

Simple Linguistic Inferences of Large Language Models (LLMs): Blind Spots and Blinds

Anonymous ACL submission

Abstract

We evaluate LLMs’ language understanding capacities on simple inference tasks that most humans find trivial. Specifically, we target (i) grammatically-specified entailments, (ii) premises with evidential adverbs of uncertainty, and (iii) monotonicity entailments. We design evaluation sets for these tasks and conduct experiments in both zero-shot and chain-of-thought setups, and with multiple prompts and LLMs. The models exhibit moderate to low performance on these evaluation sets. Subsequent experiments show that embedding the premise in syntactic constructions that should preserve the entailment relations (presupposition triggers) or change them (non-factives), further confuses the models, causing them to either under-predict or over-predict certain entailment labels regardless of the true relation, and often disregarding the nature of the embedding context. Overall these results suggest that, despite LLMs’ celebrated language understanding capacity, even the strongest models have blindspots with respect to certain types of entailments, and certain information-packaging structures act as “blinds” overshadowing the semantics of the embedded premise.

1 Introduction

LLMs have gained immense popularity thanks to their unprecedented ability to understand user queries and generate fluent seemingly-human responses. At the same time, people constantly report LLMs’ failures, anecdotal (Borji, 2023) and systematic, e.g. the lack of reliability and consistency (Shen et al., 2023; Jang and Lukasiewicz, 2023; Plevris et al., 2023), contradictory or unreasonable answers (Zhong et al., 2023), inability to detect false assumptions (Shen et al., 2023), wrong information in prompts (Zuccon and Koopman, 2023), contradictory responses to identical queries (Jang and Lukasiewicz, 2023; Plevris et al., 2023).

However, humans are prone to some failures as

well, e.g., overlooking false assumptions in questions beyond their area of expertise, or failing to find the correct solution to a math problem.

In this work we focus on tasks that are trivial for humans, and do not require any specialized expertise beyond proficiency in English. For example, it is obvious to a human that *Her brother was singing* entails *Someone was singing*, and *Fred’s tie is very long* implies *Fred’s tie is long*, but not vice versa. However, as shall be seen shortly, LLMs fail to establish such systematic relations correctly. LLMs’ errors on such simple tasks are much more indicative of absence of *human-like* text understanding.

We experiment with several types of natural language inferences (NLI), (a.k.a. recognizing textual entailment (Dagan et al., 2005; Bowman et al., 2015)), that are easy for humans, and show that they pose a challenge to LLMs. These NLI tests reveal some of the models’ blind spots, and indicate that they are far from a genuine *human-level* understanding. Moreover, some information-packaging structures, where a statement is embedded within a surrounding linguistic context, may act as “blinds” that hinder the semantics of embedded premises, again in contrast to human-like behavior.

A summary of the conditions we test, as well as the trends observed for one model (gpt-3.5-turbo-0301) is available in Figure 1.¹

We focus on inference types that are solely based on common linguistic phenomena and “trivial” world-knowledge such as class membership (“a dog is an animal”, “navy blue is a shade of blue”). Specifically, we test LLMs’ ability to make three inference types: (i) *Grammatically-specified entailments*, i.e. replacing a constituent of the premise with an indefinite pronoun as *somebody* or *something*. (ii) Premises with *evidential adverbs of un-*

¹Results for other models are presented in Table 2 and discussed in Section 3. While directions and magnitudes differ, all models are far from human level, and do not respond correctly to embedding environments.

Condition	Section	Standalone		Embedded within a Presupposition Trigger		Embedded within a Non-factive Clause	
		Expected	Model	Expected	Model	Expected	Model
Pronouns	3.1	100% Entail	53% Entail	100% Entail	72.3% Entail	100% Neutral	56.7% Entail
Monotonicity Positive	3.2.3	100% Entail	43% Entail	100% Entail	55.7% Entail	100% Neutral	60.2% Entail
Monotonicity Negative	3.2.4	100% Neutral	37% Entail	100% Neutral	52.8% Entail	100% Neutral	52.9% Entail
Uncertainty Adverbs	3.3	100% Neutral	79.9% Entail	100% Neutral	86.9% Entail	---	---

Pronouns (grammatically-specified entailments) Sue was hungry → someone was hungry	<div style="border: 1px solid gray; padding: 5px; margin-bottom: 5px;"> Embedded within a presupposition trigger <i>John realized</i> Sue was hungry → someone was hungry </div> <div style="border: 1px solid gray; padding: 5px;"> Embedded within a non-factive clause <i>John thought</i> Sue was hungry → someone was hungry </div>
Monotonicity (positive) John saw a dog → John saw an animal	
Monotonicity (negative) John saw an animal → John saw a dog	
Uncertainty adverbs John allegedly saw a dog → John saw a dog	

Figure 1: High-level summary of the experiments and results (reported numbers are for gpt-3.5-turbo-0301).

certainty (*supposedly, allegedly* etc.), that block the entailment of the rest of the clause, and (iii) *Monotonicity entailment* (see MacCartney and Manning (2008)) of two kinds: upward, i.e. from subsets to supersets (“Jack is a dog” entails “Jack is an animal”), and downward, i.e. from supersets to subsets (“Jack isn’t an animal” entails “Jack isn’t a dog”). We manually curate test sets for these inference types and experiment with them in a zero-shot setup, observing that LLMs struggle with these phenomena, leading to low accuracy.

We next check how embedding of the premise in a larger grammatical context affects the prediction. Such embedding can take several forms. Contexts consisting of presupposition triggers (e.g. *He realized that [...], They were glad that [...], Something happened before [...]*) serve to *strengthen* the embedded premise, while similarly structured non-factives (e.g. *I feel that [...], He imagined that [...]*) may *cancel* it. We experiment with both context types and show that in most cases they affect the LLMs’ predictions incorrectly. E.g., ChatGPT² has a hard time discerning the two cases, incorrectly treating both as hints towards entailment (for regular prompting) or against it (for chain-of-thought prompting). These or similar trends are observed across different prompts and models (GPT-3.5, GPT-4, LLaMA 2). While GPT-4 clearly improves over other LLMs in terms of accuracy and shows some more sound trends, it still leaves

much room for improvement compared to the human baseline.

These results demonstrate that state-of-the-art LLMs were unable to learn simple linguistic inferences that humans find trivial: they did not acquire them automatically in pre-training, and also in the process of instruct-tuning or human-feedback tuning. Persistence of the problem across prompts and LLMs implies that this is a systematic issue.

While current LLMs struggle with these trivial phenomena, we hope future models will show improvements. Our methodology and the dataset we release³ can help track this progress.

2 Linguistic phenomena considered

We focus on the following linguistic phenomena.

Grammatically-specified entailments The set of the entailments of any sentence includes so-called *grammatically-specified entailments* (Wilson and Sperber, 1979), i.e., entailments where a constituent of the premise is substituted with a variable (such as an indefinite pronoun like *somebody, something* etc.). For instance, the entailments of “*You’ve eaten all my apples*” include, among others:

You’ve eaten all someone’s apples.

You’ve eaten all of something.

