

When Truth Matters – Addressing Pragmatic Categories in Natural Language Inference (NLI) by Large Language Models (LLMs)

Reto Gubelmann

University of St. Gallen
reto.gubelmann@unisg.ch

Aikaterini-Lida Kalouli

CIS - LMU Munich
katerina.kalouli@hotmail.com

Christina Niklaus

University of St. Gallen
christina.niklaus@unisg.ch

Siegfried Handschuh

University of St. Gallen
siegfried.handschuh@unisg.ch

Abstract

In this paper, we focus on the ability of large language models (LLMs) to accommodate different pragmatic sentence types, such as questions, commands, as well as sentence fragments for natural language inference (NLI). On the commonly used notion of logical inference, nothing can be inferred from a question, a command, or an incomprehensible sentence fragment. We find MNLI, arguably the most important NLI dataset, and hence models fine-tuned on this dataset, insensitive to this fact. Using a symbolic semantic parser, we develop and make publicly available, fine-tuning datasets designed specifically to address this issue, with promising results. We also make a first exploration of ChatGPT’s concept of entailment.

1 Introduction: “I didn’t say that!”

Committing oneself to the truth of a certain claim always implies or suggests one’s commitment to the truth of a number of other claims, and it precludes one’s commitment to the truth of a second set of claims. This is the essence of the logical notion of entailment (here used synonymously with inference) and contradiction. For instance, somebody who claims “Loral did harm national security” is also committed to the less specific claim “Something or somebody did harm national security”.

The concept of inference is itself quite complex. As Gubelmann et al. (2022) detail, valid inferences can be divided into deductively valid inferences, where it is not logically (see Plantinga 1974) possible that the premise is true while the conclusion is false, and inductively valid inferences (also called abductions), where it is possible that the premise is true while the conclusion is false, but where the truth of the premise is in general a good reason for the truth of the conclusion.

For two utterances to be able to (deductively or inductively) entail or contradict each other, they have to be of the correct pragmatic category. While

assertions or claims are able to stand in these logical relationships, the same does not hold for other pragmatic kinds of utterances, such as questions or commands, as they do not involve making a claim that could be true or false and hence commit their author to the truth of certain claims that could then entail or contradict other claims. For instance, uttering (P) in example (1) does not commit the speaker to Loral’s harming of national security – she’s simply asking a question, not making a claim. Hence if, after uttering (P) from (1), somebody replies with “So you claim that somebody did harm national security”, the appropriate response would be “I didn’t say that!”.

Neither does she commit herself to everybody’s, as a matter of fact, having a happy Hanukkah when uttering (P) in (2), that is, wishing everyone a happy Hanukkah: Rather, she is expressing the wish that everyone is going to have a happy Hanukka. Committing oneself to some state of affairs, i.e., to express a claim that can be true or false, is not the kind of thing one does when uttering a question or a command (which is not to dispute that questions and commands come with specific presuppositions, including factual presuppositions, that need to be fulfilled for the speech act in question to succeed).

(1) (P) Did Loral harm national security? (H) National security was not in danger. (contradiction)

(2) (P) Happy Hanukkah, everybody! (H) Everyone, have a happy Hanukkah! (entailment)

Properly distinguishing between pragmatic kinds of utterances that can and cannot stand in logical relations is important in several areas of application. First, we can consider the legal context, where fact and claim verification is of critical importance. In the setting of the automatic extraction of claims from testimonies, a system should be able to dis-

tinguish between claims and statements with other pragmatic functions such as questions, which do not commit their speakers to the truth of any claims (see, e.g., the overview in Ashley (2018)). Additionally, an essential application area is education: Using large language models (LLMs) to give formative feedback on students’ arguments requires that the LLMs be able to distinguish between claims made in the text, which can be used to infer other claims, and questions and commands, which cannot (see Rapanta et al. (2013) for an illustration of the importance for this logical concept of entailment in education science).

Most recently, the introduction of general-domain, openly-available conversational systems such as ChatGPT (OpenAI) shows the need for such a distinction even more clearly: a chatbot, which collects its information from web resources but does not perform any reasoning steps itself, can falsely spread non-claims as claims if it cannot differentiate between the two.

Current NLP research conceives natural language inference (NLI) as a three-way classification task between two sentences (or sentence-fragments), called premise (P) and hypothesis (H). LLMs are trained to predict *contradiction* (P and H cannot both be correct), *entailment* (If P is correct, then H must be correct as well), or *neutral* (neither of the two). While much of the very early research focused on deductively valid inferences, more recent research has also taken into account inductive inference, which are called applied entailments (Dagan et al., 2005) or informal reasoning (MacCartney, 2009).

The Multi-Genre Natural Language Inference Dataset (MNLI) Williams et al. (2018) has arguably become the most widely used dataset for fine-tuning LLMs for NLI. This means that many (perhaps the majority of) LLMs that are fine-tuned for the task of NLI are fine-tuned using MNLI and thereby pick up MNLI’s concept of inference. The instructions given to the crowdworkers who worked to create the dataset as well as explicit comments by the authors support the conclusion that MNLI’s target notion of entailment dovetails with the one detailed here, applicable to claims but not to questions and commands.

However, MNLI contains prompts that are questions, such as in (1), or commands, such as in (2), as well as fragments such as (3), which are entirely incomprehensible if they are presented, as in

MNLI, without any context. In fact, all examples are from MNLI’s training split with their respective gold-labels in brackets. Hence, there seems to be a conceptual gap between the notion of entailment as explicitly embraced by the authors of MNLI and the pragmatic kinds of some of the prompts used to create the dataset.

(3) (P) The kids. (H) The adults. (contradiction)

In this paper, we study the extent of the phenomenon, the consequences that this set-up of MNLI has for LLMs that are fine-tuned on MNLI, and we explore ways to acquaint the LLMs with these core pragmatic categories. Our paper makes three contributions. First, after detailing the notion of inference as well as the conceptual gap in MNLI on a theoretical level (section 3), we empirically assess the extent of the phenomenon of non-assertive premises in MNLI (section 4). Second, relying on the existing semantic parser GKR, we show a promising path towards acquainting LLMs with these pragmatic categories (section 5). Third, we publish both an expert-curated gold-standard evaluation dataset as well as 7 different fine-tuning datasets to further research in this field.¹ Additionally, we also take the very first steps toward exploring ChatGPT’s concept of entailment.

2 Related Research

2.1 Inference in Logic and Semantics

Both deductive and inductive inferences require claims with determinate truth-values for their functioning. This means that it is necessary for any relationship of inference to be possible that both relata are constituted by a claim with determinate truth conditions: it needs to be clear in which situations premise and hypothesis are true. Otherwise, it would be impossible to assess whether the truth of the premise guarantees/makes reasonable the truth of the hypothesis, which is the essence of both deductive and inductive inferences. We propose that only sentences fulfilling the following conditions C1 and C2 can express such determinate claims.

C1 Only sentences whose pragmatic force is assertive can express determinate claims.

