Journal of Logic, Language and Information (2024) 33:21–48
https://doi.org/10.1007/s10849-023-09410-4

Capturing the Varieties of Natural Language Inference: A Systematic Survey of Existing Datasets and Two Novel Benchmarks

Reto Gubelmann[1] · Ioannis Katis[1] · Christina Niklaus[1] · Siegfried Handschuh[1]

Published online: 20 November 2023

Abstract
Transformer-based Pre-Trained Language Models currently dominate the field of Natural Language Inference (NLI). We first survey existing NLI datasets, and we systematize them according to the different kinds of logical inferences that are being distinguished. This shows two gaps in the current dataset landscape, which we propose to address with one dataset that has been developed in argumentative writing research as well as a new one building on syllogistic logic. Throughout, we also explore the promises of ChatGPT. Our results show that our new datasets do pose a challenge to existing methods and models, including ChatGPT, and that tackling this challenge via fine-tuning yields only partly satisfactory results.

Keywords NLI · Inference · Transformer · MNLI · Survey · ChatGPT

1 The Generalization Problem of Neural NLI & Kinds of Inference

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 models are tasked to categorize their relationship as either one of *contradiction* (P and H cannot both be true or are unlikely to both be true), of *entailment* (If P is

✉ Reto Gubelmann
reto.gubelmann@unisg.ch

Ioannis Katis
ioannis.katis@unisg.ch

Christina Niklaus
christina.niklaus@unisg.ch

Siegfried Handschuh
siegfried.handschuh@unisg.ch

[1] ICS, University of St. Gallen (HSG), St. Gallen, Switzerland

true, then H must be true as well or is likely true as well), or as being *neutral* (neither of the two).

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

While example (1) is an example for a syntactically rather simple, fallible, common-sense inference, example (2) falls on the other side of the spectrum: It is a syntactically complex, deductively valid inference that seems rather remote from common-sense. Despite their differences, however, both are examples for valid inferences (see Sect. 2.1 for a systematic representation of this concept). The true challenge of NLI is to develop methods that can cope with this diversity, perhaps by recognizing, as human experts do, what kind of inference is at issue in a given context and argument.

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

As we will show below (see Sect. 2.2), 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.

The work in this article can be seen as a contribution to addressing this problem of generalization. On our analysis, what contributes to the problem are the conceptions of inference inherent in the datasets that dominate current research: they tend to confine themselves to a part of the conceptual space spanned by the concept of inference, which means that models that have been trained on them struggle when tasked to cope with the phenomenon of inference in its full breadth. More specifically, our article makes three contributions. First, after detailing the concept of valid inference, we give a systematic view on the NLI datasets that are currently available, which allows us to identify two gaps in this dataset landscape, namely the lack of direct inductive inferences (such as example (1) and a scarcity of syntactically complex, quantifier-heavy samples (such as (2)).

Our second and third contributions aim at filling these gaps. Regarding the first gap, we introduce to the field of NLI a well-established dataset from the argumentative writing literature, which contains almost exclusively direct inductive inferences; we examine the performance of state-of-the-art models on this dataset. Third, regarding the second gap, we propose and make publicly available a fine-tuning and challenge dataset that is based on syllogistic logic (and therefore made up of syntactically complex, formally valid inferences), and we evaluate the performance of both neural NLI models and a symbolic approach on this dataset. We also evaluate ChatGPT on both the argumentative writing as well as our syllogistic dataset.

With regard to the overarching conceptual scheme, we propose to conceive NLI as a kind of common-sense reasoning. Common-Sense reasoning is sometimes seen as a use-case for formal logical methods (Davis & Marcus, 2015; Davis, 2017), and even more often as an umbrella term for cognitive tasks that require some world knowledge or common-sense knowledge (Storks et al., 2019; Trinh & Le, 2019; Zellers et al., 2018). Hence, on our proposal, NLI is a kind, but not the only kind, of common-sense reasoning.

2 Survey of the Current state of Research

In this survey, we first detail the different kinds of valid inference (Sect. 2.1), then we survey the current state of the art regarding models and datasets (Sect. 2.2), before discussing the problem of generalization in NLI (Sect. 2.3).

2.1 Kinds of Inference

A first and central distinction to be drawn within the concept of valid inference is the one between deductively valid inferences and inductively valid inferences (see (Koons, 2021) for an introduction to the distinction and to the concept of inductive, or 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 inductive inference, this condition does not hold: for such inferences to be valid, it is sufficient if the truth of the premise **gives good reason** to accept the truth of the conclusion (which means that it is possible, but unlikely, that the premise is true while the conclusion is wrong). Example (1) is a case of inductive 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) for a critical discussion of the distinction). Example (2) is a case of a formally deductively valid inference: It does not matter what 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).

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.

There are different proposals to systematize the domain of inductive inferences. Currently, a prominent one is that inductive 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

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

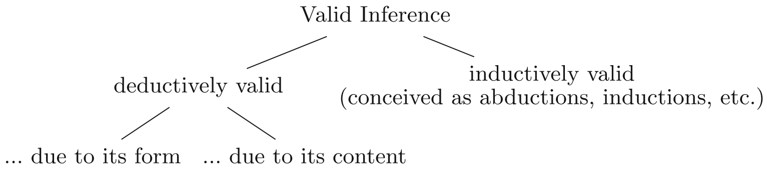


Fig. 1 Kinds of valid inferences

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 (1999), 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.

Having a clear conception of inference is obviously important for generalization. For instance, a model fine-tuned to identify deductively valid inferences might not excel at identifying inductively valid inferences: Pairs such as (1) should be labelled *entailment* if inductively valid inference is at issue, but *neutral* from the perspective of deductive validity. In other words, the validity of inductive inferences is simply invisible to a model trained on deductive inferences, which might suggest a low recall with such inferences. Furthermore, a model that is not used to inferences that are valid due to their (perhaps syntactically complex form) might struggle to accurately classify them.

2.2 Neural NLI: Models & Datasets

2.2.1 The Models

Transformer-based 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 the Multi-Genre Natural Language Inference (MNLI) dataset, 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 (https://paperswithcode.com/sota/natural-language-inference-on-multinli), for instance, shows that the top ten PLMs are without exception transformer-based.

With ChatGPT (OpenAI), a general-purpose chatbot trained by OpenAI, things are a little different. Due to OpenAI's non-disclosure of crucial parameters of the model and its training (and its refusal to publicly release the model), little details are known about it, and it is not possible to fine-tune it. Hence, we simply evaluate the version available via the API on April 13, 2023.

2.2.2 The Datasets

Given the importance of fine-tuning for the entire method of tackling NLI as it is currently practiced, it is clear that this method is squarely based on the availability – and quality – of large NLI datasets. Thanks to their sheer size, the Stanford Natural Language Inference (SNLI) datasets (Bowman et al., 2015) and MNLI (Williams et al., 2018) have come to dominate the field, as their size is suitable for fine-tuning large PLMs for NLI. For an example, see (4) (the example is from MNLI, SNLI is identically structured, with the main difference lying in the genres of the premises).