You’ve eaten something.

²<https://openai.com/chatgpt>

³URL will be provided in camera ready

135	<i>You've done something.</i>		
136	<i>Someone's eaten all my apples.</i>		
137	Monotonicity entailments hold when less specific predicates are substituted with more specific ones, or vice versa. They can be of two types:		
138			
139			
140	• Upward: more specific predicates can be substituted with less specific ones.		
141			
142	<i>Jack is a dog.</i> \models <i>Jack is an animal.</i>		
143	• Downward: less specific predicates can be substituted with more specific ones.		
144			
145	<i>All animals need water.</i> \models <i>All dogs need water.</i>		
146			
147	Evidential Adverbs “express degrees of certainty with respect to the speaker’s subjective perception of the truth of a proposition” (Haumann, 2007). We test LLMs’ ability to understand evidential adverbs expressing <i>uncertainty</i> (<i>allegedly</i> , <i>purportedly</i> , <i>supposedly</i> etc.). Introducing such adverbs into a clause cancels the entailment of the rest of the clause. E.g., <i>Mike allegedly worked all night</i> does not entail <i>Mike indeed worked all night</i> . The relation between the two statements is neutral.		
148			
149			
150			
151			
152			
153			
154			
155			
156			
157	Presuppositions and Presupposition Triggers		
158	<i>Presupposition</i> (Beaver et al., 2021; Jeretic et al., 2020; Parrish et al., 2021) is a type of inference “whose truth is taken for granted in the utterance of a sentence” (Huang, 2011). Below, a presupposes b (i.e. if b is false, a cannot be felicitously uttered):		
159			
160			
161			
162			
163	a. <i>Jane returned to New York.</i>		
164	\models b. <i>Jane has been to New York before.</i>		
165	Presuppositions are not presented as at-issue content of the utterance, but rather as part of the background, mutually known or assumed by the speaker and the hearer (even if in reality it is not the case). The speaker of a does not <i>inform</i> the hearer that Jane has been to New York before: she <i>assumes</i> it; and if the hearer does not know it, she <i>accommodates</i> it upon hearing the utterance (Fintel, 2008).		
166			
167			
168			
169			
170			
171			
172			
173	Presuppositions are normally evoked by constructions or lexical items, called <i>presupposition triggers</i> (Karttunen, 2016). In sentence a above, the presupposition is triggered by the verb <i>returned</i> , from the class of <i>iterative verbs</i> which presuppose that the action has happened before. Other iterative verbs are <i>relearn</i> , <i>reread</i> , <i>reapply</i> etc. Presupposition triggers used in this work are factives, temporal and other adverbial clauses and embedded <i>wh</i> -questions.		
174			
175			
176			
177			
178			
179			
180			
181			
182			
		Non-factive Verbs and Expressions (Kiparsky and Kiparsky, 1970), such as <i>believe</i> , <i>claim</i> , <i>feel</i> , <i>hope</i> , <i>suspect</i> , <i>think</i> , do not entail either truth or falsity of their complements . For example, given:	183
			184
			185
			186
		a. <i>Jane thinks that Bill bought bread.</i>	187
		b. <i>Bill bought bread.</i>	188
		c. <i>Bill didn't buy bread.</i>	189
		Sentence a does not entail either b or c . The relation between a and b is neutral, and so is the relation between a and c .	190
			191
			192
		Presupposition Triggers, Non-Factives and NLI	193
		It is important to note that embedding a premise under a presupposition trigger does not affect the relations between the premise and hypothesis. By contrast, if we embed the premise under a non-factive, the relation becomes neutral. For example:	194
			195
			196
			197
			198
		a. <i>A balloon hit a light post.</i>	199
		\models b. <i>Something hit a light post</i>	200
		Premise a above entails hypothesis b . If we embed premise b under a presupposition trigger as in a' :	201
			202
		a' . <i><u>She realized that</u> a balloon hit a light post</i>	203
			204
		\models b. <i>Something hit a light post</i>	205
			206
		the relation does not change: the new premise a' still entails b . However, when embedding premise a under a non-factive verb:	207
			208
		a''. <i><u>I suspect</u> a balloon hit a light post</i>	209
		$\not\models$ b. <i>Something hit a light post</i>	210
		the relation becomes neutral: without additional context the new premise a'' does not entail b .	211
			212
		3 Main Experiments	213
		Benchmark Design. We construct our benchmark in the form of several groups of NLI pairs, where each group focuses on a single linguistic phenomenon, and all the items within a group share the same label. This allows us to isolate the models’ performance on the different conditions, rather than just providing aggregate accuracy numbers, which hide the underlying linguistic trends.	214
			215
			216
			217
			218
			219
			220
			221

Zero-shot Setup. We deliberately choose a zero-shot setup rather than in-context learning, for the following reasons. 1) All tested patterns are easily learnable from examples due to their consistent structure within each group, so predictions based on such learning would be a form of “parrotting” rather than “understanding”. 2) It is zero-shot capabilities that indicate if an LLM is truly a general-purpose system (Qin et al., 2023). In real-world scenarios, the goal is not necessarily to handle these inferences directly, but rather to use them as a means to an end. Hence, zero-shot capability becomes crucial. 3) Finally and most importantly, our goal was not to answer “can we make an LLM perform well on these simple cases through prompting” but rather to answer “is this semantic picked up by an LLM automatically as part of its training, from text alone” as some works seemingly suggest (e.g., Merrill et al., 2022).

Models. We test a range of OpenAI models: GPT-3.5 (text-davinci-003), ChatGPT (gpt-3.5-turbo-0301 and gpt-3.5-turbo-0613), and GPT4 (gpt-4-0314 and gpt-4-0613). We access these LLMs through OpenAI’s API⁴ with the default settings⁵, using a single prompt which receives two texts and asks if, given text 1, text 2 is true, false or neutral (see Appendix B for details).⁶

We also evaluate the 70B LLaMA-2 Chat model (Touvron et al., 2023).⁷

Data Collection Methodology. For monotonicity entailment, we used a sample from the MED dataset (Yanaka et al., 2019a) (see §3.2.2, §3.2.3). For grammatically-specified entailments and uncertainty adverbs, the data was created by the first author by manually collecting naturally-occurring sentences from web-pages retrieved by random keywords and selecting a subset of sentences in the page that fit our desired criteria (e.g., for grammatically-specified entailment - sentences containing noun phrases in different positions that could be substituted by indefinite pronouns), and potentially re-writing them (for example, remov-

⁴<https://openai.com/product>

⁵For text-davinci-003 we set the temperature at 0.

⁶This prompt yields a 71% accuracy on 300-instance sample from the SNLI dataset, consistent with previous SNLI prompting literature (Qin et al., 2023; Wang et al., 2023; Jang and Lukasiewicz, 2023).

⁷We access LLaMA 2 through the Replicate API: <https://replicate.com>. Since the prompt used for the OpenAI models yields very poor accuracy (39%) on SNLI for LLaMA 2, we used a different prompt template, with SNLI accuracy of 61%. We use temperature of 0.01 and top-k=1.

ing embedding contexts) to obtain simple sentences. We did not consult an LLM during this process. We then derived premise-hypotheses pairs from these sentences as described in §3.2.1 and §3.2.4. For the embedding contexts, we manually curated a set of candidate environments by collecting naturally occurring presupposition triggers and non-factives from textual material on the same web-pages.

