial expressions and their counts per entailment class in the SpaceNLI patterns

ment. To create a representative and challenging NLI dataset for machine learning, we applied several *revision phases* to the problems: introducing new problems that either cover new semantic aspects of spatial expression or serve as a perturbed version of an existing problem.

In the initial revision phase, four annotators divided the extracted problems and created slightly modified versions of them with an inference label different from the original.[5] This was motivated by the current trends in the literature on adversarial, stress, and debiased datasets (Naik et al. 2018; Ribeiro et al. 2020; Kaushik et al. 2020; Gardner et al. 2020, *inter alia*). For example, (16) is a perturbed example of (15). Where possible, NLI problems of a new type were also created using the similar spatial expressions found in the extracted problems.

To validate the resulting pool of NLI problems (in total 162), following (Zhang et al., 2017), they were labeled on a 5-point Likert scale by three annotators.[6] After collecting the 5-point annotations, for each annotator, we picked a mapping of 5-point to 3-point that maximizes the inter-annotator agreement (avg. Cohen's $\kappa = .71$). The problems without majority labels were discarded and 111 problems remained.

To better balance the inference labels and increase the coverage of spatial expressions, a second revision phase was carried out on the remaining problems. In several cases, problems with low annotator agreement were revised, e.g., changing the tense where it caused confusion or replacing a preposition with a weaker version (*at*↦*near*). All the new and revised problems (in total 63) were validated based on three samples: each problem was manually converted into a pattern by replacing NPs with variables, and three random NLI samples per pattern were generated (see §3.2 for details), which were subsequently validated by three annotators.

Finally, a third revision phase was carried out on the remaining problems to additionally decrease the overall and spatial type-specific label imbalance. The collected problems (in total 160) were treated as a seed by converting them into NLI patterns to generate a large amount of sample NLI problems from them. To illustrate the coverage of spatial expressions in the collected patterns, Table 2 gives the complete list of spatial expressions for each entailment class.

3.2 Sample generation

We manually created NLI patterns from the initially collected NLI problems (§ 3.1) by replacing NPs with placeholders and specifying selection restrictions for them imposed by the verbs, spatial expressions, and gold inference labels (see Figure 1). The selection restrictions imposed by spatial expressions are subtle and can affect gold labels or the naturalness of sentences. For example, if the figure is much larger than the ground, it can make the sentence infelicitous: *the apple on the fridge* and *the apple near the fridge* are preferred to *the fridge under the apple* and *the fridge near the apple*. Inferences driven by proximity-related spatial expressions are sensitive to the size of the objects. For instance, based on our conducted validations, *Cindi is opposite to the cat* is more likely to be

[5]The annotators for the pattern construction consist of the authors of the paper, two linguist students, and one AI student. The guideline for creating inference problems can be found in the supplementary material.

[6]The question was to what extent the hypothesis sentence is true, given that the premises are true, with choices: *definitely false*, *most likely false*, *unknown*, *most likely true*, *definitely true*. We used two additional choices, *difficult* (unable to annotate due to the complex reasoning it requires) and *skip* (presence of an ungrammatical or nonsensical sentence). We used the brat annotation tool (Stenetorp et al., 2012) for labeling. The annotation guideline is included in the supplementary material.

neutral to *Cindi is far from the cat*, but *the school is opposite to the house* is more likely to contradict *the school is far from the house*.

To meet selection restrictions and allow relative diversity of NPs in the generated samples, we defined a mini world with a domain containing 171 entities corresponding to common and proper nouns. The entities are organized in a taxonomy with 20 subclasses covering general types of entities (e.g., person, animal, vehicle), the projections of an argument in certain argument structures (e.g., enter in X, be in X, throw X), compatibility with projective spatial expressions, and size categories (S for entities comparable to small objects like book and cat, M to persons, and L to vehicles). Binary and ternary relations are defined based on the set unions of the products of entity sets and subclasses.

To automatize the sampling of sound NLI problems from the patterns, we formatted the mini world in YAML and NLI patterns in XML. We implemented a procedure that samples problems from the patterns by filling in NP placeholders with definite NPs from the mini world and respecting the pattern-specific selection restrictions. For sanity checking, the procedure verifies that it can generate corresponding seed NLI problems for each pattern.