(4) 55785e (P) I burst through a set of cabin doors, and fell to the ground- (H) I burst through the doors and fell down. (entailment)

As a consequence of their popularity, as we shall see in the following Sect. 2.3, most of the research on generalization issues focuses on MNLI and SNLI. In contrast, our goal is to give a full picture of the variety of NLI datasets currently available. We try to give this overview on Table 1 (we dive deeper into these datasets in the appendix, Sect. A). We take the multitude of approaches that the field has developed in recent years to be a clear advantage: Human ability to draw logical inferences is a complex, multi-faceted ability, ranging from drawing strict deductive inferences to very implicit and fallible common-sense inferences. However, we also think that being aware of the different shades of the concept of inference is relevant for generalization issues. A system that is trained on a dataset representing deductive validity will classify inductively valid inferences as invalid, for instance and therefore have a low recall with these types of inferences. Note that not all of the datasets listed in table 1 are explicitly proposed as NLI datasets. We elaborate on this in the following.

What Qualifies as an NLI Dataset? We have included PIQA among the NLI datasets even though their authors (unlike the creators of (Hella)SWAG and SIQA) do not explicitly consider themselves as creating an NLI dataset. However, it seems helpful to group this carefully designed dataset together with its social-reasoning-focused counterpart, SIQA, and its frame-completing further counterpart (Hella)SWAG. PIQA is in good company here in requiring that subjects infer the most appropriate means to reach a given goal using their store of common-sense physical knowledge. The OBQA and ARC datasets are included here in accordance with the stated opinions of their authors in the papers.

Table 1 Matrix representation of current NLI datasets

	Deductive validity (Formal and material inf.)	Inductive validity (Only indirect found)
yes-no-questions	BoolQ (Clark et al., 2019), FraCas (Cooper et al., 1996)	
Classification (3-, 2- and 5-class)	SNLI (Bowman et al., 2015), MNLI (Williams et al., 2018), SICK (Marelli et al., 2014), PPDB 2.0 (Pavlick et al., 2015), RTE (Wang et al., 2018), WNLI (Wang et al., 2018)	
Multiple-choice	OBQA (Mihaylov et al., 2018), ARC (Clark et al., 2018)	SWAG, HellaSWAG (Zellers et al., 2018, 2019), PIQA (Bisk et al., 2020), SIQA (Sap et al., 2019)

Furthermore, and perhaps slightly more controversially, we have not included the QNLI dataset, which was designed by Wang et al. (2018) as an NLI benchmark as well as CB, which was included into SuperGLUE (Wang et al., 2019) as an NLI task (even though Marneffe et al. (2019), the creators of CB, do not see it as an NLI dataset). With QNLI, the reason for not including it as an NLI dataset is rather simple: It is one task to determine whether a given sentence can be logically inferred from another one, and quite another task to determine whether a given sentence answers a given question. This is why, incidentally, researchers such as Demszky et al. (2018) develop sophisticated transformation methods to convert question-answer pairs into premise-hypothesis pairs.

With the so-called commitment bank (CB), things are different. The problem here is that the entire dataset is centered around the notion of commitment. For instance, there just is no contradiction between “I don’t think they’ve ever made a movie, do you?" and “they’ve ever made a movie”, simply because somebody might very well think not-A while A is true (for more details on this dataset, see the appendix, Sect. 1). If I think that the earth is flat, while it is in fact round, then there is a conflict between my belief and common knowledge, but there is no contradiction between the sentence “I think that the world is flat” and the claim “The world is round”.

Deductive or Inductive Validity?

The BoolQ and FraCas dataset are explicitly designed with deductive validity in mind. With the RTE challenge dataset from Wang et al. (2018), careful qualitative inspection shows clearly that it represents a deductive notion of inference; example (5) is representative in this regard: Usually the entailment pairs have a hypothesis that is a paraphrase of the premise or one whose truth-conditions are a proper subset of the premises. The contradiction pairs usually contain a hypothesis that is incompatible with the premise.

(5) 85 This growth proved short-lived, for a Swedish invasion (1655-56) devastated the flourishing city of Warsaw. Warsaw was invaded by the Swedes in 1655, and the city was devastated. (entailment)

With SNLI and MNLI, the explicit commentary by the creators as well as the crowdworker-instructions have convinced us that they are best conceived as aiming at deductive validity. The instructions to the crowdworkers by Bowman et al. (2015) for SNLI require that, given a prompt, the workers write one caption that is "definitely a true description" (entailment), "might be a true description" (neutral), and a third one that is "definitely a false description" of the photo described by the prompt. The use of "definitely" as opposed to "highly plausibly" or similar clearly signals that the dataset to be created will represent a concept of inference close to standard deductive validity. Similarly, for MNLI, probably the most popular NLI dataset of our days, the crowdworkers are given a prompt and instructed to write three prompts, one that is "definitely correct", one that "might be correct", and one that is "definitely incorrect" about the situation in the prompt. Furthermore, Williams et al. (2018) state that, for their entailment labels, they are aiming at pairs where the hypothesis is "necessarily true or appropriate whenever the premise is true". This adds to the evidence that these two very influential datasets are squarely aiming at deductively valid inferences rather than inductive inference. We give the full crowdworker instructions in the appendix, Sect. A.15.

The OBQA and ARC datasets are open-book. This means that the correct answer can be inferred from collections of factuality statements that are provided with the challenge datasets. Looking at the specific examples provided by the authors of the dataset, it is clear that they, too are aiming at deductively valid inferences with these open-book question-answering challenges, see (Mihaylov et al., 2018, 2381) and (Clark et al., 2018, 6) respectively.

Direct or Indirect Inductive Inference?

Our decision to classify datasets such as (Hella)SWAG and P/SIQA as aiming at inductive validity through indirect reasoning might raise eyebrows. Our reasoning is as follows. The usual way to aim for inductively valid inferences is to go at it directly. This means to support a contested hypothesis directly with (defeasible) grounds, such as at the very beginning in example (1), where we defeasibly infer from the wetness of the streets that it has rained: To point to the wetness of the street directly (but defeasibly) supports the claim that it has rained.

In contrast, with the four datasets mentioned, the prompts never provide *direct* grounds to choose the appropriate answer from the choices. A (Hella)SWAG prompt directly triggers merely a certain frame in the competent language user, which then lends certain assumptions more plausibility than others. For instance (see the example (6)), if a woman walks on stage and takes a seat at the piano, this is probably because she wants to give a piano concert, a goal that makes options a)-c) very implausible.

(6) On stage, a woman takes a seat at the piano. She
a) sits on a bench as her sister plays with the doll. b) smiles with someone as the music plays. c) is in the crowd, watching the dancers. d) nervously sets her fingers on the keys.

This choice of option (d), however, is not supported directly by the initial prompt in the way the wetness of the streets directly supports the hypothesis that it has rained in example (1). Indeed, this is explicitly advertised as a strength of the datasets: They are intended to be particularly difficult for PLMs by requiring this indirect way of reasoning.