The labels were assigned based on the linguistic properties of the phenomena (e.g. entailment-cancelling property of uncertainty adverbs). The resulting labels for each data group are confirmed by the majority vote in our human annotation experiment (see Section 3.1) in 99.8% of the cases and by all three annotators unanimously in 95.8% of the cases⁸ with a fleiss kappa of 0.94, which corresponds to almost perfect agreement, attesting to the high quality of the data.

3.1 Human Performance

To verify that the inference types we consider are indeed trivial for humans, we sample 100 examples from each of the 11 data types detailed in Figure 1 and have each example labeled by three human annotators from Amazon Mechanical Turk, using majority vote to obtain the final label. We obtain an overall accuracy of 97.55%, and a full agreement among the annotators for 91% of the examples. These very high numbers confirm that these inference types pose no challenge for humans. For results grouped by inference type see Table 2. More details are available in Appendix A.

3.2 Testing LLMs in an NLI setting

Table 2 (top) summarizes the results.⁹ For brevity, we include only one ChatGPT and one GPT-4 version in the table. Results for the other versions can be found in Appendix C.

3.2.1 Grammatically-Specified Entailments

Data: We manually curated a dataset of 100 pairs with grammatically specified entailments replacing an NP in each premise with an indefinite pronoun in the hypothesis (see Table 1, example (1)). This is a seemingly very easy dataset, trivial for any human, where all gold labels are “ENTAILMENT”.

Results: All LLMs except GPT4 perform poorly on these items (maximum 53% accuracy). GPT-4

⁸These numbers differ from the ones reported in §3.1, because here we do not consider the monotonicity data, which was taken from an existing dataset.

⁹All the results reported in this paper are from a single run.

Inference type	Standalone	Under presupposition triggers (premise modified; label unmodified)	Under non-factives (premise modified; label - always neutral)
Grammatically-specified entailments	(1) P: <i>Crown Princess Mary of Denmark has given birth to a healthy baby boy.</i> H: <i>Someone has given birth to a healthy baby boy.</i> L: entailment	(5) P: <i>We are happy that Crown Princess Mary of Denmark has given birth to a healthy baby boy.</i> H: <i>Someone has given birth to a healthy baby boy.</i> L: entailment	(9) P: <i>I hope Crown Princess Mary of Denmark has given birth to a healthy baby boy.</i> H: <i>Someone has given birth to a healthy baby boy.</i> L: neutral
Monotonicity positive	(2) P: <i>She planted blue and purple pansies in the flower bed.</i> H: <i>She planted pansies in the flower bed.</i> L: entailment	(6) P: <i>After she planted blue and purple pansies in the flower bed, she started planting other flowers.</i> H: <i>She planted pansies in the flower bed.</i> L: entailment	(10) P: <i>I think she planted blue and purple pansies in the flower bed.</i> H: <i>She planted pansies in the flower bed.</i> L: neutral
Monotonicity negative	(3) P: <i>Susan made a dress for Jill.</i> H: <i>Susan made a long dress for Jill.</i> L: neutral	(7) P: <i>They are aware that Susan made a dress for Jill.</i> H: <i>Susan made a long dress for Jill.</i> L: neutral	(11) P: <i>They believe Susan made a dress for Jill.</i> H: <i>Susan made a long dress for Jill.</i> L: neutral
Adverbs	(4) P: <i>These persons were allegedly inhabiting the home.</i> H: <i>These persons were inhabiting the home.</i> L: neutral	(8) P: <i>The owner was aware that these persons were allegedly inhabiting the home.</i> H: <i>These persons were inhabiting the home.</i> L: neutral	N/A

Table 1: Inference types (P - premise, H - hypothesis, L - label)

versions handle them much better, with 84% and 100% (human-level) accuracy respectively.

3.2.2 Positive Monotonicity Entailment

Data: We sample 100 positive examples (the gold label is “ENTAILMENT”) from the Monotonicity Entailment Dataset (MED) (Yanaka et al., 2019a) - see example (2) in Table 1.

Results: Except for GPT-4, all models yield very low accuracy (far below 50). The highest result (by gpt-4-0613) of 70% accuracy is still far below human performance.

3.2.3 Negative Monotonicity Entailment

Data: We sample 100 negative examples (the gold label is “NEUTRAL”) from MED (Yanaka et al., 2019a) - see example (3) in Table 1.

Results: Here, the earlier GPT-4 version (gpt-4-0314) only slightly underperforms humans (88%). All other models (including gpt-4-0613) have much lower accuracy: 28%- 56%.¹⁰

3.2.4 Adverbs of Uncertainty

Data: We manually create a dataset of 100 sentence pairs where the premise contains an uncertainty adverb, while the hypothesis omits it (exam-

ple (4) in Table 1). We apply 9 adverbs (*allegedly, hopefully, possibly, presumably, probably, purportedly, reportedly, seemingly, supposedly*) to each of the 100 pairs, obtaining 900 examples, 100 per adverb. The gold label for all pairs is “NEUTRAL”.

Results: Strongest result (70% accuracy) is obtained by the earlier GPT-4 (0314), and is substantially below human performance. All other models, including the newer GPT-4, achieve very low accuracy on this data (4.67% - 38%).

Bottomline: Except for GPT-4, all tested models perform very poorly on all the explored inference types. In many cases the LLMs (especially LLaMA 2) predict mostly *the opposite of the expected label* (“ENTAILMENT” where “NEUTRAL” is expected and vice versa). While the two GPT-4 versions improve significantly over other LLMs, neither handles all the inference types equally well: the later version (gpt-4-0613) is very good at handling the cases where “ENTAILMENT” is expected, but its predecessor (gpt-4-0314) outperforms it on “NEUTRAL” cases.

3.3 Embedding the Premises under Presupposition Triggers

Table 2 (middle) summarizes the results (except for gpt-3.5-turbo-0613 and gpt-4-0314 covered in

¹⁰Similar results on MED are reported by Liu et al. (2023).

358 Appendix C).

359 **Data:** Next we modify all the datasets described
360 above by embedding the original premises under
361 presupposition triggers (see examples (5)–(8) in
362 Table 1). We use 23 trigger types: 21 factives, such
363 as *know*, *realize*, *be glad*; embedded wh-questions,
364 such as *This explains why he came*; and adverbial
365 clauses, e.g. *After he came, I cooked dinner*.

366 For uncertainty adverbs, we randomly sample
367 100 examples from our dataset of 900 sentence
368 pairs, and apply each of the 23 triggers to each
369 sampled pair, obtaining 2300 pairs. For other data
370 types (grammatically specified entailments, posi-
371 tive and negative monotonicity entailments) we
372 embed the 100 original premises under 23 triggers,
373 which also results in 2300 pairs per data type. The
374 hypotheses remain unchanged (see Table 1).¹¹

375 Presupposition triggers do not change the rela-
376 tion between the premise and the hypothesis, so the
377 gold label remains unchanged: “ENTAILMENT”
378 for grammatically specified and positive mono-
379 tonicity entailments; “NEUTRAL” for negative
380 monotonicity entailments and uncertainty adverbs.

