A Philosophically-Informed Contribution to the Generalization Problem of Neural Natural Language Inference: Shallow Heuristics, Bias, and the Varieties of Inference

Reto Gubelmann
University of St.Gallen
Rosenbergstrasse 30
9000 St.Gallen

Christina Niklaus
University of St.Gallen
Rosenbergstrasse 30
9000 St.Gallen

Siegfried Handschuh
University of St.Gallen
Rosenbergstrasse 30
9000 St.Gallen

{reto.gubelmann, christina.niklaus, siegfried.handschuh}@unisg.ch

Abstract

Transformer-based pre-trained language models (PLMs) currently dominate the field of Natural Language Inference (NLI). It is also becoming increasingly clear that these models might not be learning the actual underlying task, namely NLI, during training. Rather, they learn what is often called bias, or shallow heuristics, leading to the problem of generalization. In this article, building on the philosophy of logics, we discuss the central concepts in which this problem is couched, we survey the proposed solutions, including those based on natural logic, and we propose or own dataset based on syllogisms to contribute to addressing the problem.

1 Introduction

Current natural language inference (NLI) is typically conceived as a three-way classification problem. With samples such as (1), consisting of a premise (P) and a hypothesis (H), the PLMs are tasked to categorize their relationship as either one of *contradiction* (P and H cannot both be true), of *entailment* (If P is true, then H must be true as well), or as being *neutral* (neither of the two).

(1) (P) The streets are wet. (H) It has rained.

As we will show below (see section 2), transformer-based pre-trained language models (PLMs) are currently the standard to approach this task of NLI. What is emerging as neural NLI's most pressing problem is the fact that these neural PLMs might almost outperform the crowdworker-based human baseline for the dataset on which they were fine-tuned, but perform worse than random at out-of-dataset-samples. We call this, following standard usage, the problem of generalization.

In this article, we focus on this problem of generalization, contributing a perspective that is informed by the philosophy of logic. More specifically, our article makes three contributions. First, after developing a conceptual background from the philosophy of logic, we give a comprehensive and systematic view on the extent of the problem of generalization in NLI, and we survey the different extant proposals to address this problem. Second, we propose and make publicly available a new fine-tune and challenge dataset that is based on syllogistic. Third, we evaluate the performance of both neural NLI models (including models fine-tuned on our syllogistic dataset) and a symbolic approach on this dataset. In the remainder of this section, we introduce the philosophical concept of inference.

The first and central distinction to be drawn regarding the concept of valid inference is the one between deductively valid inferences and defeasibly valid inferences (see Koons 2021 for an introduction to the distinction and to the concept of defeasible reasoning).[1] An inference is deductively valid if it is not possible that the premises are true while the conclusion is false (for the concept of necessity involved here, see Plantinga 1974, 1ff.). With defeasible inference, this condition does not hold: for such inferences, it is possible that the premises are true, while the conclusion is wrong. Example (1) is a case of defeasible inference: the streets could be wet, but this could have other causes than rain.

Within the domain of deductively valid inferences, it is common to distinguish inferences that are deductively valid due to the form of the propositions that constitute the inference, and others that are valid due to the content of these propositions (see Quine 1980 [1951] for a critical discussion of the distinction). Example (2) is a case of a formally deductively valid inference: It does not matter what

[1] For an early discussion of the distinction between deductively valid inferences, especially as opposed to conventional and conversational implicatures, see Zaenen et al. (2005).

you plug in for "Germans", "childcare workers", and "fingerprint collectors", you will always get a deductively valid inference (note that the truth of either premise or hypothesis is not required for an inference to be deductively valid. The concept of validity applies only to the truth-functional relationship between premise and hypothesis. A deductively valid inference with true premises is called a *sound* inference).

(2) (P) All Germans are childcare workers and all childcare workers are fingerprint collectors. (H) All Germans are fingerprint collectors.

In contrast, example (3) is deductively valid because of the content, the meaning of "bachelor" and "unmarried": replacing these concepts with others will likely result in an invalid inference.

(3) (P) Peter's marital status is that of a bachelor. (H) Peter is unmarried.

Formally valid inferences can further be classified according to the formal logical apparatus that is needed to prove its validity: propositional calculi, propositional calculi of different orders, and modal calculi are the most common options (see Smullyan 1968 for introductions to propositional and first-order logic, Garson 2006 for modal logic). Briefly, natural logic can be understood as the program to successively cover all of these areas without having to resort to translation into a formal language (for details, see section 2.2 and appendix, section C).

There are different proposals to systematize the domain of defeasible inferences. Currently, a prominent one is that defeasible inferences are inferences to the best explanation, that is, abductive inferences (for an excellent introduction to the concept, see Lipton 2004). Example (1) evinces the plausibility of this perspective: It is reasonable to conceive the hypothesis there as an explanation for the premise. The inference is defeasible because there could emerge a better explanation for the premise (in example (1), this could be the information that a street cleaning crew just passed through the street). An alternative conception is that such inferences are inductive in nature, that is, based on a number of previous observations of similar situations. Ever since Hume, it has been painfully clear that, without further metaphysical argument, such inductive inferences are not deductively valid. Figure 1 gives an overview on these kinds of valid inference.