Similarly, with SIQA and PIQA, to understand why one option is more plausible than another, one has to be able to, as it were, think sideways. For instance, to understand that Remy will likely want to get her Room key in the example given in example (7) for the SIQA dataset, we have to understand that a concierge is usually checking in for somebody other than herself, that, given the frame that has been activated, this is probably Remy, and that you usually want to have a key as a main outcome of the check-in process.

(7) Remy gave Skylar, the concierge, her account so that she could check into the hotel. What will Remy want to do next? (a) lose her credit card (b) arrive at a hotel X (c) get the key from Skylar

In sum, it is remarkable that there are no datasets that focus on direct reasoning aiming at inductive validity such as example (1), and, to the best of our knowledge, no large datasets involving syntactically complex, formally valid inferences, such as example (2) (this is also pointed out by Bernardy and Chatzikyriakidis (2019)). We see two reasons for this.

Formally Simple Deductive inferences are cost-efficient First, writing up materially deductively valid inferences is much easier and hence faster and cheaper than writing up inductively valid inferences. Just exploit a simple conceptual hierarchy to materially infer, say, from the fact that a dog is playing outside the fact that an animal is playing outside. Similarly, if you want to create a contradiction, just negate the prompt and voilá, you've got yourselves a contradiction – a deductive one, of course. It would require much more reflection to come up with a counterclaim that seriously questions the prompt without downright contradicting it.

Inductive Inferences Should be Challenging Second, if researchers take the pains to create inductively valid pairs, they want it to have good chances at outsmarting the PLMs. This is antecedently more probable with the kind of indirect reasoning used in the datasets just discussed. It seems less promising to use direct reasoning, where less world knowledge and intuition seems to be required to fill the gap.

This results in two research gaps: (1) a scarcity of datasets and challenges for what might be the most mundane use case for NLI tout court: direct inductive inferences such as the very first example (1). (2) A scarcity of syntactically complex, formally valid inferences, such as example (2).

2.3 The Generalization Problem of Neural NLI & Kinds of Inference

The basic problem that has emerged 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 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. For an overview on the current approaches to mitigate this situation, see Gubelmann et al. (2022).

We suggest that the multi-facetedness of inferences, which has so far ben addressed only partially by current datasets, is another cause of the problem of generalization, and one that has been largely overlooked in the current debate. Building on this insight, we make two contributions to the ongoing research effort to overcome the problem of generalization. In the following two sections, we will present two datasets that are intended to contribute towards filling the two research gaps identified and hence towards solving the problem of generalization by acquainting the models with a more comprehensive conception of inference. First, the argumentative writing dataset provides a substantial number of direct inductive inferences (see Sect. 3). Second, the dataset that we will present in Sect. 4 is intended to remedy both the lack of quantifiers and the lack of complex, syntax-based inference-patterns, given its grounding in syllogistic logic.

In a sense, our two contributions fall onto the two extreme ends of the spectrum of valid inferences introduced above (Sect. 2.1). The argumentative writing dataset contains inferences of the kind of (1), while example (2) is directly taken from our syllogistic dataset. We hope that these datasets contribute to training scenarios that allow PLMs to implicitly represent the kinds of inference that exist and to learn to predict which of these kinds is at issue in a given context.

3 Argumentative Writing: Direct Inductive Inferences

Given the lack of direct inductive inference datasets, we hypothesize that *models fine-tuned on datasets that are representing a deductive concept of inference would have a low recall with entailment relations that are not deductively valid, but merely inductively valid*. The reason for this is that an inference that is inductively valid is often deductively invalid. Hence, when looking on it from the perspective of deductive validity, it would receive the label neutral. An analogous hypothesis can be proposed for contradiction.

3.1 Dataset

To investigate this hypothesis, we build on the argumentative writing literature. Arguably the most prominent dataset in this area is the one by Stab and Gurevych

(2014), expanded upon by Stab and Gurevych (2017). The dataset consists of annotations of 402 persuasive essays from students, taken from the online portal called essayforum. Overall, there are 6089 argument components annotated that are connected by 3832 relationships. Three annotators annotated a subset of 80 essays, the remaining 322 essays were annotated by just one annotator. For details regarding the annotation procedure and inter-annotator-agreement scores, see Stab and Gurevych (2017).

The dataset fits the bill because the arguments that are annotated are of the right type of inference, namely inductive inference, and because the creators of the dataset use a labelling structure that can easily be mapped onto ours. The relationships that are annotated between them are "support" and "attack". These are non-symmetric relationships that can hold between "Premises" and "Claims" as well as between "Claims" and "MajorClaims". These relationships of support and attack map nicely onto the NLI labels of entailment and contradiction, conceived inductively. Neutral labels are then simply any relationships between sentences that are not annotated with support and attack. Example (8) (Stab & Gurevych, 2014, 1504) shows a sample annotation for both attack and support.

(8) **Living and studying overseas is an irreplaceable experience when it comes to learn standing on your own feet.** *One who is living overseas will of course struggle with loneliness, living away from family and friends*$_1$ but *those difficulties will turn into valuable experiences in the following steps of life.*$_2$ Moreover, *the one will learn living without depending on anyone else.*$_3$

In this example, the claim in focus is the one put in boldface. According to the annotations, which are in a separate file, it is attacked by premise 1, which is in turn attacked by premise 2. Premise 3 then adds further support to the main claim. The example nicely shows that the kind of arguments in focus of this dataset is exactly the kind lacking in current NLI literature, namely cases of direct inductive reasoning. The entire reasoning in this example is fallible - it might be that somebody who studies overseas ends up with a highly protective aunt that leaves absolutely no room for personal development. Still, in general, the third premise directly supports the claim of the paragraph.

With the goal of examining the generalization abilities of NLI models fine-tuned on MNLI (as is currently the standard), we wanted to assess how well the models cope with the kind of annotations in the dataset. To this end, we have let three selected PLMs predict *any* relationship between any of the sentences in the dataset compiled by Stab and Gurevych (2017), in the following referred to as "AAE-DS". Taking the absence of an annotation as evidence for the neutral label yields a strong label imbalance, as most of the relationships between sentences in AAE-DS lack a label and are hence treated as neutral. More specifically, we obtain 54,657 neutral, 3259 entailment, and 156 contradiction pairs. While this imbalance is strong indeed, it is quite common among natural language texts, even in argumentative essays: given two sentences, there probably is no relationship existing between them. Therefore, in real-world use cases, this is the kind of label distribution that one could expect. Still, when diving

Table 2 Performance of the models on the argument annotated essays dataset by Stab and Gurevych (2017). We report accuracy for the overall performance

PLM (# Parameters)	Global		Entailment		Contradiction		Neutral	
	W-Acc	U-F1	Prec	Rec	Prec	Rec	Prec	Rec
ce-deberta-large (435 M)	87%	0.32	0.07	0.02	0.00	0.07	0.94	0.92
ta-fb-bart-large (406 M)	1%	0.06	0.06	0.95	0.00	0.01	0.96	0.04
ce-MiniLM2 (66 M)	86%	0.32	0.11	0.03	0.00	0.07	0.94	0.91
ChatGPT (175B 1k smpl.)	60%	0.51	0.51	0.5	0.33	0.34	0.69	0.69

deeper into the results, we will also look at label-specific figures, giving unweighted F1-scores in addition to weighted measures.