To measure how faithfully the inference labels are transferred from seed and pattern NLI problems to the corresponding NLI samples, we used sampled problems in the second phase of validation when validating new NLI problems (see §3.1). The results showed that 79% of samples were unanimously labeled with the original label. After filtering out patterns with a relatively low agreement, this ratio increased to 97% for the samples generated from the validated patterns.

The NLI problems sampled from the same pattern or related patterns are string-wise very close to each other, sometimes differing only in terms of occurrences of a single NP. Regardless of this similarity, we expect such problems to pose a challenge for NLI systems based on large language models (LLMs) as it has been shown that their predictions can be sensitive to a single-word substitution (Glockner et al., 2018; Gururangan et al., 2018). In addition to NPs, one could have allowed the replacement of other phrases in the NLI patterns, but this would have significantly complicated the definition of the mini world and generation of natural and sound NLI samples.

Property	E %	N %	C %	All % (#)	
Dir	39.6	35.4	25.0	30.0	(9600)
NonP	25.0	41.7	33.3	22.5	(7200)
Proj	29.4	26.5	44.1	21.2	(6800)
ArgO	47.6	28.6	23.8	26.2	(8400)
+ neg	48.0	28.0	24.0	15.6	(5000)
1prem	41.8	26.5	31.6	61.3	(19600)
2prem	25.0	42.9	32.1	35.0	(11200)
3prem	50.0	50.0	0.0	3.8	(1200)
All	36.2	33.1	30.6	100.0	(32000)

Table 3: Statistics of several properties of the sampled NLI dataset. The statistics also apply to the collection of NLI patterns as the samples are evenly distributed over the patterns. The properties consist of the spatial inference types, whether including negation, and the number of premises.

LLM-based NLI models	Training data	SNLI + MNLI	SpaceNLI Acc	$PA_{0.95}$	$PA_{1.0}$
DeBERTaV3-L#1 Joelzhang/deberta-v3…	SMFA	**91.8**	59.6	47.5	**37.5**
ALBERT-XXLv2 ynie/albert-xxlarge-v2…	SMFA	90.8	57.8	**48.1**	36.2
DeBERTa-L He et al. (2021)	M	90.7	54.1	42.5	36.2
RoBERTa-L Nie et al. (2020)	SMFA	90.6	55.6	40.0	31.9
BART-L ynie/bart-large-snli_mnli…	SMFA	90.4	55.4	39.4	29.4
DeBERTaV3-L#2 Laurer et al. (2022)	MFALW	90.3	**66.5**	44.4	33.8
XLNet-L-cased Nie et al. (2020)	SMFA	90.3	55.8	42.5	30.0

Table 4: Performance of SOTA NLI systems on SpaceNLI. SNLI+MNLI shows the average score on these datasets. Training data names are denoted with the initial letters: **S**NLI, **M**NLI, **A**NLI, **F**ever-NLI, **W**ANLI, and **L**ingNLI. The best system per problem accuracy on SpaceNLI, DeBERTaV3-L_{MFALW} (with $\Delta \geq 6.9\%$), doesn't turn out to be the best at the consistency threshold ≥ 0.95. Table 5 in Appendix A represents an extended version of the table with more threshold points.

4 Experiments

4.1 Sample dataset

We uniformly generated a spatial dataset of 32,000 NLI samples from 160 NLI patterns, i.e., 200 samples per pattern. We used the mini world as described in § 3.2. The dataset statistics are given in Table 3. The inference labels are relatively balanced: each label being represented by at least 30% of the problems. Each spatial inference type counts at least 20% of the overall problems and 23% of

label-specific problems. In contrast to the common biases in NLI datasets, a majority of the problems with negation are labeled as entailment, not contradiction. This is due to perturbed problems introduced in the revision phases (§ 3.1). Around 39% of problems have multiple premises, where three-premised problems occur only in the directional problems, the argument orientation problems contain only single-premised problems, and most of the multi-premised problems are in the non-projective problems. We refer to the generated dataset as SpaceNLI and use it in subsequent experiments.[7]

4.2 Evaluating SOTA NLI systems

4.2.1 Standard accuracy

We selected NLI models that have results comparable to the state of the art in NLI and evaluate them on SpaceNLI. The models were chosen based on their availability, tractable size, and high average accuracy ($> 90\%$) on the SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018) datasets (see Table 4). The models are based on various large language models (LLMs) like DeBERTaV3 (He et al., 2023), BART (Lewis et al., 2020), ALBERT (Lan et al., 2020), XLNet (Yang et al., 2020), etc. (see Table 4). The LLMs are fine-tuned on several NLI train datasets: SNLI, MNLI, FEVER-NLI (Nie et al., 2019), ANLI (Nie et al., 2020), LingNLI (Parrish et al., 2021), WANLI (Liu et al., 2022). We use the models from the HuggungFace model hub[8] and provide them with the corresponding hub names in Table 4.