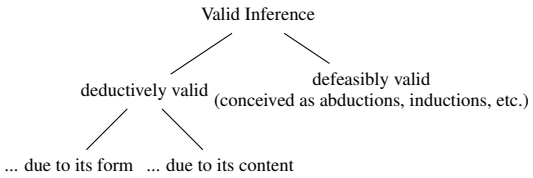

Figure 1: Kinds of valid inferences.

We will see that current practice oscillates between deductively and defeasibly valid inferences. Our own dataset focuses on the area of formal deductive validity.

To conclude our terminological survey, we mention that we will propose to distinguish between bias and shallow heuristics in a way suggested by Blodgett et al. (2020): We use bias as preconceptions that are potentially harmful, intrinsically normative, and always couched in a wider worldview. In contrast, a shallow heuristic is a local tactic to succeed at a given task without any understanding or mastery of the actual task that is explicitly not part of an intrinsically normative worldview.

2 State of the Art in Neural NLI

2.1 Neural NLI: Models & Datasets

In this section, we introduce the state of the art in neural NLI. As the focus of critical attention increasingly shifts to the datasets, we consider them in detail as well.

The Models Transformer-Based pre-trained language models (PLMs) have become the *de facto* standard in a variety of natural language processing tasks, including NLI. Based on the encoding part of the transformer (Vaswani et al., 2017), researchers have proposed a number of highly successful NLU architectures, starting with BERT (Devlin et al., 2019), quickly followed by others, including RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2019), DeBERTa (He et al., 2020), and smaller versions such as DistilBERT (Sanh et al., 2019) and Albert (Lan et al., 2019). Additionally, a number of sequence-to-sequence architectures have been proposed that are more similar to the original transformer than to BERT in that they directly try to transform one sequence to another, much like the basic set-up of neural machine translation. These include T5 (Raffel et al., 2019) and BART (Lewis et al., 2020).

These PLMs are then fine-tuned on specific

datasets, such as MNLI, which means that, while predicting labels on the dataset in question, a part of their parameters is being optimized. Fine-tuning usually takes several thousand times less computations than pre-training.

Such transformer-based PLMs fine-tuned to specific datasets perform impressively at standard natural language understanding (NLU) benchmarks, which include natural language inference (NLI) tasks. The MNLI Leaderboard, for instance, shows that the top ten PLMs are without exception transformer-based. Notably, in contrast to GLUE as a whole, the PLMs did not yet manage to outperform the human baseline at MNLI (as of June 15, 2022).

The Datasets Given the importance of fine-tuning for the entire method as it is currently practiced, it is clear that this method is squarely based on the availability – and quality – of large NLI datasets. Table 1 gives an overview on the currently most widely used datasets.[2]

Name of Dataset	Total Size	Genre
RTE (Wang et al., 2018)	6k	News, Wikipedia
QNLI (Wang et al., 2018)	116k	Wikipedia
WNLI (Wang et al., 2018)	852	hand-written
SICK (Marelli et al., 2014a)	9.8k	video & image captions
SNLI (Bowman et al., 2015)	570k	image captions
MNLI (Williams et al., 2018)	433k	10 genres, written & spoken

Table 1: Overview on Datasets used. Under "size", we report the total number of samples in train, test, and validation splits.

Thanks to their sheer size, SNLI and MNLI have come to dominate the field, as their size is suitable for fine-tuning large PLMs for NLI. As a consequence, as we shall see in the following section 2.2, most of the research on generalization issues focuses on these datasets.

There is a number of studies that critically assess the SNLI and MNLI datasets for their bias and thereby provides the groundwork for proposals following option 1 below (section 2.3). Williams et al. (2018) themselves note that their dataset contains a negation bias: if the hypothesis contains a negation, then it is more likely to be part of a contradiction pair (most likely, because simply negating the premise provides an efficient way for annotators to create contradiction pairs). Poliak et al. (2018) systematically investigate the prospects of hypothesis-only approaches (methods that only consider the hypothesis for predicting the label) to NLI in different datasets, finding better-than-random performance at most of them, which suggests the broad presence of statistical irregularities. Gururangan et al. (2018) show that SNLI and, to a lesser extent, MNLI, contain clues that make hypothesis-only approaches quite successful. Chien and Kalita (2020) focus on syntactic bias for PLMs fine-tuned on SNLI and MNLI, also finding that these bias are strong. Bernardy and Chatzikyriakidis (2019) argue that both SNLI and MNLI only cover a part of the entire range of human reasoning. In particular, they suggest that they do not cover quantifiers, nor strict logical inference.

The dataset that we will present in this study is intended to remedy both the lack of quantifiers and the lack of strict logical inference, given its focus on formally valid inferences.

Furthermore, we emphasize that, thanks to their near-ubiquitous use for fine-tuning, SNLI and to a greater extent MNLI determine the precise shape of the concept of inference that state-of-the-art models employ. On the one hand, the instructions given to crowdworkers are such that it seems reasonable to conclude that MNLI is about deductively valid inference: given a premise, crowdworkers are asked to "[w]rite a sentence that is definitely correct about the situation or event in the line [containing the premise]" (for the full instructions, see the appendix, section A). Requiring that the hypothesis be definitely correct given the correctness of the premise seems to require that it is not possible that the hypothesis could be false, given that the premise is true.

This reading, however, is contradicted by the fact that the creators of MNLI deliberately selected bits of text at random, not filtering for grammaticality, etc. These bits then served as prompts for the crowdworkers: they were tasked to write other bits of text for each prompt that either contracts, is entailed by, or is neutral vis-à-vis this prompt.