C2 Only assertions which are sensible (that is, where it is clear what has to be the case for

¹To access the datasets, please consult: <https://github.com/retoj/whentruthmatters>.

the claim to be true) can express determinate claims.

C1 is violated by questions and commands, such as the premises in examples (1) and (2). The question “Did Loral harm national security?” lacks determinate truth conditions because questions cannot be true or false, but rather sensible or nonsensical. Similarly, uttering a command like “Happy Hanukkah!” does not aim to make a determinate claim about the state of affairs but rather aims to bring about a certain state of affairs.

We owe the insight that one can do different things with different types of sentences (the theory of so-called “speech acts”) to Austin (1962, 1975) and Searle (1969, 1985), continuing a basically Wittgensteinian outlook Wittgenstein (2006/1953, §43).² For a more recent survey of this approach, see Levinson (2017). Speech acts, like any actions, can succeed or fail to reach the goal that the agent intends with it. If some presuppositions for an act are not met, then it cannot possibly succeed.

With regard to fulfilling C2, the bare minimum needed for a sentence to express a claim with determinate truth conditions is some entity that is identified with sufficient precision (call it “subject”) as well as something, again sufficiently precise, that is predicated of that entity (call it “predicate”). This conception of a minimal claim as consisting of some specific entity of which something is said is a standard in logical textbooks, see, e.g., Smullyan (1968, 43) or Garson (2006, 29), but also in everyday human communications. For an overview, see Shapiro and Kouri Kissel (2021, sec. 2.2). For an influential contemporary statement of this minimal notion of a determinate claim, see Burge (2010, 537-547). For example, C2 is clearly violated by the premise of (3). Without any further context, and *MNLI does not provide any such context*, it is not clear whether the fragment “the kids” is intended as subject or predicate, but it is clear that one of the two is missing.

To see that nothing can be inferred from anything that violates C1 and C2, it is crucial to be aware of the distinction between *inference* and *presupposition*. For instance, one might be tempted to say that from the question (1), it can be inferred that Loral potentially endangers national security, which would contradict the hypothesis of this exam-

²Mastery of this multitude of language games in a flexible and adaptive manner is a key challenge for AI systems to come to really understand language, see Gubelmann (2023).

ple. This, however, would be to confuse inference with presupposition (the subtleties of the notion of presupposition, going back to Russell 1905 are still lively discussed in linguistics, see Dryer 1996 for a more recent influential contribution).

To claim that national security was never in danger would not, as the gold label for example (1) suggests, contradict the question: questions can be answered, rejected, ridiculed, etc., but not contradicted in the relevant logical sense. Rather, the claim would (at least on some readings of the question) show that the question fails to make proper sense, as one of its presuppositions, that national security was ever endangered, is not met.

A phenomenon similar to presuppositions has been described by Grice (1975) as conventional implicature.³ Roughly, conventional implicatures, unlike presuppositions, do not affect the sensibility of the utterance in question (this follows (Potts, 2015, 31), who argues that an implicature, unlike a presupposition, is independent from the primary content of the utterance.⁴ Consider example (4).

- (4) (a) Bern, the capital of Switzerland, is the largest city of the country. (b) This is not true.

On a first level, the claim expressed by sentence (a) in example (4) is simply wrong: Zurich, not Bern, is the largest city of Switzerland. On the second level, however, it is also not the case that Bern is the capital of Switzerland: The founders of Switzerland deliberately avoided designating an official capital city due to existing rivalries between the candidates for such a role. This second level is beyond the reach of the challenge (b), it only reaches the actual claim being made about the relative population of Bern.

In fact, this availability for direct challenge is what helps to clearly identify the determinate, claimed content in an utterance – and it also helps to establish whether there is any such claimed content in the first place. With commands such as the one in example (2), you cannot respond with “This is not true”, nor can you do so in response to a question such as in (1), or in response to an incomprehensible fragment such as in (3). In contrast, to

³See Davis (2019) for an overview. For more recent discussions of the concept, see Potts (2004); Lepore and Stone (2010). For an original perspective on the concept from formal semantics, see Peters (1979).

⁴Karttunen and Peters 1979, in contrast, use the two concepts almost equivalently.

access and reject the presuppositions behind some of these utterances, one has to do more linguistic work. For instance, you could reject a possible presupposition of the question in example (1) with: “Your question is beside the point because Loral has never produced anything else than toilet paper; therefore, the very idea that it could have been even a potential danger for national security is misguided.”

The second important distinction that we want to point out is that between inference and meaning-preserving paraphrase. The two relationships are orthogonal: If one claim with determinate truth-conditions is the meaning-preserving paraphrase of another such claim, then they can be mutually inferred from each other. If, in contrast, what is being paraphrased is a question, a command, or an incomprehensible fragment, then no relationship of inference exists between the original and the paraphrase – regardless of how synonymous they are.

We would, finally, like to note that we recognize the usefulness of a broader, non-truth-functional notion of entailment for uses beyond NLI. For instance, [Groenendijk and Stokhof \(1984, p.47f.,p481f.\)](#) define entailment between questions by resorting to a very general, non-truth-functional notion of entailment as a kind of semantic inclusion. Such a notion is very useful for question-answering or information retrieval tasks, but it is not how the NLI task was originally defined [Dagan et al. \(2005\)](#); [MacCartney \(2009\)](#), nor what the MNLI instructions to the crowdworkers specify, as we shall see.

2.2 Inference in NLP

LLMs based on the transformer architecture ([Vaswani et al., 2017](#)) have become the *de facto* standard in a variety of NLP tasks, including NLI. Highly successful architectures, starting with BERT ([Devlin et al., 2019](#)) and followed by others such as RoBERTa ([Liu et al., 2019](#)), XLNet ([Yang et al., 2019](#)), DeBERTa ([He et al., 2020](#)) as well as smaller versions such as DistilBERT ([Sanh et al., 2019](#)) and Albert ([Lan et al., 2019](#)), but also sequence-to-sequence architectures, e.g., T5 ([Rafel et al., 2019](#)) and BART ([Lewis et al., 2020](#)), have shown state-of-the-art performance on NLI.

Thanks to their sheer size, SNLI ([Bowman et al., 2015](#)), 570k premise-hypothesis pairs from image captions, and MNLI ([Williams et al., 2018](#)), 433k premise-hypothesis-pairs from 10 genres, written

and spoken, dominate the field, as their size is suitable for fine-tuning large LLMs. There is a number of studies that critically assess SNLI and MNLI for their bias. [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 (this bias is most likely due to the fact that simply negating the premise provides an efficient way for crowdworkers 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 LLMs 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. Furthermore, [Pavlick and Kwiatkowski \(2019\)](#), [Zhang and de Marneffe \(2021\)](#), and [Jiang and de Marneffe \(2022\)](#) all address the topic of disagreement among annotators. [Jiang and de Marneffe \(2022\)](#) focus on MNLI and suggest using a fourth category, namely “complicated”, along with the known ones of entailment, contradiction, and neutral. Similarly, [Kalouli et al. \(2019, 2023\)](#) discuss the annotation artifacts and quality of such datasets, especially concerning the distinction between neutral and contradiction pairs, and propose a refinement of the task.