381 **Results:** The OpenAI models show the same pat-
382 tern across all data types: under presupposition trig-
383 gers they predict “ENTAILMENT” more *regard-*
384 *less of the correct label*. This improves accuracy
385 where “ENTAILMENT” is expected (grammati-
386 cally specified and positive monotonicity entail-
387 ments), but reduces accuracy for “NEUTRAL” data
388 (negative monotonicity, uncertainty adverbs).¹²

389 LLaMA 2 shows a “healthier” trend: under pre-
390 supposition triggers it seems to capture all infer-
391 ence types better than in the standalone experi-
392 ments, predicting more entailment for the “posi-
393 tive” datasets and more “neutral” relations for the
394 “neutral” datasets. However, its accuracy remains
395 consistently low (below 40%) across all data types.
396 Additionally, under presupposition triggers, it con-
397 sistently predicts more “contradiction” labels for
398 all inference types, which is always wrong.

399 **Bottomline:** Under presupposition triggers the
400 OpenAI LLMs tend to predict more entailment ir-
401 respective of the correct label. LLaMA 2 improves

¹¹In experiments with both embedding context types we use full data for text-davinci-003, gpt-3.5-turbo-0301 and LLaMA 2, and 100 examples per inference type for other models.

¹²The only exception is gpt-3.5-turbo-0613: under presupposition triggers it predicts less entailment (except for negative monotonicity), which in most cases hurts the accuracy (see Appendix C).

402 in this setting, but shows a different wrong trend:
403 predicting more “contradiction” labels in all cases.

404 3.4 Embedding the Premises under 405 Non-Factives

406 Table 2 (bottom) summarizes the results (gpt-3.5-
407 turbo-0613 and gpt-4-0314 are covered in Ap-
408 pendix C).

409 **Data:** We modify the original datasets (see §3.2),
410 embedding the premises under non-factives (see
411 examples (9)–(11) in Table 1). For grammatically-
412 specified entailments and both types of mono-
413 tonicity data, we embed each of the 100 original
414 premises under 23 different non-factives, e.g. *feel*,
415 *hope*, *believe* etc., obtaining 2300 pairs for each
416 data type. We omit this experiment for adverbs for
417 semantic reasons: including both an uncertainty
418 adverb and a non-factive into the premise (*I guess*
419 *he allegedly worked all night*.) results in double
420 expression of uncertainty, creating a tautology.

421 The hypotheses remain unchanged (see Table 1).

422 Since non-factives cancel entailment, the gold
423 label in all the cases changes to “NEUTRAL”.

424 **Results:** Surprisingly, GPT-3.5 and ChatGPT¹³
425 again predict “ENTAILMENT” more often for all
426 the data groups.¹⁴ This hurts the accuracy for all
427 data groups, since now the correct label is always
428 “NEUTRAL”. Both GPT-4 versions show the *cor-*
429 *rect* trend predicting *less* “ENTAILMENT” and
430 *more* “NEUTRAL” labels under non-factives. Yet,
431 the now-incorrect “ENTAILMENT” label remains
432 very frequent reducing the accuracy.

433 LLaMA 2 again (as with presupposition trig-
434 gers) seems more accurate compared to standalone
435 experiments, but a closer look reveals it is likely
436 “right for the wrong reason”: the proportion of neu-
437 tral predictions has changed very little compared
438 to the standalone experiments (maximum by 7%),
439 but now the “NEUTRAL” label is always correct,
440 which seemingly boosts the accuracy.

¹³The newer ChatGPT version, gpt-3.5-turbo-0601, shows the same trends as under presupposition triggers, predicting less entailment (except for negative monotonicity) (see Appendix C). It shows that, just like its predecessors, this LLM is unable to distinguish the two types of embedding contexts.

¹⁴This inclination to overpredict entailment might imply the use of lexical overlap heuristics, given the lexical similarity between premises and hypotheses in our data. However, a closer examination reveals a contrary trend: standalone premises have greater lexical overlap with hypotheses than embedded ones, yet the models predict more entailment for embedded premises. Our evaluation on a subset of HANS (McCoy et al., 2019) also suggest that lexical overlap does not significantly influence ChatGPT towards entailment.

3.4.1 Bottomline

Only the GPT-4 models capture the overall trend correctly reducing entailment predictions and increasing neutral predictions under non-factives. However, while, ideally, the entailment predictions should be now reduced to 0, they still remain quite frequent, especially in the supposedly more advanced gpt-4-0613, reducing the models’ accuracy.

Across all the experiments described above, GPT-4 models show the best results and the most human-like trends. They are also the only ones consistently refraining from the always incorrect “CONTRADICTION” label. However, the earlier version (see Appendix C) proves biased towards neutral predictions, while the later one constantly leans towards entailment. Under presupposition triggers they share their predecessors’ erroneous trend: predicting more entailment regardless of the correct label. In spite of the favourable dynamics both GPT-4 versions still perform far below the human level on most inference types we explored.

Notably, overall, the newer ChatGPT and GPT-4 versions prove less accurate than the respective earlier ones (see Appendix C).

4 Prompt Variations

We also experiment with prompt variations.

Prompt paraphrasing. We ask ChatGPT (gpt-3.5-turbo-0301) to rephrase our prompt template, obtaining two templates that we verify to be semantically equivalent w.r.t. the task (details in Appendix D). While accuracies vary between prompts, the scores on all tasks remain low.

Chain-of-Thought Prompting (CoT). Using gpt-3.5-turbo-0301, we investigate CoT prompting (Kojima et al., 2023), and find that it *reverses* the LLM’s trend to predict more entailment for embedded premises (see Section 3): now the LLM predicts *less* entailment under presupposition triggers or non-factives, while the number of neutral predictions grows - again, regardless of the correct label, i.e. the reversed trend is equally inaccurate. CoT prompting improves the accuracy, but in a one-sided way: scoring higher on “neutral” test sets, but lower on almost all the “entailment” ones.

Analysis of CoT Results. The CoT technique allows us to explore the model’s “reasoning”. We manually evaluate a subset of the CoT explanations. In half of the cases (50.9%) both the final decision and the CoT explanation were wrong. In 23.6% a

correct explanation was followed by a correct decision; in 23.6% a wrong explanation was followed by a correct decision. In 1.86% of the cases a correct explanation was followed by a wrong decision. 81% of the cases reflected a correct understanding of the task expressed in the prompt. In half of the cases (49.1%) the CoT mentioned the underlying linguistic phenomena explicitly, but only in half of those (23.6% of the total) reflected their correct understanding and only 14.5% of the time used them as a basis for the final prediction.

The details of the CoT experiments and the manual analysis are available in Appendix E.