[2]Note that the RTE ("Recognizing textual entailment") dataset has been compiled from RTE1 (Dagan et al., 2005), RTE2 (Bar Haim et al., 2006), RTE3 (Giampiccolo et al., 2007), and RTE5 (Bentivogli et al., 2009). The QNLI ("Question-answering Natural Language Inference") dataset was created based on Rajpurkar et al. (2016). The WNLI ("Winograd Natural Language Inference") dataset was created based on Rahman and Ng (2012).

A consequence of the diversity of genres and this near-absence of preprocessing in MNLI is that the corpus contains premises such as (4).

(4) iuh-huh how about any matching programs

It is incoherent to say that questions entail any other statements: to entail something, a statement has to have determinate truth conditions; questions are textbook cases of sentences that have no determinate truth conditions. So, it is simply not possible for (4) to be part of any valid inference, let alone a deductively valid one. This issue stems from the very idea of MNLI, which is to represent the full variety of American English, using only minimal pre-processing.

Furthermore, crowdworkers are incentivized to produce large number of samples, which makes it rational to assume that a number of samples they produce are like example (5). Intended as a case of contradiction, it is clear that the premise does not contradict the conclusion in any logical sense: The speaker could simply have been lying, and no contradiction between premise and hypothesis would exist.

(5) (P) Oh, my friend, have I not said to you all along that I have no proofs. (H) I've always had the proof that he did it.

In sum, from a philosophical perspective, a qualitative inspection of the MNLI dataset shows that there might be some deeper problems in the set-up of the dataset. Furthermore, despite appearances to the contrary as per crowdworker-instructions, MNLI itself focuses on defeasible reasoning, that is, samples where the premise gives grounds to believe the hypothesis but does not entail it.

2.2 The Generalization Problem of Neural NLI

The basic problem that begins to emerge with this currently dominant approach to NLI is the problem of generalization. By this, we understand the inability of the PLMs to transfer the impressive performance on datasets on which they have been fine-tuned to out-of-dataset samples. Of course, a drop in performance is natural (even for humans) if the PLM is asked to perform the same task on substantially different data. If, however, the performance of a PLM simply collapses entirely when applied to out-of-dataset-samples, then it is *conceptually wrong* to say that the PLM has learned the task, namely correctly predicting logical relationships between statements, in the first place during fine-tuning: The task itself remains stable regardless of whether the samples are in or out of dataset. Together with the PLM's performance's lack of stability, this implies that it has learned something other than the task itself.

The problem of generalization in NLI is broadly acknowledged in the literature, see Zhou and Bansal (2020), Bras et al. (2020), Utama et al. (2020), Asael et al. (2021), He et al. (2019), Mahabadi et al. (2019), and Bernardy and Chatzikyriakidis (2019). It is generally assumed that the underlying cause of the problem of generalization is the PLMs' overfitting (see Goodfellow et al. 2016) on the training set. This overfitting, so the assumption goes, leads to the PLMs' picking up on spurious idiosyncrasies of the datasets, leading to the use of shallow heuristics and ultimately to a lack of generalization. Romanov and Shivade (2018) detail the generalization problems of pre-transformer PLMs in a highly specialized domain, namely medical history reports used by doctors.

If the models do not learn the central logical concepts during fine-tuning, what are they learning? The dominant view in the field is that they are learning so-called shallow heuristics, or bias: rules of thumb that work for the dataset due to some kind of bias in the data, but which do not apply to out of dataset samples, causing performance to collapse. In a much-discussed study, McCoy et al. (2019) conduct experiments to the conclusion that state-of-the-art PLMs use three kinds of syntactic heuristic at NLI tasks, which they call the lexical overlap, the subsequence, and the constituent heuristics. McCoy et al. (2019) also present a new stress test dataset called HANS ("Heuristic Analysis for NLI systems") that is built so that PLMs' use of the three heuristics will come to light in cases where the heuristics suggest entailment, but where the true label is not entailment.

2.3 Two Options to Address the Generalization Problem

In this section, we will consider the two main options that researchers have explored to address the problem of generalization in NLI.

Option 1: Debias the Dataset or the PLM The first option represents the mainstream of current thinking on NLI: It accepts the diagnosis that the models are merely picking up shallow heuristics

because there is a technical shortcoming in the method, and it tries to solve the problem by debiasing the datasets or the PLMs themselves. In table 2,[3] we list the papers, we mention whether their approach is based on a priori knowledge about the bias that one should tackle, and we report performance gains on the target dataset specified. Whenever available, we report performance gains on HANS, as this dataset has established itself as the *de facto* standard in the debiasing literature. As a consequence, these figures lend themselves best to comparisons between different approaches.

Paper	a priori knowl?	Target dataset	Acc.
He et al. (2019)	Yes	HANS	n.a.
Clark et al. (2019)	Yes	HANS	66.15 (+3.7)
Mahabadi et al. (2020)	Yes	HANS	71.95 (+10.1)
Yaghoobzadeh et al. (2019)	Yes	HANS	70.5 (+7.4)
Zhou and Bansal (2020)	Yes	Custom	+4.5
Belinkov et al. (2019)	Yes	Various	no gain
Dranker et al. (2021)	Yes	Various	no gain
Bras et al. (2020)	No	Various	+3.6
Utama et al. (2020)	No	HANS	69.7 (+8.2)
Nie et al. (2020)	(Yes)	Various	appr. +2
Bowman et al. (2020)	(Yes)	MNLI	no gain

Table 2: Overview on the extant approaches in option 1. Where no performance figures are given (n.a.), the paper doesn't report overall figures per dataset and it was not possible to extract these figures with simple, undisputable computations; "no gain" is a shorthand for "no significant gains".