We contribute to this ongoing research by focusing on the pragmatic categories of sentences (questions, commands, claims) which determine whether they can stand in the logical relationships introduced above (section 2.1). We use GKR ([Kalouli and Crouch, 2018](#); [Kalouli, 2021](#)) to automatically categorize premises from MNLI that violate C1 or C2. GKR (Graphical Knowledge Representation) is the semantic representation generated by the corresponding parser. In GKR the sentence information is split into six subgraphs: a) the dependency graph holding the syntactic dependencies, b) the lexical graph holding lexical information such as synonyms and antonyms of the words of the

sentence, c) the properties graph holding morpho-syntactic information such as the numerus of nouns and quantifiers, d) the concept graph holding the basic predicate-argument-structure of the sentence, the “who-is-doing-what-to-whom” information, e) the context graph making existential commitments over the concepts of the concept graph, e.g., for the sentence “the dog is not eating the bone” it says not only that there is the concept of *eating* involved in the sentence, but it also commits to its non-existence, its non-*instantiation* (due to the negation) and f) the coreference graphs capturing coreference links between entities.

In addition to its performance, what makes this parser particularly suitable for our goal is that it also identifies the type of sentence that is being parsed (assertion, question, or command). By default, the parser also categorizes subject-less sentences as imperatives (which helps to identify incomprehensible sentence fragments).

3 Analyzing MNLI’s Concept of Inference

After describing how current research in linguistics and logic conceives inference and separates it from presupposition, implicature, and paraphrase, and after situating MNLI in the current way how NLP approaches the task of NLI, we now detail the tension that we see in MNLI’s concept of inference.

Given how Williams et al. (2018, 1114) specify the tasks for the crowdworkers creating MNLI, the goal seems to be premise-hypothesis pairs that are deductively valid. We give the part of the instruction that is relevant for entailment in bold (for the full instructions, see the Appendix, section A):

[...] 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.**

According to this passage, the hypothesis to be written should be such that it is definitively correct about the situation or event described in the premise. World knowledge is allowed to be used, presumably to make room for implicit but uncontroversial premises. From a logical point of view, this means that whenever the state of affairs described in the premise obtains, the one described in the hypothesis must obtain as well. Hence, MNLI seems indeed to follow earlier NLI research and aim at inference in the deductive or inductive sense detailed

above (sections 1 and 2.1). This also agrees with the stated goal of Williams et al. (2018), according to which they are aiming at pairs where the hypothesis is “necessarily true or appropriate whenever the premise is true”.

However, a manual inspection of the collected examples shows a different picture: the dataset contains entailment and contradiction pairs with premises that are non-assertive because they express commands or questions, or because they are fragmentary beyond comprehension (see examples above (1), (2), and (3)). Note that these examples are not resulting from cherry-picking: The creators of MNLI deliberately selected bits of text at random from 10 different genres, emphasizing that they only applied minimal pre-processing (e.g., removing sentences with less than eight characters, mathematical formulae, bibliographical references, see (Williams et al., 2018, 114f.)). No grammaticality checks or parsing of sentence types are done. Hence, including incomprehensible fragments as well as questions and commands results from an explicit design decision by the authors.

Unfortunately, this design decision seems to be in tension with the instructions to the crowdworkers as well as with the stated goal to find premises that are true or appropriate whenever the premise is true. The examples (1), (2), and (3) do not contain premises that can be true or false, making it exceedingly difficult for the crowdworkers to follow the instructions and write a sentence that is definitively correct about the situation or event in the line: Commands and Questions do not aim to describe situations, incomprehensible fragments cannot describe such situations.

The crowdworkers did their best. Sometimes, as in (1), they developed a hypothesis that contradicts one of the presuppositions of a question (see, for example, (1)), developed a largely synonymous command for a premise containing a command (see example (2)), or just wrote down a concept that differs from the concept in the premise (see (3)). None of this, of course, amounts to developing entailment or contradiction pairs.

In our pre-study, we try to quantitatively assess the extent of this problem and develop a solution for it.

4 Pre-Study: Non-Assertive Premises in MNL

With this pre-study, we pursue two goals. First, we would like to obtain a more reliable estimate for the amount of non-assertive premises (that is, premises that do not express a determinate claim because they violate C1 or C2 from section 2.1 above) in MNL. Second, our main experiment relies on GKR correctly categorizing premises from MNL that violate C1 or C2 (for details, see section 2.1) or that don't violate them and thus represent assertive sentences. Thus, before starting with the main experiment, we conduct a pre-study to evaluate our choice of using GKR.

We randomly select 1000 premise-hypothesis-pairs from MNL and submit the premises of each of the samples to the parser. The output of GKR (more specifically its context graph, where the type of sentence is specified) is then compared to our gold-standard annotations. The precision (P) and recall (R) results of this pre-study are shown in Table 1. The table also compares the results to a simplistic baseline approach, where we only count as non-assertive all premises that end with a question mark (?) or an exclamation mark (!).

	Assertive		Non-Assertive		Total	
	P	R	P	R	P	R
GKR	93.6	97.7	83.6	63.3	88.6	80.5
?, !	88.4	96.5	54.6	22.8	71.5	59.6

Table 1: The P(precision) and R(recall) results from the evaluation of 1000 MNL premises.

Table 1 shows that GKR's precision is at 83.6% and its recall at 63.3%, when it comes to identifying non-assertive premises that cannot stand in a logical relationship. Identifying assertive premises is achieved with an even higher precision and recall (93.6% and 97.7%, respectively). These results heavily outperform the simplistic baseline approach, in which both the precision and the recall for non-assertive premises do not overcome chance. This is not surprising though. First, in MNL, questions and commands are not always marked with question and exclamation marks, respectively, so this is no reliable method. Second, many questions do have a question mark, but are direct speech embedded in indirect speech, e.g., *How much? asked the northerner*. In these cases, the premises are indeed assertions (with an embedded non-assertive content). Third, there are premises with exclamation marks, which are no commands, e.g., *You were*

just wonderful!. Finally, this simplistic approach cannot capture any cases of fragmentary premises. These results confirm the quality of the GKR parser and the need for such a tool.

With our pre-study we find that 153 of the 1000 samples are non-assertive (based on their gold label). This suggests that approximately 15% of all pairs in MNL are indeed not assertions, meaning that they cannot entail or contradict any other assertions⁵. This however also means that any of these pairs having an entailment or a contradiction label (assuming a balanced dataset, this would mean around 10%) is indeed mislabeled as there can only be the neutral relation for non-assertions. Note that even if we do not consider the gold labels but only the true positives of GKR (since GKR's output is what will be considered in the main experiment), the percentage of non-assertions in MNL only drops to 10% (97 out of 1000 samples are true positives). This would again mean that around 2/3 of these 10%, that is, some 6.6%, of MNL is incorrectly annotated.