Due to financial limitations, we selected an evaluation dataset of 1'000 labels in total for ChatGPT, composed of 100 contradiction labels, 300 entailment labels, and 600 neutral labels. This is not the same label distribution as in the full AAE-DS, however, we wanted to have at least 100 samples per label and therefore had to compromise.

3.2 Experiment

We used the PLMs that are hosted by Huggingface (Wolf et al., 2019), one of them is fine-tuned by Morris et al. (2020), prefixed with "ta" for "textattack", and two by Reimers and Gurevych (2019), prefixed with "ce" for "crossencoder" ("facebook" is abbreviated by "fb"). We mapped predictions by the models in the way already suggested: *entailment* onto *support*, *contradiction* onto *attack* and neutral onto *(no relation)*. Details on dataset creation and relationship prediction can be found in the appendix, Sect. B.

3.3 Results & Discussion

The results of our experiments can be seen in Table 2. For crossencoder-PLMs, the results show hardly any drop in performance between MNLI-matched and our AAE-DS (with MNLI-matched, their accuracies are 87% and 88% respectively, see the Appendix, Table 5 for our own figures on MNLI-M): the variation is less than 2%. For textattack's bart-large, in contrast, the accuracy completely collapses to 1%.

The striking drop in performance by bart-large can be ascribed to a very poor recall (0.04) in the large neutral class, which is tantamount to much hallucination of contradiction and entailment relationships: its precision with these labels is at 0.06 and 0.01. One could of course argue that the strong label imbalance of our AAE-DS is the root cause of this very poor performance. However, we emphasize again that this situation is not artificially construed, but rather directly taken from a well-respected argumentative writing dataset.

The two ce-models, in contrast, have high recall and precision with neutral, which implies, due to label imbalance, high overall accuracy. Their precision and recall with the other two labels is, however, very low.

ChatGPT, finally, copes rather well with its 1k dataset, resulting in an accuracy of 60%. Taking a closer look at precision per label, we see that it performs better than the other models – with one important exception: the neutral class which makes up for the majority of labels. Its precision with contradiction and entailment is 0.3 and 0.5 respectively, implying that it is hallucinating much less than the other models. We have also run the other models on the 1k dataset that ChatGPT was tested on, and it turns out that ChatGPT outperforms the CE-Models by 3% in accuracy there. Note, again, that the label imbalance, while favoring the ce-models, is a realistic assumption.

With regard to our hypothesis, the figures are very encouraging: recall on entailment and contradiction relations is low with deberta and minilm2 – and it would likely also be low for BART-large if it would not hallucinate as much as it did. With ChatGPT, finally, recall is substantially higher than with the other models when looking at entailment and contradiction. However, the figures are still around.5 and thus still very low for a cutting-edge model performing a standard NLU task.

If we compare unweighted F1-score (column 3) instead of weighted accuracy (column 2), then ChatGPT performs considerably better: it leads the field, distancing both ce-models by nearly 0.2. This implies that ChatGPT would perform best with a dataset where the labels are balanced. However, even there, it only reaches an F1-score of 0.51 which is clearly still unsatisfactory.

We conclude therefore that the need for more direct inductive inferences as training data that we have identified *a priori* via our systematic Table 1 has been supported by our experiment. From the empirical data, we can directly, if only inductively (and hence defeasibly), infer that the models do not see the inductive validity of these inferences because they were fine-tuned mostly on deductive inferences.

4 The Syll-DS: Bias, Shallow Heuristics, and Formal Complexity

In our second experiment, we address the second research gap identified by the survey of NLI datasets (see Sect. 2.2), namely that there is currently a lack of large datasets that center on quantifiers as well as deductively valid inferences, by providing a dataset that focuses on these very domains. While we might not reach the syntactic sophistication of FraCas (Cooper et al., 1996), our dataset is still squarely focused on syntactically rather complex patterns of formal validity, large and hopefully contributes to filling the research gap identified.

Furthermore, our dataset provides a simple way to distinguish two properties of models that are often conflated: bias and shallow heuristics. As we have seen above (Sect. 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 usually viewed as potentially harmful. 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 and hence part of a larger 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. 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.

4.1 Dataset

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) in Sect. 1 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 (9). By starting out with (2) and changing one single word, three letters in total, we have switched the relationship from entailment to contradiction.

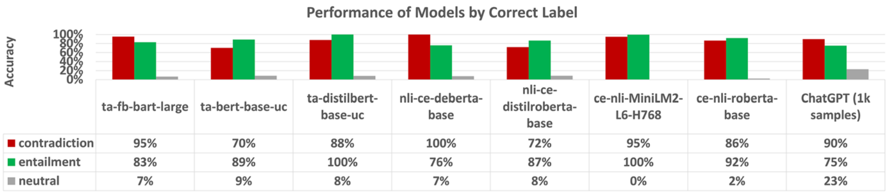
(9) (P) All Germans are childcare workers and all childcare workers are fingerprint collectors. (H) No Germans are fingerprint collectors.

Finally, consider the formal logical relationship in (10). By changing one word, four letters, we switched the relationship from entailment to neutral. Given the premise of (10), it is simply not clear whether the hypothesis is true or not.

(10) (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 patterns yields 121,500 test cases in total, each consisting of a premise and a hypothesis. 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



	ta-fb-bart-large	ta-bert-base-uc	ta-distilbert-base-uc	nli-ce-deberta-base	nli-ce-distilroberta-base	ce-nli-MiniLM2-L6-H768	ce-nli-roberta-base	ChatGPT (1k samples)
🟥 contradiction	95%	70%	88%	100%	72%	95%	86%	90%
🟩 entailment	83%	89%	100%	76%	87%	100%	92%	75%
▨ neutral	7%	9%	8%	7%	8%	0%	2%	23%

Fig. 2 Performance on our syllogistic dataset by correct label

general worldviews or racial biases, but merely local attempts to succeed at the tasks without understanding it.

We are also testing ChatGPT (with the model that is available via the API on April 13, 2023). Due to costs per inference, we were only able to test 1296 samples as opposed to the 121k samples for the other models. We highlight this by adding “1k samples” to all charts reporting ChatGPT’s results alongside the results by other models. The composition of this dataset as well as the precise prompts that we used can be found in the appendix, Sect. D.

4.2 Experiment

We run a total of seven freely available PLMs on our test dataset, all of which are fine-tuned on standard NLI datasets, namely SNLI and MNLI (see the Appendix, Table 5 for their respective performance on MNLI). Additionally, we also evaluate ChatGPT on a smaller dataset of 1296 samples. The PLMs are hosted by Huggingface (Wolf et al., 2019), three of them are fine-tuned by Morris et al. (2020), prefixed with “ta” for “textattack”, and four by Reimers and Gurevych (2019), prefixed with “ce” for “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 the appendix, Sect. C, Table 5, together with details of the evaluation. Performance ranges from 81% accuracy for distilbert to 89% accuracy for BART-large. We have not evaluated ChatGPT on MNLI.