Option 2: Hybrid Approaches Given the current generalization problem faced by purely neural approaches, some champions of symbolic methods have seen a chance to reinsert symbolic methods into the mainstream by combining neural with symbolic approaches. All current hybrid approaches rely on natural logic, an alternative to classical translation of natural language sentences into some formal langue. For details and references, see the appendix, section C. Hu et al. (2020) deliberately propose a lightweight, almost simplistic system that does not aim at setting a new state of the art, but rather at mapping out the lower bound performance of such a model. They explore its uses to provide training data for BERT.

An early approach at combining the two approaches is Raina et al. (2005). They combine classical formal logic with statistical learning for abductive reasoning (i.e., inference to the best explanation, a kind of non-monotonic inference, see (Lipton, 2004)).

Angeli and Manning (2014) introduce a seminal approach combining natural logic, monotonicity structures, WordNet and learned word probabilities as well as embeddings to conceive of NLI as a search problem. Kalouli et al. (2020) combine a classical symbolic system with a transformer-based neural PLM to achieve state-of-the-art performance on many standard datasets. Chen et al. (2021) adopt a different approach, conceiving of NLI as a path planning problem with the premise as the start and the hypothesis as the goal to be reached. They develop a system called NeuralLog that combines classical symbolic approaches using monotonicity notation (Hu et al., 2020) with, among others, Sentence-BERT embeddings to score the candidate hypotheses (Reimers and Gurevych, 2019). They report state of the art performance on both the SICK (Marelli et al., 2014b) as well as the MED (Yanaka et al., 2019) test sets; however, from among the neural approaches, they only consider BERT base. We report the results of these two only hybrid approaches post-HANS in table 3.

Paper	Target dataset	Acc.
Kalouli et al. (2020)	HANS	68.9 (+7.4)
Chen et al. (2021)	MED	93.4 (+21.8)

Table 3: Overview on the performance of the two most recent hybrid approaches. the MED dataset has been developed by Yanaka et al. (2018).

3 Dataset

In our experiment, we build on the insight gained from the qualitative inspection of MNLI as well as from research by Bernardy and Chatzikyriakidis (2019) that current NLI datasets lack samples that center on quantifiers as well as deductively valid inferences by providing a dataset that focuses on these very domains. Furthermore, our dataset provides a simple way to distinguish two properties of

[3]Note that the papers do not always report identical baseline performance, e.g., for BERT-base. We have reproduced these figures all the same, as the differences are small enough so that they do not affect our overall argument.

models that are often conflated: bias and shallow heuristics. As we have seen above (section 2.2), it is often said that the datasets or the models contain various biases. However, following Blodgett et al. (2020), we propose to use **bias** only for evaluations that are inherently normative and part of a larger worldview that is viewed critical. For instance, if a model expects that doctors are always men and therefore fails to correctly predict some logical relationships between sentences, one should attribute this to a bias: the model represents doctors as men, which is a clear case of a gender stereotype.

In contrast, a **shallow heuristic** is something that the models use irrespective of any such worldview, simply to succeed at a given task without fully learning it. The so-called negation bias is a clear case for such a shallow heuristic: It is not connected to any larger and problematic worldview but a simple instance of a rule of thumb.

While it has so far not been used to assess NLI capacities of NLU models, the systematic behind our dataset dates back to Aristotle. In his *Prior Analytics* (composed around 350 BC), Aristotle (1984, book 1) diligently analyzes the possible combinations of subject-, predicate-, and middle-term *via* quantifiers and negations to form a number of formally valid inferences. He deduces 24 formally valid patterns of inferences, so-called syllogisms. Example (2) is an instance of such a syllogism, belonging to the mood of the first figure that goes by the name of "BARBARA", the capital "A" signifying affirmative general assertions ("All X are Y.").

Now, consider the formal logical relationship in (6). By starting out with (2) and changing one single word, three letters in total, we have switched the relationship from entailment to contradiction.

(6) (P) All Germans are childcare workers and all childcare workers are fingerprint collectors. (H) No Germans are fingerprint collectors.

Finally, consider the formal logical relationship in (7). By changing one word, four letters, we switched the relationship from entailment to neutral.

(7) (P) All Germans are childcare workers and some childcare workers are fingerprint collectors. (H) All Germans are fingerprint collectors.

We are using a total of 12 formally valid syllogisms – called BARBARA, CELARENT, DARII, FERIO, CESARE, CAMESTRES, FESTINO, BAROCO, DISAMIS, DATISI, BOCARDO, FERISON – and we manually develop 24 patterns that are very similar to these 12 syllogisms, but where the first and the second sentence together contradict or are neutral to the third sentence. This yields a total of 36 patterns, 12 of which are valid syllogisms, 12 are contradictory, and 12 are neutral. To fit the premise-hypothesis structure expected by the models, we combine premise one and two to form a single premise.