5 Main Experiment: Probing LLMs for Pragmatic Understanding

For our main experiment, based on the tension found in MNL's concept of inference (see above, section 3), we hypothesize that *models fine-tuned on MNL lack any sensitivity to the fact that non-assertive premises cannot entail or be contradicted by other premises* (research hypothesis 1), and that *this deficit can be amended using properly composed fine-tuning datasets* (research hypothesis 2). Finally, we hypothesize that *this does not significantly harm performance on the original MNL evaluation dataset* (research hypothesis 3). To empirically test these hypotheses, we compile a number of fine-tuning datasets and evaluate LLMs fine-tuned on them both on a specific, hand-corrected dataset that only contains neutral premises as well as on the original MNL-matched evaluation dataset.

5.1 Models

We use three transformer-based models that are already fine-tuned on MNL, delivering very good performance on this dataset, and that differ substantially in their architecture. We deliberately choose

⁵Since this was tested on a random sample, we can expect that this percentage will be similar for any other random sample.

models based on different architectures, sizes and fine-tuning methods.

The reason why we fine-tune models that are already fine-tuned on MNLI is that we assume that our non-assertive dataset is not large enough on its own for learning inference. Thus, we use models that are first fine-tuned on the large MNLI dataset and have thereby acquired a basic understanding of the task. Then, we fine-tune them further on the smaller non-assertive dataset to sharpen their conceptions of entailment and contradiction so that they do not predict entailment or contradiction, but rather neutral, when the premise is non-assertive. To test our assumption, we also include a version of RoBERTa-large that is not already fine-tuned to MNLI in our fine-tuning experiments. If it performs better than the others who are already fine-tuned to MNLI, our assumption is falsified, otherwise, it is verified.

The models are DeBERTa-base (He et al., 2020), XLNET-base (Yang et al., 2019), both 110M parameters, and RoBERTa-large (Liu et al., 2019), 355M parameters. Our DeBERTa-model is fine-tuned to MNLI using the method proposed by Reimers and Gurevych (2019), the XLNET-model by the adversarial method proposed in Morris et al. (2020), and for RoBERTa, we use the original fine-tuned version by Liu et al. (2019). We obtain all of our models from Huggingface (Wolf et al., 2019).

5.2 Datasets

We run GKR over randomly chosen premises of the train split of MNLI. As compute time per sample is rather high (about 30 sec per sample in our setting), we stopped the process after receiving 1875 premises that GKR classified as either interrogative or imperative (a label also given to sentence fragments lacking a subject), and hence non-assertive. The same run also yielded 8546 premises that GKR classified as assertions. Based on this, we develop a manually corrected evaluation dataset as well as a number of systematically varied fine-tuning datasets.

Evaluation-Datasets We use a subpart of the 1875 non-assertive premises, namely 636 premises, to compile an evaluation dataset: We manually verify that these premises are indeed not expressing a determinate claim (either because they are questions, commands, or incomplete beyond understanding), resulting in 406 premises. We then select the three premise-hypothesis-pairs correspond-

ing to each premise in the MNLI dataset (for each premise there was an entailment, a contradiction and a neutral pair created). This results in 1218 pairs whose correct relationship should be neutral due to their premise, but which were written up by crowdworkers to be evenly split among the labels of entailment, contradiction, and neutral.⁶ We call this evaluation dataset “GKR-n” for “GKR-neutral”. The second dataset that we use to evaluate our fine-tuned models is MNLI-Matched (“MNLI-M”), the matched evaluation dataset provided by Williams et al. (2018).

Fine-Tuning-Datasets We compile 8 different fine-tuning datasets, each consisting of a train split containing 6000 samples and a validation split containing 600 samples, evenly distributed across the three labels. In addition to a dataset that solely consists of unfiltered MNLI-train samples (mnli_u) as well as a dataset consisting solely of samples whose premises GKR classified as assertive (GKR_a), we compile six datasets combining these two sources (see Table 2). These six datasets are combinations from two different datasets for entailment and contradiction labels (entailment-contradiction 1 & 2, in short, ec1 & ec2) with three different datasets for neutral labels (neutral 1,2, 3: n1,n2,n3).

Name	Ent. & Contr.	Neutral
mnli_u	(no filtering)	(no filtering)
ec1_n1	1/3 GKR-ass.	1/3 GKR non-ass.
ec1_n2	1/3 GKR-ass.	2/3 GKR non-ass.
ec1_n3	1/3 GKR-ass.	3/3 GKR non-ass.
ec2_n1	2/3 GKR-ass.	1/3 GKR non-ass.
ec2_n2	2/3 GKR-ass.	2/3 GKR non-ass.
ec2_n3	2/3 GKR-ass.	3/3 GKR non-ass.
GKR_a	3/3 GKR-ass.	3/3 GKR-ass.

Table 2: Fine-tuning datasets used in the main experiment. They result from combining unfiltered MNLI samples with samples filtered by GKR; if nothing is specified, the quantity in question is provided by unfiltered pairs from the MNLI training split. All datasets contain 2.2k samples originally labeled entailment, 2.2k contradiction, and 2.2k neutral. For example, in ec1_n1, 1/3 of its 4.4k entailment and contradiction samples have been filtered by GKR to make sure their premises are assertive (“1/3 GKR-ass.”), and 1/3 of its 2.2k neutral samples have been filtered by GKR to have non-assertive premises (“1/3 GKR non-ass”).

For example, the dataset ec2_n2 in Table 2 is composed of 6600 samples in total, of which 2200 are entailment, 2200 contradiction, and 2200 neu-

⁶Note that some premises in MNLI are used in more than three pairs, leading to minimal distortions in the scaling up.

tral. Of the entailment and contradiction samples, making up 4400 samples, 2/3 are such that GKR has identified their premises as assertions (2nd column in T 2). Furthermore, 2/3 of the neutral samples are such that GKR has identified their premise as non-assertive (3rd column in Table 2).

The basic idea behind fine-tuning on these 8 different datasets is to see what factors influence performance on the evaluation dataset: ec1 differs from ec2 in containing only 1/3 as opposed to 2/3 of pairs whose premises have been confirmed by GKR to be assertive. From n1 over n2 to n3, the portion of neutral pairs with non-assertive premises increases from 1/3 over 2/3 to 3/3. Testing all combinations of these datasets allows us to determine whether assertive premises in entailment and contradictions samples are more important than non-assertive premises in neutral samples when it comes to performance on the evaluation dataset, where all of the labels should be neutral due to the non-assertive premises of all the samples used there. Including the mnli_u dataset as fine-tuning dataset allows us to test whether our fine-tuning method leads to any distortions: unless our fine-tuning method is flawed, models fine-tuned on mnli_u should perform on MNLI-M approximately as they did before fine-tuning. After all, here we just perform further fine-tuning with the same data that has been used for the original fine-tuning run.⁷

5.3 Fine-Tuning Experiment

We fine-tune the four different transformer-based LLMs on a single GPU of a DGX-2 cluster. We fine-tune each model on each dataset for two epochs, using the trainer API provided by Huggingface. The results shown are the average over three fine-tuning runs per model and dataset. We use a batch size of 8 throughout, and we begin with an initial learning rate of $2e-5$.