The basic idea behind the experiment is to assess whether the PLMs’ performance on our dataset reveals any shallow heuristics learned by the models during fine-tuning on MNLI and SNLI.

The results of our experiments are shown in Fig. 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*, with ChatGPT being a bit of an outlier, as it is strongest on contradiction labels and achieves double-digit performance on neutral labels as well (albeit still below a purely random baseline).

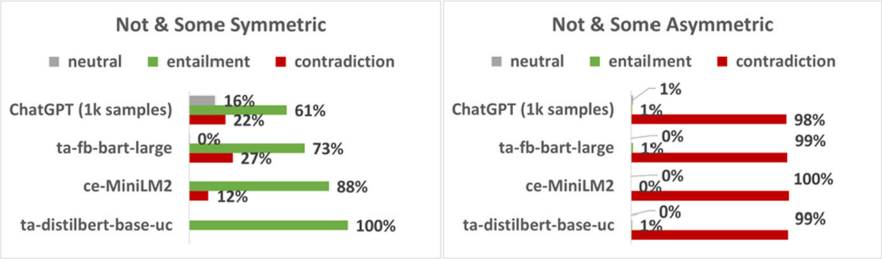


Fig. 3 Left: Predicted labels for patterns that are symmetric between premise and hypothesis regarding existential quantifier and negation. Right: Predicted labels for patterns that are asymmetric between premise and hypothesis regarding existential quantifier and negation

4.3 Results & Discussion

Overall, Fig. 2 shows that textattack's distilbert leads the field with a accuracy of 65%, which might be surprising just because it was among the smallest models evaluated here – and also because it even beats ChatGPT, which only achieves an overall accuracy of 63%. 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).

Furthermore, Fig. 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 25%, with ChatGPT being the only model that reaches double-digits. 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 Fig. 3, left chart). 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 Fig. 3, right chart). ChatGPT clearly also follows these heuristics, with the only slight difference that it is more inclined than the other models to predict neutral with the first of these two patterns.

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

Table 3 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. *For GKR4NLI, we report accuracies on the test datasets from FT1/FT2

Model	Neutr	MNLI-M
FT1-crossencoderMiniLM2-L6-H768	100%	72%
FT2-crossencoderMiniLM2-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

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 (also by ChatGPT), as the Fig. 3 evinces.

4.4 Fine-Tuning, Testing of 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. For obvious reasons, we had to exclude ChatGPT from this experiment. 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 Fig. 3). For FT2, we used 71k samples for training and validation, leaving the neutral subset of the remaining 50k samples, about 13k samples, for testing.[2]

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 3.

The results shown in Table 3 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.

[2] 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.

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 as well as for ChatGPT. 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), (9), and (10) 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. Perhaps we could see the difficulty as an asset: Maybe we have made some progress towards what Richardson et al. (2020) explicitly ask for, namely more difficult fragments? 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.

5 Conclusion

We have surveyed current NLI datasets and depicted the problem of generalization from the background of a systematic view on the kinds of inferences that exist. We have suggested that current datasets are light on direct inductive inferences as well as on syntactically complex, formally valid inferences. We have then proposed two steps towards addressing these research gaps. First, using a dataset from argumentative writing research, we could add empirical support to our hypothesis that there is a shortage of directly inductive inference datasets. To address the second research gap, we have proposed our own syllogistic dataset. This dataset allows to distinguish between bias and shallow heuristic, it focuses on syntactically complex, formally valid inferences, and our results suggest that it can help to improve both neural and symbolic approaches.

We have also found that, on both experiments, ChatGPT is outperformed by much smaller PLMs, adding further evidence to the hypothesis that larger models do not per se perform better with logical tasks (note, however, that the datasets used to evaluate ChatGPT are of limited size of about 1k). In the future, we would like to further analyze the hallucination phenomena that we have observed in the AAE-DS and work towards providing a large-scale dataset focusing on direct inductive generalization.

Acknowledgements Research for this article has been supported by The University of St. Gallen's basic research fund, grant number 2260281. This article builds on Gubelmann et al. (2022).

Author Contributions Conceptualization: Reto Gubelmann, Christina Niklaus; Methodology: Reto Gubelmann; Formal analysis and investigation: Reto Gubelmann, Ioannis Katis; Writing - original draft preparation: Reto Gubelmann; Writing - review and editing: Reto Gubelmann, Christina Niklaus; Funding acquisition: Siegfried Handschuh, Reto Gubelmann; Resources: Siegfried Handschuh; Supervision: Siegfried Handschuh.

Funding Open access funding provided by University of St. Gallen. See above, previous section

Data Availability After publication, datasets and scripts can be freely accessed under https://github.com/retoj/nli-generalization.

Declarations

Conflict of interest No conflicts of interest of any of the authors of this paper.

Ethics Approval No ethics approval is needed for this kind of research.

Consent for Publication The authors hereby consent to publication.

Code Availability See above, section Data Availability.

Appendix: A Further Details on NLI Datasets

Table 4 gives more information on the NLI datasets systematized above, Table 1. In the following, we comment on these datasets, highlighting what sets them apart from the well-known MNLI and SNLI datasets.

A.1 FraCas

The FraCas test suite consists of 346 challenges that are carefully hand-crafted following a systematic of a variety of different linguistic phenomena. For instance, the example (11) is exploiting disjunction distribution (Cooper et al., 1996, 116).

Table 4 Size & source of existing NLI datasets

Name	Size	Source(s)
FraCas (Cooper et al., 1996)	346	Hand-crafted
SICK (Marelli et al., 2014)	9.8k	Video & Image captions
PPDB 2.0 (Pavlick et al., 2015)	100 M/ 26k	Word-Based
SNLI (Bowman et al., 2015)	570k	image captions
RTE (Wang et al., 2018)	6k	News, Wikipedia
WNLI (Wang et al., 2018)	852	Hand-written
SWAG (Zellers et al., 2018)	113k	Activity Net Captions
HellaSWAG (Zellers et al., 2019)	70k	ActivityNet, WikiHow
MNLI (Williams et al., 2018)	433k	10 genres, written & spoken
BoolQ (Clark et al., 2019)	16k	Queries to Google Search
PIQA (Bisk et al., 2020)	21k	Crowdworker-Written guided by Instructables
SIQA (Sap et al., 2019)	38k	Crowdsourced
ARC (Clark et al., 2018)	7.8k	Science Textbooks
OBQA (Mihaylov et al., 2018)	6k	WorldTree (Jansen et al., 2018) facts, Crowdsourced Questions

(11) Smith saw Jones sign the contract and his secretary make a copy.
Did Smith see Jones sign the contract?

A.2 SICK

This dataset is based on image and video captions, which are normalized and then expanded to fit the three-way-classification scheme that has become the quasi-standard in NLI. See example (12).