5 Relation to previous findings

A vast literature evaluated linguistic abilities of earlier generations of pretrained models¹⁵, including some phenomena we consider here: Jiang and de Marneffe (2019); de Marneffe et al. (2019); Ross and Pavlick (2019); McCoy et al. (2019); Yanaka et al. (2021) (**embedded clauses**); Jeretic et al. (2020); Parrish et al. (2021); Kabbara and Cheung (2022) (**presuppositions and implicatures**); Yanaka et al. (2019a,b); Goodwin et al. (2020); Yanaka et al. (2020); Jumelet et al. (2021); Geiger et al. (2020) (**monotonicity entailment**); Ettinger (2020); Guerin and Chemla (2023) (**hypernymy**). Some works focused on other phenomena: **negation, synonyms, antonyms, tenses, temporal adverbials, active vs. passive voice, symmetric and asymmetric relations, comparatives, superlatives** etc. (e.g., Kassner and Schütze, 2020; Hartmann et al., 2021; Hossain et al., 2020, 2022; Ettinger, 2020; Truong et al., 2022; Jang et al., 2022; Ravichander et al., 2022; Ribeiro et al., 2020). Many of these studies make use of the NLI task to evaluate LMs’ basic linguistic abilities (e.g., McCoy et al., 2019; Jiang and de Marneffe, 2019; Yanaka et al., 2021; Kabbara and Cheung, 2022; Jeretic et al., 2020; Parrish et al., 2021; Kabbara and Cheung, 2022; Yanaka et al., 2019a,b). We draw inspiration from this line of work.

In the remainder of the section, we focus on work that evaluates SOTA LLMs (starting with GPT-3).

Truong et al. (2023) seek to determine if **negation** remains a challenge in the age of LLMs.¹⁶ They assess GPTneo, GPT-3, and InstructGPT

¹⁵There are comprehensive overviews of such literature, for example, Rogers et al. (2020); Chang and Bergen (2023).

¹⁶In fact, their work covers a range of negation-related phenomena, such as synonym/antonym lexical relations, monotonicity entailment and hypernymy under negation.

		human			GPT-3.5			ChatGPT			GPT-4			LLaMA 2 70b		
		e	n	c	e	n	c	e	n	c	e	n	c	e	n	c
Standalone	pronouns	100.0	0.0	0.0	39.0	56.0	5.0	53.0	34.0	13.0	100.0	0.0	0.0	31.0	61.0	8.0
	monotonicity positives	89.0	11.0	0.0	25.0	46.0	29.0	43.0	51.0	6.0	70.0	30.0	0.0	36.0	59.0	5.0
	monotonicity negatives	5.0	94.0	1.0	22.0	28.0	50.0	37.0	42.0	21.0	33.0	56.0	11.0	46.0	35.0	19.0
	uncertainty adverbs	0.0	100.0	0.0	77.6	4.7	17.8	79.9	9.1	11.0	61.0	38.0	1.0	89.9	9.0	1.1
Under presupposition triggers	pronouns	100.0	0.0	0.0	65.4	27.8	6.8	72.3	23.3	4.3	100.0	0.0	0.0	32.0	58.1	9.9
	monotonicity positives	97.0	3.0	0.0	38.0	25.1	36.8	55.7	37.2	7.1	71.0	29.0	0.0	36.9	44.9	18.3
	monotonicity negatives	5.0	95.0	0.0	39.8	19.0	41.2	52.8	29.4	17.8	36.0	53.0	11.0	32.4	38.5	29.0
	uncertainty adverbs (sample)	1.0	99.0	0.0	87.8	2.7	9.6	86.9	7.5	5.6	70.0	30.0	0.0	58.4	29.4	12.3
Under non-factives	pronouns	0.0	100.0	0.0	51.0	41.0	8.0	56.7	40.0	3.3	97.0	3.0	0.0	29.7	64.7	5.5
	monotonicity positives	1.0	99.0	1.0	28.8	27.3	43.9	60.2	30.2	9.7	44.0	56.0	0.0	34.4	52.3	13.3
	monotonicity negatives	0.0	100.0	0.0	30.3	17.4	52.3	52.9	22.3	24.7	23.0	69.0	8.0	29.3	42.0	28.7

Table 2: Experiment results across all models. Background indicates the expected label (green - "entailment", blue - "neutral"). The green and red font colors indicate increase or decrease compared to standalone premises. The numbers in **bold** also indicate *accuracy* (e.g. for uncertainty adverbs accuracy coincides with the percentage of neutrals). The **underlined bold** numbers indicate *the best accuracy for a specific inference type* across all models (human baseline excluded). **e** - entailment (%); **n** - neutral (%); **c** - contradiction (%). Model versions shown: GPT-3.5 – text-davinci-003; ChatGPT – gpt-3.5-turbo-0301; GPT-4 - gpt-4-0613. Result for additional GPT models (gpt-3.5-turbo-0613 and gpt-4-0314) available in Appendix C.

across various negation benchmarks with different model sizes and prompts, and discover a number of important limitations, including insensitivity to negation, an inability to capture synonym/antonym lexical relations and to reason under negation. Ye et al. (2023) find that LLMs¹⁷ with CoT-style prompts struggle with **lexical negation** (words with **negative suffixes**: *implausible, unreasonable* etc.). Lorge and Pierrehumbert (2023) explore the capability of LMs (including text-davinci-002) to handle **scalar adverbs** (specifically, **adverbs of frequency, modality and degree**). Despite some success they notice many shortcomings: “weak differentiation amongst the semantic classes of adverbs, poor ability to discriminate scalar adverbs from negations”, “strong effects of adverb frequencies and lack of generalisation across two logically equivalent entailment constructions”. Cong (2022) probes GPT-3 for understanding of scalar **implicatures** and **presuppositions** and find that its performance is “mostly at chance”. Stengel-Eskin and Van Durme (2022) investigate how GPT3-davinci, GPT-Neo, Jurassic, T5 for QA, T0pp handle **subject-control sentences**. Davis (2022) analyzes how autoregressive models with GPT-like architectures, such as GPT-2 XL, GPT-Neo, GPT-J, GPT-3 (text-davinci-002) handle **binding** and conclude that they consider antecedents which humans reject as structurally ungrammatical. Focusing on “language illusions”, Zhang et al. (2023) find that LMs including text-davinci-003 “have some capacity to process **comparative structures**”, but struggle with **multiple negation** and **NPI licensors**.

¹⁷They test 14 LLMs including GPT-4 and 3.5 and four LLaMA variants.

This overview suggests that for newer LLMs (GPT-3 and later) there is much less research focusing on fundamental linguistic abilities. We add to this line of work, focusing on very simple inference types that to the best of our knowledge were not previously evaluated (grammatically-specified entailments, uncertainty adverbs) as well as evaluating stability under embedding contexts (presupposition-triggers and non-factives).

At the same time, the existing research shows that current pre-training objectives do not necessarily capture the basic linguistic phenomena, thus “new training paradigms are essential to achieve better linguistic competency” (Truong et al., 2023).