5.4 Results & Discussion

The results of our experiment are shown in Figure 1. We give the individual models' results on the two evaluation datasets sorted by the fine-tuning dataset that was used.

With regard to our decision to use models previously fine-tuned to MNLI (see above, 4.1), the

⁷Note that it would not be useful for our purposes to convert non-assertive premises to assertions. We need to have them marked as neutrals/non-assertive as systems need to be able to differentiate between these pragmatic functions (see discussion in section 1).

results confirm our assumptions: roberta-large performs overall 28% worse than the version of the model that was previously fine-tuned to MNLI (abbreviated by roberta-lmnli in figure 1). As a consequence, we do not consider it in our presentation and discussion of results anymore.

Figure 1 shows that the most important factor for performance on the GKR-n evaluation dataset is the portion of neutral samples that are neutral because their premise is non-assertive. The accuracy of all models is 32% on average, and hence almost exactly random, if no such samples have been specifically selected and added to the fine-tuning dataset (as is the case in mnli_u as well as GKR_a, see the first and the penultimate columns, respectively). This accuracy increases steadily if the portion of neutral samples of said kind is increased from n1, n2, to n3, where it reaches 89% for roberta-large-mnli. Performance on MNLI-matched decreases from n1 to n2 and n3, but in much smaller steps: from 86% to 82% to 77%.

These results allow for three main insights. First, without fine-tuning on our datasets, the LLMs do indeed fail to show any sensitivity for the fact that questions, orders, or incomprehensible fragments cannot entail or contradict anything. This follows from the random accuracy that the models reach after being fine-tuned on mnli_u, it confirms our first hypothesis, and it lends further support to the tension found in MNLI's concept of inference. Considering the fact that MNLI is the *de facto* standard fine-tuning dataset, this means that the standard method used currently to fine-tune LLMs to NLI tasks very likely results in models that falsely classify pairs such as (1), (2), or (3) as contradicting or entailing each other.

The second insight, confirming our second hypothesis, is that our fine-tuning approach shows much promise in getting the models to understand that nothing follows logically from non-assertive premises. In particular, this applies to the models fine-tuned to ec1_n3, i.e., to a dataset that contains 1/3 of pairs with a premise that is assertive according to GKR in entailment and contradiction and 3/3 of pairs with non-assertive premises in neutral.

The third insight is that acquiring this sensitivity does not take a heavy toll on the accuracy of MNLI-matched, with an effective accuracy difference of 3.3%. While the performance difference seems greater at first sight (10% from 87% to 77%), it must be noted that 10% of the MNLI-premises

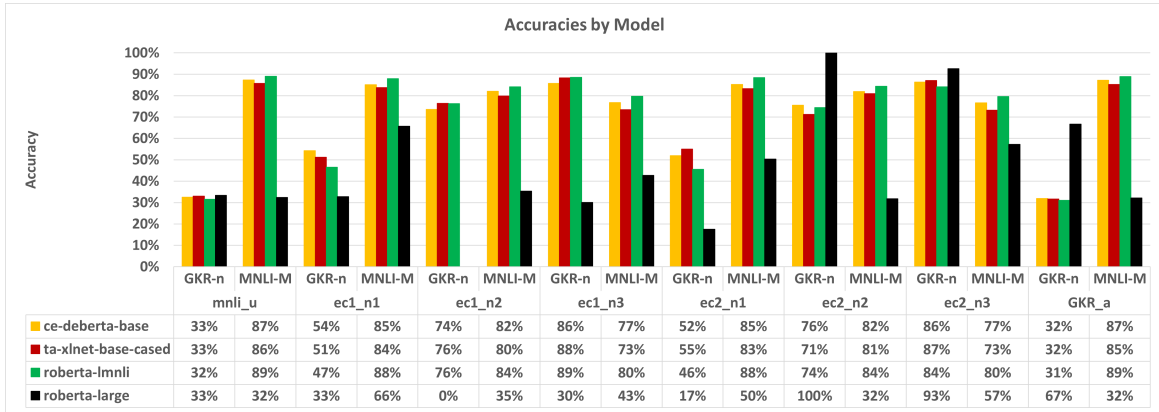


Figure 1: Accuracy by model and fine-tuning dataset (each fine-tuning was run three times and the average of the three runs is reported). For instance, the column on the very left gives the accuracies on our GKR-n evaluation dataset after fine-tuning on mnli_u for each of the three models. Performance of roberta-large, which was not already fine-tuned on MNLI, is on average 28% lower than roberta-large-mnli.

are anyway non-assertive (see Section 4) and thus those of them (approximately $2/3$, 6.6%) that are labeled as entailments or contradictions are mislabeled, yielding an actual loss in accuracy of 3.3%. We emphasize that performing well at GKR-n requires that the LLMs predict neutral for any of the pairs in that dataset (because their premises are all non-assertive), while performing well at MNLI-M of course requires to predict all of the three labels with equal frequency. It is reassuring that the same LLMs manage to perform well at both evaluation datasets, confirming our research hypothesis 3.

6 Outlook: Exploration of ChatGPT

Following recent advances in the area we are curious to see whether ChatGPT, a general-purpose chatbot trained by OpenAI⁸ a) has a better notion of entailment than other LLMs, and b) can correctly identify non-assertive statements and treat them accordingly. To address these questions, we manually explore ChatGPT: We manually prompt the Chatbot with 96 premise-hypothesis-pairs with non-assertive premises according to GKR that were originally intended as entailment pairs by the creators of MNLI. We join premise and hypothesis to obtain a question (see also Appendix C).

We find that, although ChatGPT gets the general definition of logical entailment perfectly right (“entailment is a relationship between two propositions, in which the truth of the premise guarantees the truth of the conclusion”, something that it tends to

assert quite often, see the Appendix, section B), it often fails to apply it to the given examples: it states that the hypothesis is logically implied (entailed) by the non-assertive premise in 54% of the cases. Interestingly, from the remaining 46% of the cases, where ChatGPT indeed answers negatively, i.e., that there is no entailment, we observe that in 27% of the samples the non-assertiveness of the premise stems from its being too fragmentary to express a specific proposition. This indicates that ChatGPT is indeed able to tell when a premise is too incomplete to express a determinate claim. Hence, this very small sample might suggest that ChatGPT has a more accurate notion of entailment than the best models tested in the main experiment, but this does not fundamentally alter the scene: ChatGPT can perform better with incomplete sentences, realizing that nothing can be logically inferred from them, but relatively poorly with questions or commands.

7 Conclusion

In sum, we take the results of our experiments to be very encouraging. While LLMs that have been fine-tuned only on MNLI show no sensibility for the fact that nothing follows logically from questions, commands, or incomprehensible fragments, fine-tuning on our datasets can address this potentially consequential shortcoming without losing too much accuracy on MNLI-matched. In the future, we plan to experiment with more sophisticated fine-tuning methods and datasets to maximize performance on non-assertive prompts. Furthermore, we hope to probe ChatGPT using larger datasets, as the preliminary results obtained here are very interesting.