(12) "The young boys are playing outdoors and the man is smiling nearby"
"The kids are playing outdoors near a man with a smile" 0 (entailment)

A.3 PPDB

This paraphrase database is different in two important ways from the other NLI datasets considered so far. First, it is word-based (as opposed to sentence-based); second, it differs from the typical three-way-classification in NLI detailed above, Sect. 1. Furthermore, it provides 26k hand-annotated word-pairs. The remaining 100 M word-pairs are obtained by training a regression classifier based on these 26k. Note that example (13) only shows a small part of the entire row dedicated to the relationship between transplant and transplantation. The final column contains the logical relationship (here: OtherRelated).

(13) [NN] ||| transplant ||| transplantation ||| [...] ||| OtherRelated

A.4 RTE

The RTE Dataset has been Compiled by Wang et al. (2018) from RTE1 (Dagan et al., 2005), RTE2 (Bar Haim et al., 2006), RTE3 (Giampiccolo et al., 2007), and RTE5 (Bentivogli et al., 2009). Compare example (14) for an illustration of the kind of pair found in this dataset.

(14) 85 This growth proved short-lived, for a Swedish invasion (1655-56) devastated the flourishing city of Warsaw. Warsaw was invaded by the Swedes in 1655, and the city was devastated. (entailment)

A.5 (Hella)SWAG

(Zellers et al., 2018) emphasize that their task is not typical inference but what they call grounded commonsense inference. As can be seen in example (15), the task consists not in assigning one of three labels to pairs of sentences, but in choosing one out of four options to continue a description of a scene. Hence, success at this task depends heavily on common-sense knowledge about how certain events usually develop. The authors cite the notion of object affordances developed by Gibson (2014) as well as frame semantics by Baker et al. (1998). As detailed above (Sect. 2.2), we suggest to conceive this challenge as centering around indirect inductive inferences. HellaSWAG is like SWAG, but with better adversarial filtering and longer sentences to make it harder for PLMs to succeed.

(15) On stage, a woman takes a seat at the piano. She
a) sits on a bench as her sister plays with the doll. b) smiles with someone as the music plays. c) is in the crowd, watching the dancers. d) nervously sets her fingers on the keys.

A.6 WNLI

In this challenge, which was created by Wang et al. (2018) based on Rahman and Ng (2012), the task consists in assigning two labels (1 and 0 in the datasets, mapping onto *entailment* and *non_entailment*) to pairs filtered from the original Winograd Challenge. As Winograd challenge is originally a pronoun resolution challenge, the correct label is entailment if the pronoun refers correctly and non_entailment otherwise. See example (16) for an illustration.

(16) 36 I tried to paint a picture of an orchard, with lemons in the lemon trees, but they came out looking more like telephone poles. The lemons came out looking more like telephone poles. 0

A.7 BoolQ

The BoolQ (Clark et al., 2019) dataset contains a text with yes-or-no (Boolean) questions whose correct answer can be deductively inferred from the text. For example, see (17)

(17) Q: Has the UK been hit by a hurricane? P: The Great Storm of 1987 was a violent extratropical cyclone which caused casualties in England, France and the Channel Islands A: Yes.

A.8 PIQA

The PIQA dataset consists of multiple-choice questions where the subjects have to choose the best means to reach a given goal. See example (18) for an illustration.

(18) [Goal] How do I find something I lost on the carpet? [Sol1] Put a solid seal on the end of your vacuum and turn it on. [Sol2] Put a hair net on the end of your vacuum and turn it on.

A.9 SIQA

The SIQA dataset consists of common-sense multiple-choice questions about social interactions. See example (19) for an illustration.

(19) Remy gave Skylar, the concierge, her account so that she could check into the hotel. What will Remy want to do next? (a) lose her credit card (b) arrive at a hotel X (c) get the key from Skylar

A.10 ARC

The arc dataset contains open-book grade-school-style multiple choice questions from the science domain. Open-book here means that the dataset comes with a knowledge base that contains the elements from which answers to all questions can be deductively inferred. For example, see (20).

(20) Which factor will prompt an animal's fight-or-flight response? (A) population size (B) competition for food [correct] (C) seasonal temperatures (D) protection of the environment

A.11 OBQA

The OBQA dataset is composed of 4-way-open-book multiple choice questions. Open-book here means that the dataset comes with a knowledge base that contains the elements from which answers to all questions can be deductively inferred. For example, see (21).

(21) Which of these would let the most heat travel through?
A) a new pair of jeans. B) a steel spoon in a cafeteria. C) a cotton candy at a store. D) a calvin klein cotton hat.

A.12 QNLI

The QNLI dataset was developed by the creators of the influential GLUE benchmark (Wang et al., 2018), who reference the work by Demszky et al. (2018). However, unlike the latter, Wang et al. (2018) simply extract the question and answer pairs from each of the samples in SQuAD (Rajpurkar et al., 2016) and apply some filtering for lexical overlap. The label *entailment* is then applied to all pairs where the answer to the question is present in the second sentence, *non_entailment* to all where it is not so present. Compare example (22). As a consequence, the notion of entailment at work here is not that of commonsense or strict logical entailment, but rather of semantic inclusion. It has therefore not been included in the overviews on Tables 1 and 4.

(22) 17 What percentage of farmland grows wheat? More than 50% of this area is sown for wheat, 33% for barley and 7% for oats. entailment

A.13 CommitmentBank

The 1.2k samples from the CommitmentBank (CB) corpus are not about entailment, but rather about the commitment of a speaker to the truth of an embedded claim. For instance, in example (23), the question is whether John's complex assertion commits him to the truth of the claim that Tess crossed the finish line, that it commits him to the contrary claim, or to none of both of them. Wang et al. (2019), the originators of SuperGLUE, mapped these commitment labels onto the three well-known labels of entailment, contradiction, and neutral, depending on whether the speaker is committed, is committed to the contrary claim, or is uncommitted.

(23) "premise": "A: I do too, so she couldn't possibly turn them out like some of these popular writers, B: Huh-uh. A: but oh, her books are just incredible. I don't think they've ever made a movie, do you?",
"hypothesis": "they've ever made a movie", "label": "contradiction"

A.14 Dataset Creation From Other Sources

There are a number of systematic proposals how to create datasets from other sources, see White et al. (2017), Chatzikyriakidis (2017) reflect about the possible way forward, Demszky et al. (2018) present a combination of rule-based and neural approach to convert question answering tasks in NLI tasks.

A.15 Instructions Given to Crowdworkers in MNLI

In the following, we quote in full how (Williams et al., 2018, 1114) specify the 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. "

Appendix B: Details on Experiment with Argumentative Writing Dataset

Algorithm 1 gives the algorithm for deriving the large dataset, algorithm 2 does the same for the 1k dataset used with ChatGPT, and algorithm 3 gives the procedure used to derive predictions.