We then use a pre-compiled list of occupations, hobbies, and nationalities to fill the subject- middle- and predicate-terms in these patterns. Using 15 of each of them and combining them with the 36 pattern yields 121500 test cases in total, each consisting of a premise and a hypothesis.[4] This variation allows us to capture the influence of any bias on model prediction, that is, any expectations of the models that certain nationalities are only likely to entertain certain hobbies and certain jobs, regardless of any valid inferences suggesting otherwise. Furthermore, it allows us to systematically distinguish it from shallow heuristics, rules of thumb that are not connected to any general worldviews or racial biases, but merely local attempts to succeed at the tasks without understanding it.

4 Experiment

We run a total of seven models on our test dataset, all of which are fine-tuned on standard NLI datasets, namely SNLI and MNLI (see table 4 for details: PLMs marked with one star "*" have only been fine-tuned on MNLI, PLMs marked with two stars have been fine-tuned on both SNLI and MNLI). The models are hosted by by Huggingface (Wolf et al., 2019), three of them are fine-tuned by Morris et al. (2020), prefixed with "textattack", and four by Reimers and Gurevych (2019), prefixed with "crossencoder".

The models' performances on MNLI, per our own evaluation (not all of the models provide evaluation scores, and we did not find precise documentation on how the scores were obtained), are given in table 4, for details of the evaluation, see the appendix, section B.

The basic idea behind the experiment is to assess

[4] The datasets can be found on the following github-repo: retoj/philo_nli.

PLM	N-Par.	MNLI-M
textattack-facebook-bart-large-MNLI*	406M	0.8887
crossencoder-deberta-base**	123M	0.8824
crossencoder-roberta-base**	123M	0.8733
crossencoder-MiniLM2-L6-H768**	66M	0.86602
textattack-bert-base-uncased-MNLI*	109M	0.8458
crossencoder-distilroberta-base**	82M	0.8364
textattack-distilbert-base-uncased-MNLI*	66M	0.8133

Table 4: Performance of the models in focus on the MNLI-Matched validation set. PLMs marked with one star "*" have only been fine-tuned on MNLI, PLMs marked with two stars have been fine-tuned on both SNLI and MNLI.

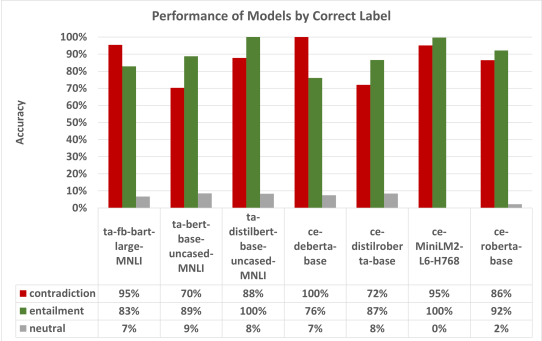

Figure 2: Performance on our syllogistic dataset by correct label.

whether the PLMs' performance on our dataset reveals any shallow heuristics learned by the models during fine-tuning on MNLI and SNLI.

5 Results

The results of our experiments are shown in figure 2. For instance, the model whose performance is represented on the very left, textattack's fine-tuned version of BART large, predicts the correct label in only 7% of cases for neutral labels, while doing so in 95% for entailment samples and still 83% for contradiction labels.

Figure 2 shows clearly that the models' predictions are quite accurate for labels *entailment* and *contradiction*, but very poor for *neutral*.

6 Discussion & Further Probes

6.1 Discussion of Experimental Results

Overall, figure 2 shows that textattack's distilbert leads the field with a accuracy of 65%, which might surprising just because it was among the smallest models evaluated here. However, there is growing evidence that NLI, and its more formal-deductive parts in particular, cannot be solved by simply increasing model size. Researchers at DeepMind find that larger models tend to generalize worse, not better, when it comes to tasks involving logical relationships. The large study by Rae et al. (2021, 23) strongly suggests that, in the words of the authors, "the benefits of scale are nonuniform", and that logical and mathematical reasoning does not improve when scaling up to the gigantic size of Gopher, a model having 280B parameters (in contrast, Gopher sets a new SOTA with many other NLU tasks such as RACE-h and RACE-m, where it outperforms GPT-3 by some 25% in accuracy).

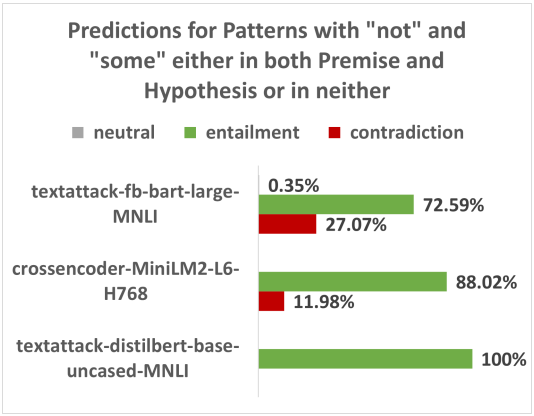

Figure 3: Predicted labels for patterns that are symmetric between premise and hypothesis regarding existential quantifier and negation.

Furthermore, figure 2 also shows that all of the models perform very poorly with neutral samples; indeed, none of the models is able to recognize such neutral relationships with a accuracy of more than 10%. Given that pure chance would still yield an accuracy of some 33%, this is a very poor performance.

We have therefore further probed the heuristics that the models might be using that could cause the poor performance with neutral labels. Manual inspection showed that they respond strongly to symmetries regarding quantifiers and negations between premises and hypotheses. In particular, if either both or none of the premise and the hypothesis contain a "some" (existential quantifier) or a negation (the symmetric conditions), then the models are strongly biased to predict *entailment* (see figure 3).