⁸See [this blogpost](#), last consulted on December 14, 2023. All interaction occurred on December 14, 2022, via OpenAI’s research interface: <https://chat.openai.com/chat>.

Limitations

We see two main limitations of this work. First, it is clear that the quality of the proposed fine-tuning method (with the non-assertive datasets) heavily depends on the quality of the GKR parser, used for the detection of non-assertive statements. Since the parser has an F-score of 84.3%, as evaluated within this work, it is to be expected that some of the examples used for fine-tuning will be faulty (mistakenly classified as non-assertive and thus mistakenly classified as neutrals). This means that the models are not given entirely clean input to learn from and thus, their performance could further improve if the quality of the GKR parser (or another similar parser) could improve. Note however that this only concerns the fine-tuning datasets and not the evaluation dataset, as this was manually checked for correctness (and thus, there is no impact on the results we report). The second limitation concerns our experimentation with ChatGPT: we were only able to experiment with a small amount of samples (due to the extremely expensive manual effort required) and thus, our findings will need to be reproduced through a larger-scale study. However, our findings agree with [Bang et al. \(2023\)](#), who find that ChatGPT performs reasonably well on deductive and abductive inference tasks.

Ethics Statement

For fine-tuning our models, we have used parts of the existing MNLi dataset. Thus, we see no additional risk of adding personally protected information, offensive material, or biases that could discriminate against marginalized groups.

References

- Kevin D Ashley. 2018. Automatically extracting meaning from legal texts: opportunities and challenges. *Ga. St. UL Rev.*, 35:1117.
- John Austin. 1962. *How to do things with words*. Clarendon Press.
- John Langshaw Austin. 1975. *How to do things with words*. Oxford university press.
- Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multi-task, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.
- 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.
- 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).
- Tyler Burge. 2010. *Origins of Objectivity*. Oxford: Oxford University Press.
- 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.
- Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Proceedings of the Machine Learning Challenges Workshop*, pages 177–190, Southampton, UK. Springer.
- Wayne Davis. 2019. Implicature. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Fall 2019 edition. Metaphysics Research Lab, Stanford University.
- 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.
- Matthew S Dryer. 1996. Focus, pragmatic presupposition, and activated propositions. *Journal of pragmatics*, 26(4):475–523.
- James W Garson. 2006. *Modal logic for philosophers*. Cambridge: Cambridge University Press.
- Herbert P Grice. 1975. Logic and conversation. In *Speech Acts*, pages 41–58. Brill.
- Jeroen Antonius Gerardus Groenendijk and Martin Johan Bastiaan Stokhof. 1984. *Studies on the Semantics of Questions and the Pragmatics of Answers*. Ph.D. thesis, Univ. Amsterdam.
- Reto Gubelmann. 2023. A Loosely Wittgensteinian Conception of the Linguistic Understanding of Artificial Neural Networks. *Forthcoming in Grazer Philosophische Studien*.
- Reto Gubelmann, Christina Niklaus, and Siegfried Handschuh. 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)*, pages 38–50.
- 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.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.
- Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating reasons for disagreement in natural language inference. *arXiv preprint arXiv:2209.03392*.
- Aikaterini-Lida Kalouli. 2021. *Hy-NLI : a Hybrid system for state-of-the-art Natural Language Inference*. Ph.D. thesis, Universität Konstanz, Konstanz.
- Aikaterini-Lida Kalouli, Annebeth Buis, Livy Real, Martha Palmer, and Valeria de Paiva. 2019. [Explaining simple natural language inference](#). In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 132–143, Florence, Italy. Association for Computational Linguistics.
- Aikaterini-Lida Kalouli and Richard Crouch. 2018. Gkr: the graphical knowledge representation for semantic parsing. In *Workshop on Computational Semantics beyond Events and Roles (SemBEaR 2018)*, pages 27–37.
- Aikaterini-Lida Kalouli, Hai Hu, Alexander F. Webb, Lawrence S. Moss, and Valeria de Paiva. 2023. [Curving the SICK and Other NLI Maladies](#). *Computational Linguistics*, pages 1–45.
- Lauri Karttunen and Stanley Peters. 1979. Conventional implicature. In *Presupposition*, pages 1–56. Brill.
- 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*.
- Ernie Lepore and Matthew Stone. 2010. Against metaphorical meaning. *Topoi. Orient – Occident*, 29:165–180.
- Stephen C Levinson. 2017. Speech acts. In *Oxford handbook of pragmatics*, pages 199–216. Oxford: Oxford University Press.
- 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.
- 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. 2009. *Natural Language Inference*. Ph.D. thesis, Stanford University, Stanford, CA, USA. AAI3364139.
- 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.
- OpenAI. ChatGPT.
- Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.
- Stanley Peters. 1979. A truth-conditional formulation of Karttunen’s account of presupposition. *Synthese. An International Journal for Epistemology, Methodology and Philosophy of Science*, 40:301–316.
- Alvin Plantinga. 1974. *The Nature of Necessity*. Oxford: 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.
- Christopher Potts. 2004. *The Logic of Conventional Implicatures*, volume 7. OUP Oxford.
- Christopher Potts. 2015. Presupposition and implicature. *The handbook of contemporary semantic theory*, pages 168–202.
- 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*.
- Chrysi Rapanta, Merce Garcia-Mila, and Sandra Gilabert. 2013. What is meant by argumentative competence? an integrative review of methods of analysis and assessment in education. *Review of Educational Research*, 83(4):483–520.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.
- 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*.

John Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge: Cambridge University Press.

John Searle. 1985. *Expression and meaning: Studies in the theory of speech acts*. Cambridge: Cambridge University Press.

Stewart Shapiro and Teresa Kouri Kissel. 2021. Classical Logic. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Spring 2021 edition. Metaphysics Research Lab, Stanford University.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undekasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17*, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

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.

Ludwig Wittgenstein. 2006/1953. Philosophische untersuchungen. In *Werkausgabe Band 1*. Frankfurt am Main: Suhrkamp.

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*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*, 32.

Xinliang Frederick Zhang and Marie-Catherine de Marneffe. 2021. Identifying inherent disagreement in natural language inference. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

A Full Instructions Given to Crowdworkers

Williams et al. (2018, 1114) specify 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 ChatGPT on logical relationships involving commands and hopelessly incomplete premises

All interaction occurred on December 14, 2022, via OpenAI's research interface: <https://chat.openai.com/chat>.

Human: Is there a relation of logical entailment between "What is the capital of france?" and "Paris"?

ChatGPT: Yes, there is a relation of logical entailment between the question "what is the capital of France?" and the answer "Paris". In logic, entailment is a relationship between two propositions (statements or sentences) in which the truth of one proposition (the premise) guarantees the truth of the other proposition (the conclusion). In other words, if the premise is true, then the conclusion must also be true.