Appendix C: Method used for evaluation of Models on MNLI

To evaluate the models used in the experiment from Sect. 4, we have used Huggingface's trainer API, see Huggingface (Wolf et al., 2019). In particular, we followed the instructions in the notebook here (https://colab.research.google.com/github/huggingface/notebooks/blob/master/transformers_doc/pytorch/training.ipynb). The results are given in Table 5. We evaluated the models using the API out-of-the-box, with the following exceptions:

Algorithm 1 Deriving NLI-Style sentence-pairs for three-way-classification from AAE-Annotations

```
$annotatedRelations \leftarrow load(relations)$    ▷ loading all existing annotations per text into a dictionary
$allRelations \leftarrow []$

for i in range(1, 403) do                                        ▷ Total of 403 essays
    $thisTextAnnotations \leftarrow annotatedRelations[i]$
    $essay \leftarrow load(essay_i)$
    $sentences \leftarrow spacy.splitSentences(essay)$    ▷ no such method exists - used for brevity

    for j in range(len(sentences)) do
        for k in range(j, len(sentences)) do
            $flag \leftarrow True$

            for annRel in thisTextAnnotations do
                if (annRel[1] in sentences[j] and annRel[2] in sentences[k]) or (annRel[1] in sentences[k] and annRel[2] in sentences[j]) then
                    $allRelations.append([sentences[j], sentences[k], annRel[0]])$
                                          ▷ annRel[0]: the existing, annotated label of the relation
                    $flag \leftarrow False$
                end if

                if flag then
                    $allRelations.append([sentences[j], sentences[k], neutral])$
                                          ▷ If there is no annotation: assign neutral
                end if
            end for
        end for
    end for
end for

$store(allRelations)$                ▷ Saving the relations (circa 58k) locally for reuse
```

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.

Algorithm 2 Deriving the 1k sample for ChatGPT-Probing

```
$allRelations \leftarrow load(allRelations)$        ▷ All relations stored from algorithm 1
$entailsIndices \leftarrow []$
$contradictsIndices \leftarrow []$
$neutralIndices \leftarrow []$

for i, line in enumerate(allRelations) do
    if line[2] == ’entails’ then
        $entailsIndices.append(i)$
    else if line[2] == ’contradicts’ then
        $contradictsIndices.append(i)$
    else if line[2] == ’neutral’ then
        $neutralIndices.append(i)$
    end if
end for

$randomIndices \leftarrow random.sample(entailsIndices, 300)$
$+ random.sample(contradictsIndices, 100) + random.sample(neutralIndices, 600)$    ▷ Creating
a list of 1k random indices

$sample \leftarrow []$
for randIndex in randomIndices do
    $sample.append(allRelations[randIndex])$
end for

$store(sample)$
```

Algorithm 3 Predicting

```
$data \leftarrow load(sample)$        ▷ Or load(allRelations)
$trueAndPredRelations \leftarrow []$

for line in data do
    $predLabel \leftarrow model.predict(line[0], line[1])$
    $trueAndPredRelations.append([line[0], line[1], line[2], predLabel])$
                    ▷ line[0], line[1] contains the sentences, line[2] the true label
                    ▷ For ChatGPT, we sent a prompt with the two sentences instead
end for

$store(trueAndPredRelations)$        ▷ Storing true and predicted labels for each model
```

Appendix D: Details on Prompting ChatGPT with the Syll-Dataset

To create the smaller dataset for probing ChatGPT, we started out with the same 36 syllogistic and pseudo-syllogistic patterns also used for the testing of the other PLMs, but we only expanded them using 4 nationalities (Gabonese, Georgians, Germans, Haitians) and 3 professions and hobbies (Clergys, Carpenters, Cashiers, Knife collectors, Films collectors, Element collectors). This yielded a total of 1296 samples, evenly distributed across the 36 patterns.

According to our information, the model that we were accessing through the API is GPT−3.5-turbo.[3]

The prompt used to get ChatGPT's verdict of the logical relationship between premise and hypothesis is as follows:

> In one word, is the relation between the sentences "sentence1" and "sentence2" an entailment, contradiction or neutral?

ChatGPT then promptly responded with one of the three classifying options given.

Table 5 Performance of the models in focus of the experiment in Sect. 4 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

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

References

Aristotle, J. B. (1984). Prior analytics. In J. Barnes (Ed.), *The complete works of aristotle* (pp. 39–113). Oxford University Press.

Asael, D., Ziegler, Z., & Belinkov, Y. (2021). A generative approach for mitigating structural biases in natural language inference. arXiv preprint arXiv:2108.14006

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The berkeley framenet project. In COLING 1998 volume 1: The 17th international conference on computational linguistics.

Bar Haim, R., Dagan, I., Dolan, B., Ferro, L., Giampiccolo, D., Magnini, B., & Szpektor, I. (2006). The second PASCAL RTE challenge. In Proceedings of the 2nd PASCAL challenge on RTE.

Bentivogli, L., Clark, P., Dagan, I., & Giampiccolo, D. (2009). The 5th PASCAL recognizing textual entailment challenge. In TAC.

Bernardy, J.-P., & Chatzikyriakidis, S. (2019). What kind of natural language inference are NLP systems learning: Is this enough? In ICAART (2) (pp. 919–931).

Bisk, Y., Zellers, R., & Le bras, R., Gao, J., & Choi, Y. (2020). PIQA: Reasoning about physical commonsense in natural language. In Proceedings of the AAAI conference on artificial intelligence (vol. 34, pp. 7432–7439). https://doi.org/10.1609/aaai.v34i05.6239

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (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 (pp. 5454–5476). Association for Computational Linguistics, Online. https://doi.org/10.18653/v1/2020.acl-main.485

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In Conference on empirical methods in natural language processing, EMNLP 2015 (pp. 632–642). Association for Computational Linguistics (ACL)

[3] See here: https://openai.com/blog/introducing-chatgpt-and-whisper-apis, consulted on April 14, 2023, the day of the experiment with ChatGPT.

Chatzikyriakidis, S., Cooper, R., Dobnik, S., Larsson, S. (2017) An overview of natural language inference data collection: The way forward? In: Proceedings of the Computing Natural Language Inference Workshop

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., & Tafjord, O. (2018) Think you have solved question answering? try ARC, the AI2 reasoning challenge. arXiv.

Clark, C., Lee, K., Chang, M. -W., Kwiatkowski, T., Collins, M., & Toutanova, K. (2019). BoolQ: Exploring the surprising difficulty of natural yes/no questions. arXiv.

Cooper, R., Crouch, D., Van Eijck, J., Fox, C., Van Genabith, J., Jaspars, J., Kamp, H., Milward, D., Pinkal, M., Poesio, M., et al. (1996). Using the framework. Technical report, Technical Report LRE 62-051 D-16, The FraCaS Consortium.

Dagan, I., Glickman, O., & Magnini, B. (2005). The pascal recognising textual entailment challenge. In Machine learning challenges workshop (pp. 177–190). Springer.

Davis, E. (2017). Logical formalizations of commonsense reasoning: A survey. *Journal of Artificial Intelligence Research, 59*, 651–723.

Davis, E., & Marcus, G. (2015). Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM, 58*(9), 92–103. https://doi.org/10.1145/2701413

de Marneffe, M. -C., Simons, M., & Tonhauser, J. (2019). The CommitmentBank: Investigating projection in naturally occurring discourse. In Proceedings of Sinn Und Bedeutung.