6 Conclusion

We tested LLMs with several simple linguistic inferences that humans find trivial, and further experimented by embedding the premises within (simple) linguistic contexts that should either preserve or cancel the inference. The models performed poorly in all cases. We conclude that LLMs do not learn entailment semantics “naturally”. The persistence of the issue across prompts, models and setups shows that these limitations are robust and this topic merits further systematic investigations. While the results are tested for several specific models, we do expect them to hold more generally. We believe the ability of LLMs to correctly understand simple textual inferences humans find natural is important. We provide a benchmark and methodology that can be used to track the improvement of future models.

7 Limitations

As a consequence of LLMs’ sensitivity to prompt phrasings, there may exist prompts that can potentially modify the reported results. At the same time we agree with Jang and Lukasiewicz (2023) who point out that “improvements with prompt design can be considered another violation of semantic consistency, because the prompts will deliver identical semantic meaning, i.e., task description”.

We showed that embedding the premises under presupposition triggers or non-factives affects the models’ predictions exhibiting certain patterns. However, our treatment of non-factives is non-exhaustive: for example, we did not try adding negation to the non-factives or using non-factive predicates denoting a high level of uncertainty (*I doubt, I’m skeptical*) or containing negative prefixes (*I am uncertain, I disbelieve*). It’s possible that implicit or explicit negation in the embedding predicates may change the LLMs’ behavior. Also, while considering presupposition triggers and non-factives, we did not consider other types of clause-embedding predicates (e.g., implicative verbs).

Finally, since LLMs undergo continuous updates, the test results presented here may vary over time, and the closed, black-box nature of the models’ training details and training data limits the scope of the conclusions that can be reached. We cannot, for example, know for certain if the low performance on these inference type stems already from the base model and its inability to learn these type of inferences from raw text, or was added during instruct tuning. That said, our data and methodology for benchmarking these capabilities is model-agnostic and remains intact.

Ethics Statement

Annotator Qualification and Compensation for Annotation To collect annotations for our human baseline (see §3.1) we used Amazon Mechanical Turk (AMT). We employed 5 annotators who showed themselves reliable in handling annotation tasks assigned by us in the past. All annotators had the following qualifications: (1) over 500 completed HITs; (2) 99% approval rate or higher; (3) native English speakers. We paid \$0.10 per example. Labeling each example normally (excluding outliers) took about 21.7 seconds which resulted in a payment of about \$16.5 per hour. Upon completing the task the annotators were given bonuses.

Annotation Data Collection and Usage Policy

Annotators were notified that their annotations are meant for research purposes in the field of Natural Language Processing. The task and collected annotations were objective and included no personal information.

References

- David I. Beaver, Bart Geurts, and Kristie Denlinger. 2021. Presupposition. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Spring 2021 edition. Metaphysics Research Lab, Stanford University.
- Ali Borji. 2023. [A categorical archive of chatgpt failures](#).
- Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. [A large annotated corpus for learning natural language inference](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.
- Tyler A. Chang and Benjamin K. Bergen. 2023. [Language Model Behavior: A Comprehensive Survey](#). *Computational Linguistics*, pages 1–55.
- Yan Cong. 2022. [Psycholinguistic diagnosis of language models’ commonsense reasoning](#). In *Proceedings of the First Workshop on Commonsense Representation and Reasoning (CSRR 2022)*, pages 17–22, Dublin, Ireland. Association for Computational Linguistics.
- Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer.
- Forrest Davis. 2022. [Incremental processing of principle B: Mismatches between neural models and humans](#). In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 144–156, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Marie-Catherine de Marneffe, Mandy Simons, and Judith Tonhauser. 2019. [The commitmentbank: Investigating projection in naturally occurring discourse](#). *Proceedings of Sinn und Bedeutung*, 23(2):107–124.
- Allyson Ettinger. 2020. [What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models](#). *Transactions of the Association for Computational Linguistics*, 8:34–48.
- Kai Fintel. 2008. [What is presupposition accommodation, again?](#) *Philosophical Perspectives*, 22:137–170.

701	Atticus Geiger, Kyle Richardson, and Christopher Potts.	Nanjiang Jiang and Marie-Catherine de Marneffe.	758
702	2020. Neural natural language inference models partially embed theories of lexical entailment and negation.	2019. Evaluating BERT for natural language inference: A case study on the CommitmentBank. In <i>Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)</i> , pages 6086–6091, Hong Kong, China. Association for Computational Linguistics.	759
703			760
704			761
705	Emily Goodwin, Koustuv Sinha, and Timothy J. O'Donnell. 2020. Probing linguistic systematicity.		762
706			763
707	Nicolas Guerin and Emmanuel Chemla. 2023. It is a bird therefore it is a robin: On BERT's internal consistency between hypernym knowledge and logical words. In <i>Findings of the Association for Computational Linguistics: ACL 2023</i> , pages 8807–8817, Toronto, Canada. Association for Computational Linguistics.	Jaap Jumelet, Milica Denic, Jakub Szymanik, Dieuwke Hupkes, and Shane Steinert-Threlkeld. 2021. Language models use monotonicity to assess NPI licensing. In <i>Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021</i> , pages 4958–4969, Online. Association for Computational Linguistics.	764
708			765
709			766
710			767
711			768
712			769
713			770
714	Mareike Hartmann, Miryam de Lhoneux, Daniel Herscovich, Yova Kementchedjhieva, Lukas Nielsen, Chen Qiu, and Anders Søgaard. 2021. A multilingual benchmark for probing negation-awareness with minimal pairs. In <i>Proceedings of the 25th Conference on Computational Natural Language Learning</i> , pages 244–257, Online. Association for Computational Linguistics.	Jad Kabbara and Jackie Chi Kit Cheung. 2022. Investigating the performance of transformer-based NLI models on presuppositional inferences. In <i>Proceedings of the 29th International Conference on Computational Linguistics</i> , pages 779–785, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.	771
715			772
716			773
717			774
718			775
719			776
720			777
721			778
722	Dagmar Haumann. 2007. Adverb Licensing and Clause Structure in English.	Lauri Karttunen. 2016. Presupposition: What went wrong? <i>Semantics and Linguistic Theory</i> , 26:705–731.	779
723			780
724	Md Mosharaf Hossain, Dhivya Chinnappa, and Eduardo Blanco. 2022. An analysis of negation in natural language understanding corpora. In <i>Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)</i> , pages 716–723, Dublin, Ireland. Association for Computational Linguistics.	Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In <i>Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics</i> , pages 7811–7818, Online. Association for Computational Linguistics.	781
725			782
726			783
727			784
728			785
729			786
730			787
731	Md Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, and Eduardo Blanco. 2020. An analysis of natural language inference benchmarks through the lens of negation. In <i>Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)</i> , pages 9106–9118, Online. Association for Computational Linguistics.	Paul Kiparsky and Carol Kiparsky. 1970. <i>FACT</i> , pages 143–173. De Gruyter Mouton, Berlin, Boston.	788
732			789
733			790
734			791
735			792
736			793
737			794
738			795
739	Yan Huang. 2011. 14. Types of inference: entailment, presupposition, and implicature , pages 397–422. De Gruyter Mouton, Berlin, New York.	Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners.	796
740			797
741			798
742	Myeongjun Jang and Thomas Lukasiewicz. 2023. Consistency analysis of chatgpt.	Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. 2023. Evaluating the logical reasoning ability of chatgpt and gpt-4.	799
743			800
744	Myeongjun Jang, Frank Mtumbuka, and Thomas Lukasiewicz. 2022. Beyond distributional hypothesis: Let language models learn meaning-text correspondence. In <i>Findings of the Association for Computational Linguistics: NAACL 2022</i> , pages 2030–2042, Seattle, United States. Association for Computational Linguistics.	Isabelle Lorge and Janet B. Pierrehumbert. 2023. Not wacky vs. definitely wacky: A study of scalar adverbs in pretrained language models. <i>ArXiv</i> , abs/2305.16426.	801
745			802
746			803
747			804
748			805
749			806
750		Bill MacCartney and Christopher D. Manning. 2008. Modeling semantic containment and exclusion in natural language inference. In <i>Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)</i> , pages 521–528, Manchester, UK. Coling 2008 Organizing Committee.	807
751	Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models IMPPRESSive? Learning IMPLICature and PRESupposition. In <i>Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics</i> , pages 8690–8705, Online. Association for Computational Linguistics.	Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In <i>Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics</i> , pages 3428–3448, Florence, Italy. Association for Computational Linguistics.	808
752			809
753			810
754			811
755			812
756			
757			