In the case of the question "what is the capital of France?" and the answer "Paris", the premise is the information about the country of France, and the conclusion is the name of its capital city. Since we know that Paris is the capital of France, the truth of the premise (the information about France) logically entails the truth of the conclusion (the name of its capital city). In other words, if the premise is true (i.e. if we are asking about the capital of France), then the conclusion must also be true (i.e. the answer is "Paris").

Human: Is there a relation of logical entailment between "the kids" and "the adults"?

ChatGPT: No, there is no relation of logical entailment between the phrases "the kids" and "the adults". In logic, entailment is a relationship between two propositions (statements or sentences) in which the truth of one proposition (the premise) guarantees the truth of the other proposition (the conclusion). In other words, if the premise is true, then the conclusion must also be true. In the case of the phrases "the kids" and "the adults", there is no logical connection or relationship between the two. The phrases refer to two different groups of people, and there is no inherent logic that would guarantee that one group must be true if the other is true. For example, if we know that "the kids" are present, it does not necessarily follow that "the adults" are also present. Therefore, there is no relation of logical entailment between these two phrases.

Human: Is there a relation of logical entailment between "Close the door now!" and any other sentence?

ChatGPT: Yes, there is a potential relation of logical entailment between the order "Close the door now!" and any other sentence that provides information about the intended action or outcome of the order. In logic, entailment is a relationship between two propositions (statements or sentences) in which the truth of one proposition (the premise) guarantees the truth of the other proposition (the conclusion). In other words, if the premise is true, then the conclusion must also be true. In the case of the order "Close the door now!", the premise is the command to close the door, and the conclusion is the intended action or outcome. Therefore, any sentence that provides information about the intended action or outcome of the order can be considered the conclusion of the entailment relationship. For example, if the order "Close the door now!" is followed by the sentence "The door is now closed", this sentence can be considered the conclusion of the entailment relationship, because the truth of the premise (the command to close the door) logically entails the truth of the conclusion (the door is now closed).

In general, the relation of logical entailment between "Close the door now!" and any other sentence will depend on the specific information provided by the other sentence. If the sentence provides information that is logically entailed by the premise of [here the response from ChatGPT ended rather abruptly].

C Full Output from ChatGPT Prompting with non-assertive premises

In tables 3, 4, and 5, we can see the predictions of ChatGPT for a selection of 96 pairs with non-assertive premises from the validation split of our dataset. For each of the pairs, we prompted ChatGPT with the following template:

(5) Does "PREMISE" entail "HYPOTHESIS"?

ChatGPT always responded with a clear, albeit sometimes somewhat verbose, yes (represented with "1" in the tables) or no (represented with "0" in the tables). The tables show that ChatGPT correctly predicts neutral ("0") in 44 of 96 prompts.

For example, here is the full turn with ChatGPT's explanation of its response for one sample with an incomplete sentence as premise.

(6) (P) : Heaven's Highway (H) Paradise's Road.

Human: Does ": Heavaen's Highway" entail "Paradise's Road?"

ChatGPT It is not clear whether the phrase ": Heaven's Highway" entails that "Paradise's Road". The phrase ": Heaven's Highway" is not a complete sentence and is difficult to interpret or understand. It appears to describe a specific location or path (Heaven's Highway), but it is not clear what this location or path entails or whether it relates to the second phrase ("Paradise's Road"). In order to determine whether the first phrase entails the second phrase, you would need to provide more information about the context in which the phrase is used and the intended meaning of the speaker.

ID	Premise	Hypothesis	True	Pred.
11906e	Would it not be a terrible shock to her?	Would it shock her?	0	0
145506e	Why did our emperor get up so early, and why is he sitting enthroned at the city's main gate, in state, wearing the crown?	The emperor has a crown on his head.	0	0
8021e	Inside both maps of the connections in the alleged right-wing conspiracy against Clinton, profiles of Al Gore (steely-eyed in this time of crisis), and still more pop-psychologizing about Clinton's personality.	The evaluations of Al Gore are part of the psychological profiling of Clinton	0	0
116777e	dirt and noise	mud and sounds	0	0
113150e	From the Place des Abbesses, take Rue Ravignan to 13 Place Emile-Goudeau.	To get to 13 Place Emile-Goudeau from the Place des Abbesses, take Rue Ravignan.	0	1
124577e	Says who?	Who said that?	0	0
95036e	that be all right between them and	Would they both be ok with it?	0	0
144940e	Get out and explore the streets, the open markets, the cafe.	Go and see the streets, open markets, and cafe.	0	1
61713e	Come on, Jeffrey.	Jeffery should hurry up.	0	0
20169e	Are they using financial statements?	Are financial statements being used?	0	1
86509e	You have raced him, senior ?"he asked Drew with formal courtesy.	Have you raced him, sir? he asked Drew.	0	0
57980e	Closed Sabbath.	It's closed on the Sabbath.	0	1
21565e	What day was it when you searched the prisoner's room?	You searched the prisoner's room, right? On what day?	0	0
17577e	Prepared for Office of Air Quality Planning and Standards, US Environmental Protection Agency, Research Triangle Park, NC and Air Quality Management Division, National Park Service, Denver, CO.	The National Park Service is based in Denver Colorado.	0	0
90510e	Then head back to Alicante, just 28 km (17 miles) away.	Go back to Alicante, which is only 28km away.	0	1
108189e	Can't keep even with 'em.	Can't stay even.	0	0
79930e	What was happening to her?	What happened to her?	0	0
37447e	um something with the defense uh	It's related to protection	0	0
110960e	(Thank you.)	Thanks was given.	0	0
52731e	Old values versus new, old virtues and new injustices.	It was the old vs the new.	0	0
145261e	Fuck the gravy	To hell with the gravy.	0	0
27635e	Use of Program Oversight	Overseeing the program.	0	1
144659e	a professional mother a person	A professional mother and a person.	0	0
38885e	back grind tape on and off the wafers	Take the tape on and off the wafers.	0	0
91077e	Randy's Self-Reference Wrap-Up	Randy's Wrap-Up of Himself.	0	1
60546e	in cold frames or whatever the	in cold frames.	0	0
16026e	But how come Kitchell could hide out in Apache country?	How was it the Kitchell was able to stay undetected in Apache lands?	0	0
86644e	sought and respected by the organizations' business managers.	sought and respected by the organizations' business managers.	0	1
797e	Contact the Hong Kong Yacht Club at Tel. 2832 2817 for information.	the contact number for the Hong Kong Yacht Club is 2832 2817	0	1
45789e	And the second point? I asked.	What is the second point?	0	0
65272e	well really just commune with nature	We commune with nature.	0	0
109942e	But does that mean that we face a repeat of the dark years of soup kitchens and brown-shirts leading up to world war?	Does that mean we're going to see something similar to the bad times before the world war?	0	0
123225e	Summary of Major Sections	Summary of important sections	0	1

Table 3: Full output from ChatGPT-Prompting 1/3.