Demszky, D., Guu, K., & Liang, P. (2018). Transforming question answering datasets into natural language inference datasets. arXiv preprint arXiv:1809.02922

Devlin, J., Chang, M. -W., Lee, K., & Toutanova, K. (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 (vol. 1, pp. 4171–4186). Association for Computational Linguistics, Minneapolis, Minnesota. https://doi.org/10.18653/v1/N19-1423

Giampiccolo, D., Magnini, B., Dagan, I., & Dolan, W. B. (2007). The 3rd pascal recognizing textual entailment challenge. In Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing (pp. 1–9).

Gibson, J. J. (2014). *The ecological approach to visual perception (Classic)*. Psychology Press.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Gubelmann, R., Niklaus, C., & Handschuh, S. (2022). A philosophically-informed contribution to the generalization problem of neural natural language inference: Shallow heuristics, bias, and the varieties of inference. In Proceedings of the 3rd natural logic meets machine learning workshop (NALOMA III) (pp. 38–50). Association for Computational Linguistics, Galway, Ireland.

He, P., Liu, X., Gao, J., & Chen, W. (2020). DEBERTA: Decoding-enhanced bert with disentangled attention. In International conference on learning representations.

He, H., Zha, S., & Wang, H. (2019). Unlearn dataset bias in natural language inference by fitting the residual. arXiv preprint arXiv:1908.10763

Hume, D. (1999). *An enquiry concerning human understanding*. Oxford University Press.

Jansen, P., Wainwright, E., Marmorstein, S., & Morrison, C. (2018). WorldTree: A corpus of explanation graphs for elementary science questions supporting multi-hop inference. In Proceedings of the 11th international conference on language resources and evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan. https://aclanthology.org/L18-1433

Kalouli, A. -L., Crouch, R., & de Paiva, V. (2020). Hy-NLI: A hybrid system for natural language inference. In Proceedings of the 28th international conference on computational linguistics (pp. 5235–5249). International Committee on Computational Linguistics, Barcelona, Spain (Online). https://doi.org/10.18653/v1/2020.coling-main.459

Koons, R. (2021). Defeasible reasoning. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (2021st ed.). Metaphysics Research Lab, Stanford University.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). LBERT: A lite BERT for self-supervised learning of language representations. In International conference on learning representations.

Le Bras, R., Swayamdipta, S., Bhagavatula, C., Zellers, R., Peters, M., Sabharwal, A., & Choi, Y. (2020). Adversarial filters of dataset biases. In International conference on machine learning (pp. 1078–1088). PMLR

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L.(2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. ArXiv **abs/1910.13461**

Lipton, P. (2004). *Inference to the best explanation* (2nd ed.). Routledge.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692

Mahabadi, R.K., Belinkov, Y., & Henderson, J. (2019). End-to-end bias mitigation by modelling biases in corpora. arXiv preprint arXiv:1909.06321

Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., & Zamparelli, R. (2014). A SICK cure for the evaluation of compositional distributional semantic models. In: Lrec, pp. 216–223. Reykjavik, ???

Mihaylov, T., Clark, P., Khot, T., Sabharwal, A. (2018). Can a suit of armor conduct electricity? A new dataset for open book question answering. In Proceedings of the 2018 conference on empirical methods in natural language processing (pp. 2381–2391). Association for Computational Linguistics, Brussels, Belgium. https://doi.org/10.18653/v1/D18-1260

Morris, J., Lifland, E., Yoo, J.Y., Grigsby, J., Jin, D., & Qi, Y. (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 (pp. 119–126).

OpenAI: ChatGPT.

Pavlick, E., Rastogi, P., Ganitkevitch, J., Van Durme, B., & Callison-Burch, C. (2015). PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (vol. 2, pp. 425–430). Association for Computational Linguistics, Beijing, China. https://doi.org/10.3115/v1/P15-2070

Plantinga, A. (1974). *The nature of necessity*. Oxford University Press.

Quine, W. V. O. (1980). Two dogmas of empiricism. *From a logical point of view* (pp. 20–46). Harvard University Press.

Rae, J.W., Borgeaud, S., & al, T.C. (2021). Scaling language models: Methods, analysis & insights from training gopher.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:1910.10683

Rahman, A., & Ng, V. (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 (pp. 777–789). Association for Computational Linguistics.

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. In Proceedings of the 2016 conference on empirical methods in natural language processing, pp. 2383–2392. Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1264

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084

Richardson, K., Hu, H., Moss, L., & Sabharwal, A. (2020). Probing natural language inference models through semantic fragments. In Proceedings of the AAAI conference on artificial intelligence (vol. 34, pp. 8713–8721).

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108

Sap, M., Rashkin, H., Chen, D., Le Bras, R., & Choi, Y. (2019). Social IQa: Commonsense reasoning about social interactions. 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) (pp. 4462–4472). Association for Computational Linguistics, Hong Kong, China. https://doi.org/10.18653/v1/D19-1454

Stab, C., & Gurevych, I. (2014). Annotating argument components and relations in persuasive essays.

Stab, C., & Gurevych, I. (2017). Parsing argumentation structures in persuasive essays. *Computational Linguistics, 43*(3), 619–659. https://doi.org/10.1162/COLI_a_00295

Storks, S., Gao, Q., & Chai, J.Y. (2019). Commonsense reasoning for natural language understanding: A survey of benchmarks, resources, and approaches (pp. 1–60). arXiv preprint arXiv:1904.01172

Trinh, T.H., & Le, Q.V. (2019). A Simple method for commonsense reasoning. arXiv.

Utama, P. A., Moosavi, N. S., & Gurevych, I. (2020). Towards debiasing NLU models from unknown biases. In Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP) (pp. 7597–7610).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 1.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (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 (pp. 353–355). Association for Computational Linguistics, Brussels, Belgium. https://doi.org/10.18653/v1/W18-5446

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019). SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *Advances in Neural Information Processing Systems, 32*, 1.

White, A. S., Rastogi, P., Duh, K., & Van Durme, B. (2017). Inference is everything: Recasting semantic resources into a unified evaluation framework. In Proceedings of the 8th international joint conference on natural language processing (vol. 1, pp. 996–1005). Asian Federation of Natural Language Processing, Taipei, Taiwan. https://aclanthology.org/I17-1100

Williams, A., Nangia, N., & Bowman, S. (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 (vol. 1, pp. 1112–1122). Association for Computational Linguistics, New Orleans, Louisiana.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q., & Rush, A.M. (2019). HuggingFace's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems, 32*, 1.

Zaenen, A., Karttunen, L., & Crouch, R. (2005). Local textual inference: Can it be defined or circumscribed? In Proceedings of the ACL workshop on empirical modeling of semantic equivalence and entailment (pp. 31–36).

Zellers, R., Bisk, Y., Schwartz, R., & Choi, Y.(2018). SWAG: A large-scale adversarial dataset for grounded commonsense inference. arXiv.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., & Choi, Y. (2019). HellaSwag: Can a machine really finish your sentence? arXiv.

Zhou, X., & Bansal, M. (2020). Towards robustifying NLI models against lexical dataset biases. In Proceedings of the 58th annual meeting of the association for computational linguistics (pp. 8759–8771).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.