813	William Merrill, Alex Warstadt, and Tal Linzen. 2022.	Ferrer, Moya Chen, Guillem Cucurull, David Esiobu,	869
814	Entailment semantics can be extracted from an ideal	Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller,	870
815	language model . In <i>Proceedings of the 26th Confer-</i>	Cynthia Gao, Vedanuj Goswami, Naman Goyal, An-	871
816	<i>ence on Computational Natural Language Learning</i>	thony Hartshorn, Saghar Hosseini, Rui Hou, Hakan	872
817	(<i>CoNLL</i>), pages 176–193, Abu Dhabi, United Arab	Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa,	873
818	Emirates (Hybrid). Association for Computational	Isabel Kloumann, Artem Korenev, Punit Singh Koura,	874
819	Linguistics.	Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Di-	875
820	Alicia Parrish, Sebastian Schuster, Alex Warstadt, Omar	ana Liskovich, Yinghai Lu, Yuning Mao, Xavier Mar-	876
821	Agha, Soo-Hwan Lee, Zhuoye Zhao, Samuel R. Bow-	tinet, Todor Mihaylov, Pushkar Mishra, Igor Moly-	877
822	man, and Tal Linzen. 2021. NOPE: A corpus of	bog, Yixin Nie, Andrew Poulton, Jeremy Reizen-	878
823	naturally-occurring presuppositions in English . In	stein, Rashi Rungta, Kalyan Saladi, Alan Schelten,	879
824	<i>Proceedings of the 25th Conference on Computa-</i>	Ruan Silva, Eric Michael Smith, Ranjan Subrama-	880
825	<i>tional Natural Language Learning</i> , pages 349–366,	nian, Xiaoqing Ellen Tan, Binh Tang, Ross Tay-	881
826	Online. Association for Computational Linguistics.	lor, Adina Williams, Jian Xiang Kuan, Puxin Xu,	882
827	Vagelis Plevris, George Papazafeiropoulos, and Alejan-	Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan,	883
828	dro Jiménez Rios. 2023. Chatbots put to the test in	Melanie Kambadur, Sharan Narang, Aurelien Ro-	884
829	math and logic problems: A preliminary comparison	driguez, Robert Stojnic, Sergey Edunov, and Thomas	885
830	and assessment of chatgpt-3.5, chatgpt-4, and google	Scialom. 2023. Llama 2: Open foundation and fine-	886
831	bard .	tuned chat models .	887
832	Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao	Thinh Truong, Timothy Baldwin, Trevor Cohn, and	888
833	Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is	Karin Verspoor. 2022. Improving negation detection	889
834	chatgpt a general-purpose natural language process-	with negation-focused pre-training . In <i>Proceedings</i>	890
835	ing task solver?	<i>of the 2022 Conference of the North American Chap-</i>	891
836	Abhilasha Ravichander, Matt Gardner, and Ana Maraso-	<i>ter of the Association for Computational Linguistics:</i>	892
837	vić. 2022. Condaqa: A contrastive reading compre-	<i>Human Language Technologies</i> , pages 4188–4193,	893
838	hension dataset for reasoning about negation .	Seattle, United States. Association for Computational	894
839	Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin,	Linguistics.	895
840	and Sameer Singh. 2020. Beyond accuracy: Be-	Thinh Hung Truong, Timothy Baldwin, Karin Verspoor,	896
841	havioral testing of NLP models with CheckList . In	and Trevor Cohn. 2023. Language models are not	897
842	<i>Proceedings of the 58th Annual Meeting of the Asso-</i>	naysayers: an analysis of language models on nega-	898
843	<i>ciation for Computational Linguistics</i> , pages 4902–	tion benchmarks . In <i>Proceedings of the 12th Joint</i>	899
844	4912, Online. Association for Computational Lin-	<i>Conference on Lexical and Computational Seman-</i>	900
845	guistics.	<i>tics (*SEM 2023)</i> , pages 101–114, Toronto, Canada.	901
846	Anna Rogers, Olga Kovaleva, and Anna Rumshisky.	Association for Computational Linguistics.	902
847	2020. A primer in bertology: What we know about	Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen,	903
848	how bert works .	Runkai Zheng, Yidong Wang, Linyi Yang, Haojun	904
849	Alexis Ross and Ellie Pavlick. 2019. How well do NLI	Huang, Wei Ye, Xiubo Geng, Binxin Jiao, Yue Zhang,	905
850	models capture verb veridicality? In <i>Proceedings of</i>	and Xing Xie. 2023. On the robustness of chatgpt:	906
851	<i>the 2019 Conference on Empirical Methods in Natu-</i>	An adversarial and out-of-distribution perspective .	907
852	<i>ral Language Processing and the 9th International</i>	Deirdre Wilson and Dan Sperber. 1979. Ordered Entail-	908
853	<i>Joint Conference on Natural Language Processing</i>	ments: An Alternative to Presuppositional Theories ,	909
854	(<i>EMNLP-IJCNLP</i>), pages 2230–2240, Hong Kong,	volume 11, pages 299–323.	910
855	China. Association for Computational Linguistics.	Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, and	911
856	Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang	Kentaro Inui. 2020. Do neural models learn sys-	912
857	Zhang. 2023. In chatgpt we trust? measuring and	tematicity of monotonicity inference in natural lan-	913
858	characterizing the reliability of chatgpt .	guage? In <i>Proceedings of the 58th Annual Meeting of</i>	914
859	Elias Stengel-Eskin and Benjamin Van Durme. 2022.	<i>the Association for Computational Linguistics</i> , pages	915
860	The curious case of control . In <i>Proceedings of the</i>	6105–6117, Online. Association for Computational	916
861	<i>2022 Conference on Empirical Methods in Natu-</i>	Linguistics.	917
862	<i>ral Language Processing</i> , pages 11065–11076, Abu	Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Ken-	918
863	Dhabi, United Arab Emirates. Association for Com-	taro Inui, Satoshi Sekine, Lasha Abzianidze, and	919
864	putational Linguistics.	Johan Bos. 2019a. Can neural networks understand	920
865	Hugo Touvron, Louis Martin, Kevin Stone, Peter Al-	monotonicity reasoning? In <i>Proceedings of the 2019</i>	921
866	bert, Amjad Almahairi, Yasmine Babaei, Nikolay	<i>ACL Workshop BlackboxNLP: Analyzing and Inter-</i>	922
867	Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti	<i>preting Neural Networks for NLP</i> , pages 31–40, Flo-	923
868	Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton	rence, Italy. Association for Computational Linguis-	924
		tics.	925
		Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Ken-	926
		taro Inui, Satoshi Sekine, Lasha Abzianidze, and	927