The results in Table 4 show that DeBERTaV3-L#2 trained on a large collection of training datasets (885K problems in total) generalizes best on the spatial reasoning (66.5%), achieving a substantial improvement ($\geq 6.9\%$) over the other models.[9]

4.2.2 Consistency & pattern accuracy

To evaluate the models on the consistency of their predictions for NLI problems from the same pattern, we define the pattern accuracy (PA) score and its curve. The PA curve records the PA score of a model for each consistency threshold. Informally, the PA score with a consistency threshold t is a ratio of NLI patterns for which model gets at least t portion of the samples generated from them. For example, the PA of 50% with a threshold 90% means that there are a half of the NLI patterns such that for each pattern a model is able to correctly classify at least 90% of its sample problems. The formal definition of the PA with a threshold t is:

$$PA_t(\hat{Y}, \mathbf{y}) = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{\sum_{k=1}^{M_i} \delta(\hat{y}_k^i = y^i)}{M_i} \geq t \right]$$

where $\hat{Y} = (\hat{y}_k^i)_{1 \leq i \leq N, 1 \leq k \leq M_i}$ are predictions for k^{th} sample of i^{th} pattern, N is the number of patterns, M_i is the number of samples for i^{th} pattern, $\mathbf{y} = (y^i)_{1 \leq i \leq N}$ gold labels of i^{th} pattern, and δ is the Kronecker delta.

While DeBERTaV3-L#2 gets the best score on the SpaceNLI problems, based on the PA scores in Table 4, it shows high consistency ($PA_{0.95}$ or $PA_{1.0}$) in fewer NLI patterns than the other two competing models, DeBERTaV3-L#1 and ALBERT-XXLv2. PA curves of the NLI models provide a closer look at this contrast (see Figure 2). While the curve of DeBERTaV3-L#2 outperforms other models by a margin, it is noteworthy that it does this by classifying sample problems of the patterns which it can hardly solve half of the

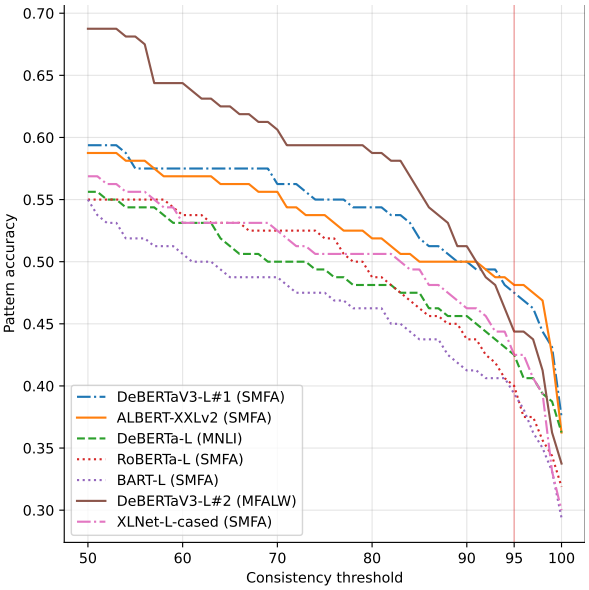

Figure 2: Pattern accuracy curves of the NLI models from Table 4. The first half, which corresponds to the scores allowing solving less than half of the samples per pattern, is omitted (see Figure 6 in Appendix A for the complete curves).

[7] We make the collection of the patterns, the generation code, and the sample dataset publicly available upon the acceptance of the paper.