Conversely, if the pattern contains an asymmetry regarding existential quantifier and negation between premise and hypothesis, then the models are very strongly inclined to predict contradiction (see figure 4).

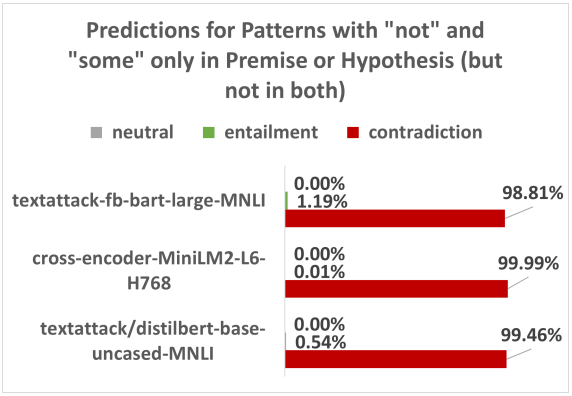

Figure 4: Predicted labels for patterns that are asymmetric between premise and hypothesis regarding existential quantifier and negation.

In the case of the contradiction and entailment pairs, these heuristics serve the models very well in our dataset, resulting in impressive performance. However, when applied to the neutral samples, the heuristics break down, performance falls far below simple guessing.

We conclude this part of our discussion by noting that the experiments did not show any significant bias in the behavior of the PLMs: Their accuracy did not change depending on existing preconceptions, say, that Germans are always engineers and like to collect stamps. What we have found, in contrast, is heavy use of shallow heuristics, as the figures 3 and 4 evince.

6.2 Fine-Tuning & Re-Evaluation, Comparison with a Symbolic Approach

In a next step, we assessed whether the models' poor performance with neutral samples in our dataset can be remedied with fine-tuning. We conducted two different fine-tuning runs, FT1 and FT2. Their sole difference consists in the way that we split up the 121k samples. For FT1, we used 110k samples for training and validation, and we tested on the neutral subset of the 10k remaining samples, which is about 3k samples ("3k" in figure 5). For FT2, we used 71k samples for training and validation, leaving the neutral subset of the remaining 50k samples, about 13k samples, for testing.[5]

We fine-tuned crossencoderMiniLM2-L6-H768 and textattack-distilbert-base-uncased-MNLI (BART-large from facebook exceeded our capacities). Furthermore, we also evaluated one of the currently leading symbolic NLI systems on both test datasets, namely GKR4NLI, introduced in (Kalouli et al., 2020). The results of all of these evaluations is shown in table 5.

Model	Neutr.	MNLI-M
FT1-crossencoder-MiniLM2-L6-H768	100%	72%
FT2-crossencoder-MiniLM2-L6-H768	62%	70%
FT1-textattack-distilbert-base-uncased-MNLI	100%	38%
FT2-textattack-distilbert-base-uncased-MNLI	61%	53%
GKR4NLI	89/23%	n.a.

Table 5: Accuracies of fine-tuned models and GKR4NLI on different test sets; For FT1-fine-tuned models, "Neutr." consists of 3k neutral samples from the syllogistic dataset, for FT2-fine-tuned models, it consists of 13k neutral samples from the same source. MNLI-M is MNLI-matched.

The results shown in table 5 show that fine-tuning does indeed help. In the first fine-tuning split FT1, both models achieve 100% accuracy; this, however, comes at rather high cost in terms of accuracy on MNLI-matched (14% and 43% respectively). GKR4NLI also performs well at this test set with 89% out of the box. With regard to the second fine-tuning split FT2, GKR4NLI's performance drops to 23%, while the two fine-tuned models achieve accuracies of around 62%, again at the cost of significantly reduced accuracy in MNLI.

These results suggest that it is not easy for the models tested to combine the representations needed to perform well at MNLI-matched with those needed to do well in our neutral samples. In particular, the results suggest that a large number of training samples is needed, as in FT1. We note that our results leave open the possibility that larger models can accommodate both kinds of sample.

At this point, we would like to compare our results with those obtained by Richardson et al. (2020). They use a cleverly chosen roaster of semantic fragments (i.e. subsets of a language translatable into formal logic, in particular first-order predicate logic) to test the models' understanding of the logical relationships of contradiction, entailment and neutral. They find that the models tested perform poorly on these tasks, but that this performance can be remedied with fine-tuning the models on sufficient amounts of training data that

[5]We adapted a huggingface-notebook found here, letting run each fine-tuning process for three epochs with a batch size of 16 on one GPU of a DGX-2.

has been synthetically generated from these fragments. In contrast to the semantic fragments used by Richardson et al. (2020), our datasets seem to pose a more difficult challenge for the models that we have tested (despite the fact that Richardson et al. 2020 only considered BERT-base, while we have also included larger and more recent models). Perhaps we have made some progress towards what Richardson et al. (2020) explicitly ask for, namely more difficult fragments?

We take these results to confirm that our dataset can make a valuable contribution to the field, as it presents a challenge for both neural and symbolic systems. Indeed, in light of these results, one could wonder whether it is not unfair to expect any NLI system to master our syllogistic dataset, as samples such as (2), (6), and (7) might be said to be very far away from ordinary language use. In response to this, we point out that, as a matter of logical fact, these are formally valid inferences which should be covered by any NLI system that aspires to cover the full extent of NLI. Furthermore, students of logics have acquired their concepts of formal validity through such examples for millennia, making it a rather natural stepping stone for AI systems. Finally, as already mentioned, it might very well be that large models could accommodate both the defeasible kinds of inferences in MNLI and our deductively valid ones.