ID	Premise	Hypothesis	True	Pred.
31686e	Interest (unless classified elsewhere), dividends, and rents (except for mineral rights) on Government property.	Mineral rights are excluded from rents on government property.	0	1
88124e	'Yes sir, Mr. Franklin?'	Can I help you Mr. Franklin?	0	0
19958e	Are they often used, may I ask?"	Are the things utilized frequently?	0	0
137712e	Personal Communication with J. Urbas, Reliant Energy, August 13, 2001.	Direct communication with J. Urbas.	0	1
34689e	A sign of failure, of a feeble economy, perhaps?	Is that a sign of a bad economy?	0	0
126927e	Department of Labor, Division of Foreign Labor Certifications, Revised June 1999 [hereinafter FY 1998 H-2A Report].	Department of Labor includes the Division of Foreign Labor Certifications.	0	1
145094e	National Saving and Investment?	Saving and Investment across the country.	0	0
110340e	A Nation of Spendthrifts?	It is a nation of spendthrifts.	0	0
48026e	Then climb (even higher!)	Then climb higher than you are.	0	0
106750e	You think he'd get after her?	The person being spoke to think he'd go after her.	0	0
134219e	U.S. airports sufficient to protect the safety of passengers and equipment?	Is protecting passengers a task that US airports aren't capable of handling?	0	0
96401e	The verdict?	The decision?	0	1
2818e	What was it?	Do you know what it was?	0	0
56782e	How did we lose our rich tradition of porcine references?	There are fewer pig references than there were in the past.	0	0
104984e	What money?	What money do you mean?	0	0
595e	Does Hillary Clinton believe her husband's denials?	It's not sure whether Clinton believes her husband or not.	0	0
123277e	For example, in lieu of hiring a large number of seasonal	Instead of hiring a lot of seasonal	0	1
141791e	So why Clinton's aggressive defense of Helms-Burton?	Why is Clinton so defensive of Helms-Burton?	0	0
125428e	um-hum treatment before for dismissal type thing	Treatment before dismissal thing.	0	0
71972e	Kaaterskill Falls ,by Allegra Goodman (Dial Press).	Goodman wrote a book called Kaaterskill Falls.	0	1
29030	That's the way you think it's gonna be, Croaker?	Is that the way you think it will be, Croaker?	0	1
71243e	Which tradition does John belong to?	John belongs to which institution?	0	0
40851e	Take the metro to Rambuteau and start at the corner of the Rue des Archives and Rue des Francs-Bourgeois, named after the poor people who were allowed to live here tax-free during the 14th century.	Take the metro to Rambuteau and start at the corner of the Rue des Archives.	0	1
85667e	: Heaven's Highway	Paradise's Road.	0	0
110009e	things that uh get you on the edge of your seat a little too much for her	She doesn't like things that get you on the edge of your seat.	0	0
40765e	Many thanks to readers Bill Moran, Darren Thornycroft, and Nicholas Lemann* (author of The Big Test) for flagging this one.	There was reason to flag this.	0	1
34376e	Understand what?	Understand what?	0	1
133820e	Could they take the place of one of the 56 channels of movies?	Could they replace one of the 56 movie channels?	0	1
102665e	Jingoistic Java Juggernaut	The Java Juggernaut is Jingoistic	0	1
75235e	So who does?	Well, who is doing it?	0	1
93725e	, annual mean PM concentration) as inputs to the health and welfare C-R functions of the benefits analysis.	PM concentration is an input to the C-R functions.	0	1
124139e	(The difference between the rates divided by the number of grams in the weight interval).	The rates are divided by the number of grams.	0	1
82736e	10 See the appendix for a further explanation about electronic signatures and GAO's review of such applications.	If you want a further explanation about GAO see the appendix.	0	0
4522e	Rival explanations	explanations that disagree.	0	1
38499e	The Blue Room , by David Hare (Cort Theatre, New York City).	The Blue Room was written by David Hare.	0	1

Table 4: Full output from ChatGPT-Prompting 2/3.

ID	Premise	Hypothesis	True	Pred.
73920e	(1) How long are seasonal agricultural workers typically in the United States?	Do seasonal agricultural workers stay in the US for a while?	0	1
64748e	If I don't, how should I handle it, given that we'll see each other around?	We will end up seeing each other around.	0	1
33128e	Kinda free with a gun, leastwise at showin' it.	They are showing that they are free with a gun.	0	1
118355e	Click on the British flag for an English version of the site.	There is an English version of the website.	0	1
105826e	just to see the show just to see the show right	Only to watch the show, correct?	0	1
58056e	Participate in the graceful tea ceremony or watch the dazzling display of skill in kendo (stick fighting), with its impressively fierce battle cries.	Join in the ceremony of tea or view the kendo performance.	0	1
48341e	Disk compression and networking into Windows.	Disk compression and networking is possible in Windows.	0	1
36136e	Can you describe him at all?"	Can you provide his characteristics?	0	1
50508e	How could productivity indexes—which basically measure the ability of workers to produce a given set of goods—properly take account of such revolutionary innovations as automobiles, antibiotics, air conditioning, and long-playing records?	Productivity indexes measure the ability of workers to make goods.	0	1
39272e	Our A Low-Wage Workforce Without the Brown People.	Brown people make up the low-wage workforce.	0	0
52111e	Acute inflammation and respiratory cell damage	Respiratory cells can be damaged.	0	1
116968e	for each household in the sample.	Individual households are sampled.	0	1
122452e	Evaluation Synthesis.	Synthesis of the evaluation.	0	1
117542e	What would you really choose as a profession, if you could just consult your inclination?	What do you want to do for a living?	0	1
83373e	from front-line employees and managers, and a variety of implementation issues, such as workload demands.	One of the implementation hurdles that will be faced is workload demand.	0	1
109228e	Continue along this road to reach the pretty coastal town of Molyvos (also known by its ancient name, Mithymna), a popular spot for tourists.	Molyvos is a coastal town and a hot spot for tourism.	0	1
46914e	Section 610(e) of the Hearing Aid Compatibility Act of 1988, 47 U.S.C.	The Hearing Aid Compatibility Act was passed in the late 1980's.	0	1
69496e	Why are you coming at me with that pillow?	Why are you tossing the pillow at me?	0	0
115705e	um oh i never heard of that	I have never heard of that.	0	1
25776e	Who was this man who held in his finger these curiously variegated links of an unknown chain?	Who was this man who held these links of chain?	0	1
98779e	Au revoir, my clever and charming young lady. Tuppence sped lightly down the stairs.	Tuppence swiftly left as the watcher whispered goodbye.	0	0
109425e	Best Practices of Leading Commercial Companies	The most dominant commercial companies' best practices	0	1
86775e	Am I an idiot?	Am I an idiot?	0	1
5396e	Greetings, Dave Hanson.	I greet you Dave Hanson.	0	1
105613e	Buchanan or Bush vs. the congressional Republicans.	There are Republicans in Congress.	0	1
132469e	Far from perfect.	It is not perfect.	0	1
112205e	Look out for Robert le Lorrain's fine sculpted horses of Apollo over the old stables in the second courtyard.	Sculptures of horses can be seen in the second courtyard.	0	1
93047e	Who knows?	Who knows?	0	1

Table 5: Full output from ChatGPT-Prompting 3/3.