[8] https://huggingface.co/models

[9] The second best, DeBERTaV3-L#1, is based on the same LLM fine-tuned on a different combination of NLI datasets. Note that Laurer et al. (2022) deliberately removed SNLI from the training set as it negatively affected the accuracy of the model in their experiments.

time (this is visible in the complete curves of Figure 6 in Appendix A). It drastically decreases after 95% of consistency while ALBERT-XXLv2 and DeBERTAV2-L#1 maintain very high consistency for $>$ 47% of NLI patterns. This demonstrates that a high-performing model is not necessarily the most consistent across patterns.

RoBERTa-L and BART-L obtain similar accuracy scores, but RoBERTa-L is more consistent in more NLI patterns than BART-L while the latter gets slightly more NLI problems for inconsistently predicted patterns. The complete curves of Figure 6 in Appendix A show how the curves swap places after the consistency threshold of 50. This shows that the standard accuracy (i.e., based on NLI problem samples) can blur the fine distinction in consistency between the models.

The dispersion of the curves at the lowest end of the consistency threshold is twice larger than at the highest end. This shows that the model predictions more diverge in coverage of patterns than in consistency per pattern. In other words, the contrast confirms the sensitivity of the models towards the inference-preserving word substitutions.

4.2.3 Few-shot learning experiments

We measured the difficulty of the SpaceNLI problems in terms of few-shot learning experiments. We used 100 samples per pattern as a test set while other 100 samples per pattern were used for drawing a few samples for each pattern. In this way, the patterns are fully shared between the training and test sets, but no sample NLI problem is in both sets. For each number of shots, we carried out the sample drawing process three times. We used two NLI models: a high performing NLI model RoBERTa-L$_{\text{SMFA}}$ from Nie et al. (2020) and a *vanilla* NLI model based on the large RoBERTa pretrained language model (Liu et al., 2019). The results of the few-shot experiments are in Figure 3.

Finetuning RoBERTa-L$_{\text{SMFA}}$ on a single sample of each pattern increases the sample-based accuracy on the test set by 14%. Each additional sample further boosts the model's accuracy. The almost perfect accuracy (>99%) is reached when 20 samples per pattern are seen during the finetuning. The results show that the lexical variability poses a challenge to the high-performing NLI model as it needs to be finetuned on at least five samples for every pattern of the test set to achieve a high score.

The challenge coming from the lexical variability and the SpaceNLI patterns is further emphasized by the relatively low results of RoBERTa Large. Even after being finetuned on the 20 samples of each NLI pattern, the model is still far from the high performance on unseen samples (but seen patterns). The relatively low results can be also partially attributed to the low ratio between the number of training samples and the large number of the model's trainable parameters.

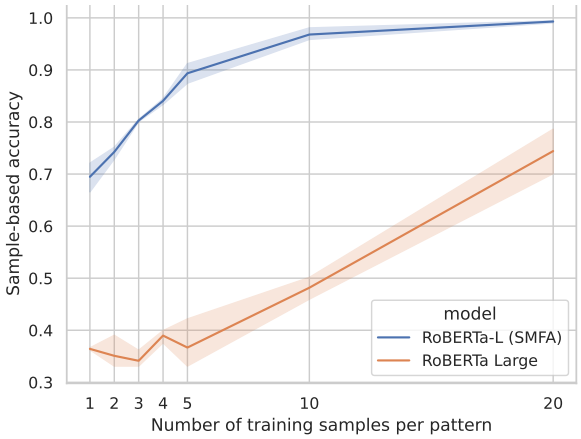

Figure 3: Average of three runs for each few-shot finetuning experiment. RoBERTa-L (SMFA, Nie et al. 2020) is already finetuned on several large NLI datasets while RoBERTa Large (Liu et al., 2019) is a pretrained language model without any previous training on NLI.

5 Analysis

To find out what type of inferences the models find challenging, we analyze the models' performance per inference type. Figure 5 shows the sample- and pattern-based accuracy scores of the models per spatial inference types as defined in §2.2. The model ranking based on the sample accuracy varies across the inference types. For instance, the best model, DeBERTaV3-L#2, remains at the top of the rankings for all inference types with quite a margin except for the projective type. On average, non-projective spatial inferences are the most challenging for the models. The easiest of the types is argument orientation, the type that is closest to the PP attachment task. For the other inference types, projective inferences are harder than directional ones. The apparent distinction in the scores between the inference types is also preserved for the $PA_{0.95}$ score (shown with the dark bars in Figure 5). The fine-grained analysis additionally shows that the best model, DeBERTaV3-L#2, suffers least in terms of consistency on the projective inferences while its performance on this inference type is not among the best.