7 Conclusion

We have detailed the problem of generalization that current neural approaches to NLI face from the background of philosophical logic. We have suggested that current datasets are light on deductively valid inferences, proposed a distinction between bias and shallow heuristics, and we have proposed our own syllogistic dataset. This dataset allows to distinguish between bias and shallow heuristic, it focuses on formally valid inferences, and our results suggest that it can help to improve both neural and symbolic approaches.

References

Gabor Angeli and Christopher D Manning. 2014. Naturalli: Natural logic inference for common sense reasoning. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 534–545.

Aristotle. 1984. Prior analytics. In Jonathan Barnes, editor, *The Complete Works of Aristotle*, pages 39–113. Oxford University Press.

Dimion Asael, Zachary Ziegler, and Yonatan Belinkov. 2021. A generative approach for mitigating structural biases in natural language inference. *arXiv preprint arXiv:2108.14006*.

R Bar Haim, I Dagan, B Dolan, L Ferro, D Giampiccolo, B Magnini, and I Szpektor. 2006. The second pascal rte challenge. *Proceedings of the 2nd PASCAL Challenge on RTE*.

Yonatan Belinkov, Adam Poliak, Stuart Shieber, Benjamin Van Durme, and Alexander Rush. 2019. Don't take the premise for granted: Mitigating artifacts in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 877–891. Association for Computational Linguistics.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The fifth pascal recognizing textual entailment challenge. In *TAC*.

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

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Johan Bos and Katja Markert. 2005. Recognising textual entailment with logical inference. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 628–635.

Johan Bos and Katja Markert. 2006. When logical inference helps determining textual entailment (and when it doesn't).

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*, pages 632–642. Association for Computational Linguistics (ACL).

Samuel R Bowman, Jennimaria Palomaki, Livio Baldini Soares, and Emily Pitler. 2020. New protocols and negative results for textual entailment data collection. In *EMNLP (1)*.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. *arXiv preprint arXiv:2002.04108*.

Zeming Chen, Qiyue Gao, and Lawrence S Moss. 2021. Neurallog: Natural language inference with joint neural and logical reasoning. *arXiv preprint arXiv:2105.14167*.

Tiffany Chien and Jugal Kumar Kalita. 2020. Adversarial analysis of natural language inference systems. *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*, pages 1–8.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4069–4082. Association for Computational Linguistics.

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

Ernest Davis. 2017. Logical formalizations of commonsense reasoning: a survey. *Journal of Artificial Intelligence Research*, 59:651–723.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yana Dranker, He He, and Yonatan Belinkov. 2021. Irm—when it works and when it doesn't: A test case of natural language inference. *Advances in Neural Information Processing Systems*, 34.

Gottlob Frege. 1892. Über sinn und bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 100:25–50.

James W Garson. 2006. *Modal logic for philosophers*. Cambridge University Press.

Gerhard Gentzen. 1935. Untersuchungen über das logische schließen. i. *Mathematische zeitschrift*, 35.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and William B Dolan. 2007. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*, pages 1–9.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*. MIT press.

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.

He He, Sheng Zha, and Haohan Wang. 2019. Unlearn dataset bias in natural language inference by fitting the residual. *arXiv preprint arXiv:1908.10763*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Hai Hu, Qi Chen, Kyle Richardson, Atreyee Mukherjee, Lawrence S Moss, and Sandra Kübler. 2020. Monalog: a lightweight system for natural language inference based on monotonicity. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 284–293.

Thomas F Icard III and Lawrence S Moss. 2014. Recent progress on monotonicity. In *Linguistic Issues in Language Technology, Volume 9, 2014-Perspectives on Semantic Representations for Textual Inference*.

Stanislaw Jaskowski. 1934. On the rules of suppositions in formal logic. *Studia Logica*, 1(1).

Aikaterini-Lida Kalouli, Richard Crouch, and Valeria de Paiva. 2020. Hy-NLI: a hybrid system for natural language inference. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5235–5249, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Robert Koons. 2021. Defeasible Reasoning. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Fall 2021 edition. Metaphysics Research Lab, Stanford University.

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. In *International Conference on Learning Representations*.

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. *ArXiv*, abs/1910.13461.

Vladimir Lifschitz, Leora Morgenstern, and David Plaisted. 2008. Knowledge representation and classical logic. *Foundations of Artificial Intelligence*, 3:3–88.

Peter Lipton. 2004. *Inference to the Best Explanation*, 2 edition. Routledge.

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. *arXiv preprint arXiv:1907.11692*.

Bill MacCartney and Christopher D Manning. 2007. Natural logic for textual inference. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 193–200.

Bill MacCartney and Christopher D Manning. 2008. Modeling semantic containment and exclusion in natural language inference. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 521–528.

Bill MacCartney and Christopher D Manning. 2009. An extended model of natural logic. In *Proceedings of the eight international conference on computational semantics*, pages 140–156.

Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2019. End-to-end bias mitigation by modelling biases in corpora. *arXiv preprint arXiv:1909.06321*.

Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2020. End-to-end bias mitigation by modelling biases in corpora. In *ACL*.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014a. The SICK (Sentences Involving Compositional Knowledge) dataset for relatedness and entailment.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, Roberto Zamparelli, et al. 2014b. A sick cure for the evaluation of compositional distributional semantic models. In *Lrec*, pages 216–223. Reykjavik.

John McCarthy. 1959. Programs with common sense. *Proceedings of the Symposium on Mechanization of Thought Processes*, (1).

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.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126.

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.

Francis Jeffry Pelletier and Allen Hazen. 2021. Natural Deduction Systems in Logic. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Winter 2021 edition. Metaphysics Research Lab, Stanford University.

Alvin Plantinga. 1974. *The Nature of Necessity*. Oxford University Press.

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. Association for Computational Linguistics.

Willard Van Orman Quine. 1980 [1951]. Two dogmas of empiricism. In *From a Logical Point of View*, pages 20–46. Harvard University Press.

Jack W. Rae, Sebastian Borgeaud, and Trevor Cai et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. DeepMind Company Publication.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Altaf Rahman and Vincent Ng. 2012. Resolving complex cases of definite pronouns: The Winograd schema challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 777–789. Association for Computational Linguistics.

Rajat Raina, Andrew Y Ng, and Christopher D Manning. 2005. Robust textual inference via learning and abductive reasoning. In *AAAI*, pages 1099–1105.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

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

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596. Association for Computational Linguistics.

Bertrand Russell. 1905. On denoting. *Mind*, 14(56):479–493.

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

Raymond M. Smullyan. 1968. *First-Order Logic*. Dover.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020. Towards debiasing nlu models from unknown biases. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7597–7610.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. 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. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Yadollah Yaghoobzadeh, Remi Tachet, Timothy J Hazen, and Alessandro Sordoni. 2019. Robust natural language inference models with example forgetting. *arXiv e-prints*, pages arXiv–1911.

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

Hitomi Yanaka, Koji Mineshima, Pascual Martínez-Gómez, and Daisuke Bekki. 2018. Acquisition of phrase correspondences using natural deduction proofs. 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 756–766. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763.

Annie Zaenen, Lauri Karttunen, and Richard Crouch. 2005. Local textual inference: can it be defined or circumscribed? In *Proceedings of the ACL workshop on empirical modeling of semantic equivalence and entailment*, pages 31–36.

Xiang Zhou and Mohit Bansal. 2020. Towards robustifying nli models against lexical dataset biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8759–8771.

A Full Instructions Given to Crowdworkers

Williams et al. (2018, 1114) specifies the following tasks for the crowdworkers:

"This task will involve reading a line from a non-fiction article and writing three sentences that relate to it. The line will describe a situation or event. Using only this description and what you know about the world:

- Write one sentence that is definitely correct about the situation or event in the line.
- Write one sentence that might be correct about the situation or event in the line.
- Write one sentence that is definitely incorrect about the situation or event in the line. "

B Method used for evaluation of Models on MNLI

To evaluate the models, we have used Huggingface's trainer API, see Huggingface (Wolf et al., 2019). In particular, we followed the instructions in the notebook here. We evaluated the models using the API out-of-the-box, with the following exceptions:

1. The textattack-models had as labels "LABEL_0, LABEL_1, LABEL_2", which could not be read by the function that ensures that the labels are used equivalently by both model and dataset; hence, we reconfigured the models to use as labels "contradiction, entailment, neutral".
2. facebook-bart-large-mnli by textattack posed two additional challenges.
   (a) Due to out of memory issues, we had to split up processing of the validation set into three chunks, averaging the accuracy received afterwards.
   (b) The logits containing the predictions issued by facebook-bart-large-mnli could not be processed by the evaluation function, which caused the need to select only the first slice of the tensor that the model was issuing, ensuring that the metric function got a 1-dimensional tensor to compute accuracy.

C From First-Order Representations to Natural Logic

Traditionally, the topic of common-sense reasoning, and later of NLI, as we understand it, was approached by the use of formal logic, predominantly first-order logic,[6] see Davis (2017) and Lifschitz et al. (2008) for extensive surveys of this approach, and McCarthy (1959) for the pioneering paper in this tradition. Bos and Markert (2005) and Bos and Markert (2006) are two typical cases in this tradition. In the latter, the authors find that, overall, adding logical processing to a shallow word-overlap approach actually hinders rather than boosts performance.

More recently, the once-dominant approach of representing premise and hypothesis in a formal language such as first-order predicate logic has been superseded by attempts to recover the logical structure of a sentence and the logical relationship between two sentences by directly annotating the natural language sentence. In particular, the so-called monotonicity calculus has been popular in a number of approaches. Icard III and Moss (2014) present an accessible and thorough review of recent theoretical work on this monotonicity approach.

The calculus stands in the tradition of natural logic (pioneered by Gentzen 1935 and Jaskowski 1934, for an overview, see Pelletier and Hazen 2021) is used by the NatLOG system developed by MacCartney and Manning (2007), MacCartney and Manning (2008), and MacCartney and Manning (2009). The basic idea behind the monotonicity calculus is to use low-level structural properties of quantifiers and predicates to assess the validity of an inference. For instance, the validity of the inference from "Every dog is a mammal" to "Every poodle is a mammal" is explained by a bottom-up combination of properties from the quantifier as well as the predicate involved – and not by translating the entire sentence into predicate calculus.

[6] As it has been pioneered by Frege (1892) and developed by Russell (1905).