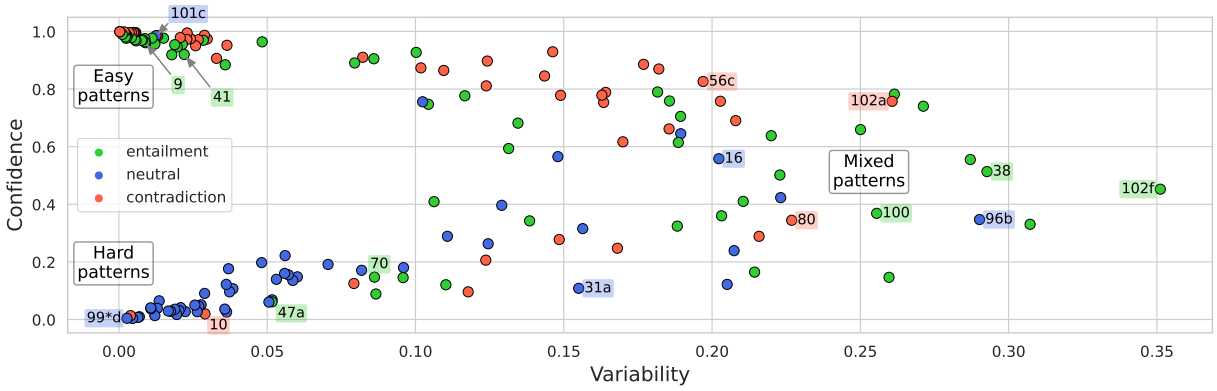

Figure 4: Prediction cartography of `RoBERTa-large` from (Nie et al., 2020). NLI patterns are characterized with *confidence* and *variability*: the mean and the standard deviation of probabilities assigned by the model to the true labels of the sample NLI problems. IDs mark NLI patterns from Figure 1 and Table 1.

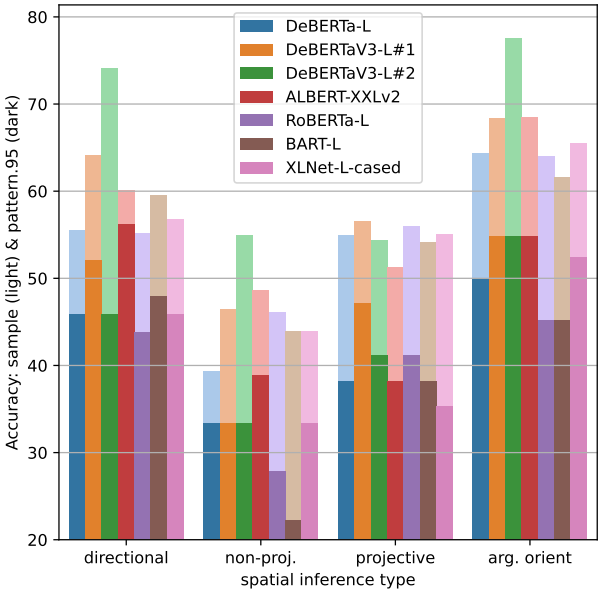

Figure 5: Sample-based (in light shades) and $PA_{0.95}$ (in dark shades) accuracy scores of the models per spatial inference type.

Based on the results in Figure 5, the non-projective NLI patterns and samples are the most challenging for the SOTA models. When looking closer at the set of non-projective problems, it turns out that it contains a high number of problems (46%) with the spatial expression "between" (as shown in Table 2), and these problems are specially challenging due to complex semantics of "between". The average accuracy of the models on such NLI samples is 41.6%. This is lower than the average sample-based accuracy (46.1%) on entire SpaceNLI and much lower than the average sample-based accuracy (54.1%) on the other part of the non-projective samples.

We further zoom in on the NLI patterns and measure a model's probabilistic predictions for the patterns. Namely, following Swayamdipta et al. (2020), we measure a model's confidence and variability. Originally the dataset cartography (Swayamdipta et al., 2020) was used to analyze the training dynamics of a model across the epochs and identify training samples that are easy or difficult for learning. In contrast, we use dataset cartography for analyzing evaluation dynamics across patterns and identifying easy and hard ones.[10]

Figure 4 illustrates the pattern-based evaluation dynamics of RoBERTa-L (Nie et al., 2020), an average model based on the evaluations. For instance, NLI pattern (102f) happens to have one of the most variable samples according to the model predictions: the mean and the standard deviation of the probabilities the model assigns to the entailment class of the samples of (102f) are 0.45 and 0.35, respectively.

(`102f`) NP_1 has hidden NP_2 behind NP_3.
`entailment` NP_2 is not in NP_3.

The evaluation cartography shows that the predictions vary mostly for entailment patterns (in green). Most of the hard patterns are neutral ones (in blue) and vice versa. Contradiction patterns (in red) tend to be easy with some variability.

6 Related work

Several works have automatically sampled NLI problems from curated patterns/templates. Jeretic et al. (2020) generated the implicature and presupposition diagnostic dataset IMPPRES from predefined templates. McCoy et al. (2019) constructed

[10] Put differently, iterative classification of the same training sample across epochs, is replaced with the classification of the same NLI pattern based on its samples.

the HANS dataset by designing templates of NLI problems that support or refute certain inference heuristics, which were later used to generate NLI problems. Richardson et al. (2020) used the template language from Salvatore et al. (2019) to produce NLI problems involving negation, Boolean connectives, quantifiers, cardinals, conditionals, and comparatives. These works all use restricted vocabulary while generating samples from the patterns.

With its pattern-based construction and restricted vocabulary, SpaceNLI comes close to the IMPPRES (Jeretic et al., 2020) and HANS (McCoy et al., 2019) datasets. Unlike these datasets, SpaceNLI involves multiple-premised problems and puts more emphasis on satisfying selection restrictions to prevent nonsensical sentences.

Based on the nature of NLI problems, SpaceNLI resembles FraCaS (Cooper et al., 1996) as both contain inference problems often found in textbooks on formal semantics. Unlike FraCaS, the inference labels of patterns in SpaceNLI are quite balanced and the number of spatial NLI patterns is twice the size of the largest section in FraCaS.

There have been attempts to identify semantic phenomena in existing NLI datasets, including aspects of spatial reasoning. By looking up certain keywords, Kim et al. (2019) automatically detect NLI problems in MultiNLI (Williams et al., 2018) that might contain spatial expressions. They create a mutated sample from the original NLI problem by negating the sentence with the potential spatial expression. Joshi et al. (2020) annotate MultiNLI problems based on the semantic aspects required by the inference label. Their taxonomic categories include the spatial subcategory, grouped with the relational, temporal, causal, and co-reference subcategories.

The problems in SpaceNLI are substantially diverse from a semantic perspective than the MultiNLI problems that were identified by Kim et al. (2019) and Joshi et al. (2020). The MultiNLI dataset is crowd-elicited and doesn't have problems with sufficient depth in spatial reasoning.

7 Conclusion

To the best of our knowledge, we have created the first spatial inference dataset that involves diverse spatial inference types. The structure and the evaluation protocol are unique as we focus on performance on the NLI patterns and consistency across the samples in the pattern, instead of focusing on mere quantitative accuracy based on the NLI problems/samples. The evaluation protocol tests models whether they can consistently recognize inference patterns while generalizing over *irrelevant* lexical substitutions. The more consistent a model is in its predictions, the less unexpected its behavior becomes.

The SOTA NLI models show moderate generalization capacity on spatial problems. While the top-performing model gets the highest overall accuracy, it is ranked third when it comes to the consistency of predictions inside the patterns: predicting at least 95% of the samples per pattern.

The introduced pattern accuracy (PA) curves provide a more fine-grained distinction between the models: the models with comparable standard accuracy scores might substantially differ in the consistency of their predictions. Overall the performance of models drops ca. 10% when raising the consistency threshold to 95%. This illustrates that the predictions of the SOTA models are sensitive to lexical replacements that have no effect on the semantics of the inference.

The evaluation results revealed that the most challenging inference type is associated with non-projective locatives mainly due to the complex semantics of "between" while the argument orientation type is the easiest. The latter is somewhat expected as the problems in the argument orientation type are close to the task of PP attachment which LLMs are expected to be good at.

Acknowledgments

This work was funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 742204). We would like to acknowledge the help from three student assistants with the data annotation and thank the anonymous reviewers for their helpful comments.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

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.

Robin Cooper, Dick Crouch, Jan Van Eijck, Chris Fox, Josef Van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, Steve Pulman, Ted Briscoe, Holger Maier, and Karsten Konrad. 1996. *FraCaS: A Framework for Computational Semantics*. Deliverable D16.

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, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention.

Annette Herskovits. 1986. *Language and Spatial Cognition: an interdisciplinary study of the prepositions in English*. Studies in Natural Language Processing. Cambridge University Press, London.

Md Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, and Eduardo Blanco. 2020. An analysis of natural language inference benchmarks through the lens of negation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9106–9118, Online. Association for Computational Linguistics.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6700–6709.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online. Association for Computational Linguistics.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Pratik Joshi, Somak Aditya, Aalok Sathe, and Monojit Choudhury. 2020. TaxiNLI: Taking a ride up the NLU hill. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 41–55, Online. Association for Computational Linguistics.

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)*.

Najoung Kim, Roma Patel, Adam Poliak, Patrick Xia, Alex Wang, Tom McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, Samuel R. Bowman, and Ellie Pavlick. 2019. Probing what different NLP tasks teach machines about function word comprehension. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 235–249, Minneapolis, Minnesota. Association for Computational Linguistics.

Thomas Kober, Sander Bijl de Vroe, and Mark Steedman. 2019. Temporal and aspectual entailment. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 103–119, Gothenburg, Sweden. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations.

Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2022. Less Annotating, More Classifying – Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT-NLI.

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.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. WANLI: Worker and AI collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. 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. Cite arxiv:1907.11692.

Bill MacCartney. 2009. *Natural language inference*. Phd thesis, Stanford University.

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

Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjamshidi. 2021. SPARTQA: A textual question answering benchmark for spatial reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4582–4598, Online. 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.

Seungho Nam. 1995. *The Semantics of Locative Prepositional Phrases in English*. Phd thesis, University of California, Los Angeles.

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.

Alicia Parrish, William Huang, Omar Agha, Soo-Hwan Lee, Nikita Nangia, Alexia Warstadt, Karmanya Aggarwal, Emily Allaway, Tal Linzen, and Samuel R. Bowman. 2021. Does putting a linguist in the loop improve NLU data collection? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4886–4901, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Kyle Richardson, Hai Hu, Lawrence S. Moss, and Ashish Sabharwal. 2020. Probing natural language inference models through semantic fragments. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8713–8721. AAAI Press.

Felipe Salvatore, Marcelo Finger, and Roberto Hirata Jr. 2019. A logical-based corpus for cross-lingual evaluation. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 22–30, Hong Kong, China. Association for Computational Linguistics.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. 2017. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223, Vancouver, Canada. Association for Computational Linguistics.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping

and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

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).

Jason Weston, Antoine Bordes, Sumit Chopra, and Tomás Mikolov. 2016. Towards ai-complete question answering: A set of prerequisite toy tasks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

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. Association for Computational Linguistics.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, and Kentaro Inui. 2020. Do neural models learn systematicity of monotonicity inference in natural language? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6105–6117, Online. Association for Computational Linguistics.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019a. Can neural networks understand monotonicity reasoning? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 31–40, Florence, Italy. Association for Computational Linguistics.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019b. HELP: A dataset for identifying shortcomings of neural models in monotonicity reasoning. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 250–255, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. Xlnet: Generalized autoregressive pretraining for language understanding.

Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017. Ordinal common-sense inference. *Transactions of the Association for Computational Linguistics*, 5:379–395.

Joost Zwarts and Yoad Winter. 2000. Vector space semantics: A model-theoretic analysis of locative prepositions. *Journal of logic, language and information*, 9:169–211.

A Results

LLM-based NLI models (train data) model names from Huggingface hub	SNLI	M_m	M_{mm}	S+M	SpaceNLI (accuracy & $\geq$ consistency score) Acc	≥ 0.5	≥ 0.67	≥ 0.9	≥ 0.95	$= 1.0$
DeBERTaV3-L#1 (SMFA) Joelzhang/deberta-v3-large-snli_mnli_fever_anli...	**92.9**	**91.4**	**91.2**	**91.8**	59.6	59.4	57.5	50.0	47.5	**37.5**
ALBERT-XXLv2 (SMFA) ynie/albert-xxlarge-v2-snli_mnli_fever_anli_R1_...	91.9	90.2	90.2	90.8	57.8	58.8	56.2	50.0	**48.1**	36.2
DeBERTa-L (MNLI) (He et al., 2021) microsoft/deberta-large-mnli	89.6	91.3	91.1	90.7	54.1	55.6	50.6	45.6	42.5	36.2
RoBERTa-L (SMFA) (Nie et al., 2020) ynie/roberta-large-snli_mnli_fever_anli_R1_R2_R...	91.8	89.9	90.0	90.6	55.6	55.0	52.5	43.8	40.0	31.9
BART-L (SMFA) ynie/bart-large-snli_mnli_fever_anli_R1_R2_R3-nli	92.0	89.4	89.6	90.4	55.4	55.0	48.8	41.2	39.4	29.4
DeBERTaV3-L#2 (MFALW) (Laurer et al., 2022) MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-l...	89.0	91.2	90.8	90.3	**66.5**	**68.8**	**61.9**	**51.2**	44.4	33.8
XLNet-L-cased (SMFA) (Nie et al., 2020) ynie/xlnet-large-cased-snli_mnli_fever_anli_R1_...	91.7	89.8	89.5	90.3	55.8	56.9	53.1	46.2	42.5	30.0

Table 5: Performance of NLI models on SpaceNLI and common NLI benchmarks: SNLI-test, MNLI-val-matched, and MNLI-val-mismatched. S+M shows the average of the three accuracy scores. Training data names are denoted with the initial letters: **S**NLI, **M**NLI, **A**NLI, **F**ever-NLI, **W**ANLI, and **L**ingNLI. The best model per problem accuracy on SpaceNLI, DeBERTaV3-L_{MFALW} (with $\Delta \geq 6.9\%$), doesn't turn out to be the best at the consistency threshold ≥ 0.95.

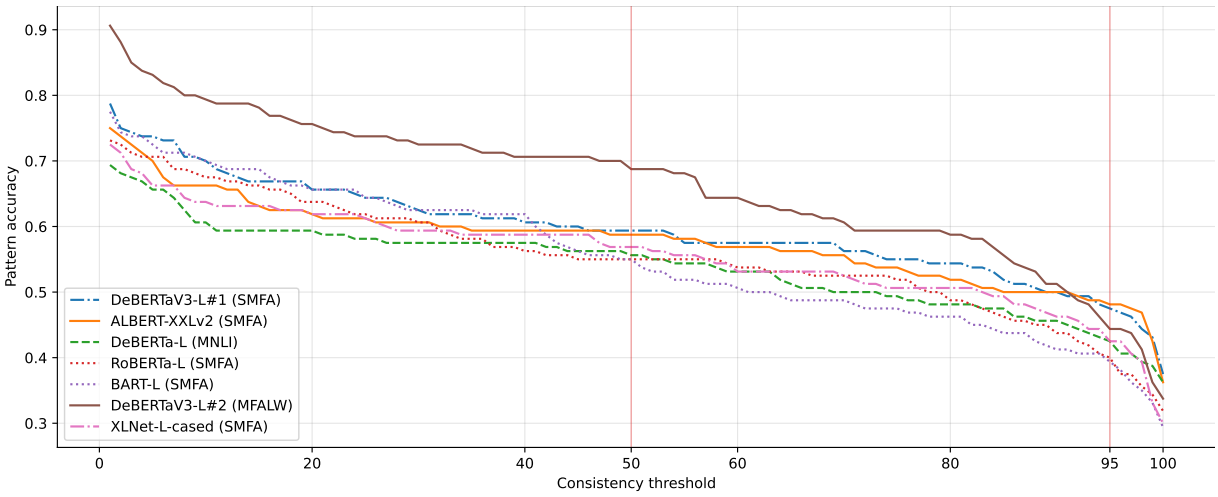

Figure 6: Pattern accuracy curves of the NLI models from Table 4. The area under the curve represents the standard accuracy based on the NLI problems.