928 Johan Bos. 2019b. [HELP: A dataset for identifying](#)
929 [shortcomings of neural models in monotonic-](#)
930 [ity reasoning](#). In *Proceedings of the Eighth Joint*
931 *Conference on Lexical and Computational Semantics*
932 *(*SEM 2019)*, pages 250–255, Minneapolis, Min-
933 nesota. Association for Computational Linguistics.

934 Hitomi Yanaka, Koji Mineshima, and Kentaro Inui.
935 2021. [Exploring transitivity in neural nli models](#)
936 [through veridicality](#).

937 Mengyu Ye, Tatsuki Kuribayashi, Jun Suzuki, Goro
938 Kobayashi, and Hiroaki Funayama. 2023. [Assessing](#)
939 [step-by-step reasoning against lexical negation: A](#)
940 [case study on syllogism](#). In *Proceedings of the 2023*
941 *Conference on Empirical Methods in Natural Lan-*
942 *guage Processing*, pages 14753–14773, Singapore.
943 Association for Computational Linguistics.

944 Yuhan Zhang, Edward Gibson, and Forrest Davis. 2023.
945 [Can language models be tricked by language illu-](#)
946 [sions? easier with syntax, harder with semantics.](#)
947 In *Conference on Computational Natural Language*
948 *Learning*.

949 Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and
950 Dacheng Tao. 2023. [Can chatgpt understand too? a](#)
951 [comparative study on chatgpt and fine-tuned bert](#).

952 Guido Zuccon and Bevan Koopman. 2023. [Dr chatgpt,](#)
953 [tell me what i want to hear: How prompt knowledge](#)
954 [impacts health answer correctness](#).

955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996

A Human annotation details

In order to construct a human baseline, we sampled 1100 examples (100 per inference type) and had 3 human annotators label each example, asking if based on Text 1, Text 2 is true, false or neutral. We then use majority vote to determine the final human label for each example.

For human annotation, we used the Amazon Mechanical Turk platform. We did not use any qualification tests and filters for annotator selection, but instead employed 5 annotators who showed themselves reliable in handling annotation tasks assigned by us in the past. We paid \$0.10 per example. Labeling each example normally (excluding outliers) took about 21.7 seconds which resulted in a payment of about \$16.5 per hour.

We obtained an overall accuracy of 97.55%. In 91% of the cases all three annotators confirmed the expected labels unanimously. The instructions received by the annotators are shown in Figure 2.

B Main entailment experiments prompts

The prompt below is used throughout the experiments described in Section 3 for all the OpenAI models (see Section 3).

You are given a pair of texts. Say about this pair: given Text 1, is Text 2 true, false or neutral (you can't tell if it's true or false)? Reply in one word.

Text 1: "text1"

Text 2: "text2"

For LLaMA 2 we use the following prompt:

You have two texts, and your task is to determine the truthfulness of Text 2 based on Text 1. Provide a one-word response indicating whether Text 2 is true, false, or neutral (indeterminable). Here are the texts:

Text 1: "text1"

Text 2: "text2"

The model outputs one of three possible labels: "true" (corresponding to "entailment"), "false" (corresponding to "contradiction") or "neutral".¹⁸

¹⁸In the rare cases when the model outputs a different label, we normalize it to one of the three expected forms. E.g. "truthful" is normalized to "true".

C Complete Results Table

All the results from all LLMs including both ChatGPT and GPT-4 versions are shown in Table 3.

Accuracy across all tested LLMs obtained in the experiments in Section 3 is compared in Table 4.

D Prompt Paraphrasing

We experimented with the following prompt variations: (Original prompt)

You are given a pair of texts. Say about this pair: given Text 1, is Text 2 true, false or neutral (you can't tell if it's true or false)? Reply in one word.

Text 1: "text1"

Text 2: "text2"

(Variation 1)

You have two texts, and your task is to determine the truthfulness of Text 2 based on Text 1. Provide a one-word response indicating whether Text 2 is true, false, or neutral (indeterminable). Here are the texts:

Text 1: "text1"

Text 2: "text2"

(Variation 2)

Assess the veracity of Text 2 based on Text 1: Is Text 2 true, false, or indeterminable? Provide a one-word response.

Text 1: "text1"

Text 2: "text2"

The comparison between the three prompts is shown in Table 5.

E Chain-of-thought prompting

E.1 Chain-of-thought experiments

For the chain-of-thought experiments, we added the phrase "Let's think step by step" (Kojima et al., 2023) to the original prompt as follows:

You are given a pair of texts. Say about this pair: given Text 1, is Text 2 true, false or neutral (you can't tell if it's true or false)?

Text 1: "text1"

Text 2: "text2"

Let's think step by step.

997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039

Instructions

Based on Text 1, determine if Text 2 is true, false or neutral (you can't tell if it's true or false).

Work Process

- Read all the examples carefully.
- For each of the 5 text pairs below:
 - Read Text 2 carefully.
 - Read Text 2 carefully.
 - Based on Text 1, determine if Text 2 is true, false or neutral (as shown in the examples)
 - Select the corresponding option

Examples (click to collapse / expand)

Example 1
Example 2
Example 3

Text 1: Jane is a cat.
Text 2: Jane is five years old.
Based on Text 1, is Text 2 True, False or Neutral (you cannot tell if it's true or false)?

True False Neutral

If Jane is a cat, she may or may not be five years old. So based on Text 1, we **cannot determine if Text 2 is True or False**. So we select **Neutral**.

Question 1:
Text 1: She took a nutritious and expensive protein shake.
Text 2: She took a protein shake.
Based on Text 1, is Text 2 True, False or Neutral (you cannot tell if it's true or false)?

True False Neutral

Figure 2: Instructions for human annotation. Note that in order to not implicitly train annotators towards the linguistic inferences we consider in this work, we provided examples that demonstrate the meaning of the neutral/entailing/contradiction labels, but on other inference types, not covered in this work.

		human			text-davinci-003			gpt-3.5-turbo-0301			gpt-3.5-turbo-0613			gpt-4-0314			gpt-4-0613			LLaMA 2 70b		
		e	n	c	e	n	c	e	n	c	e	n	c	e	n	c	e	n	c	e	n	c
SA	P	100	0	0	39	56	5	53	34	13	9	53	38	84	16	0	<u>100</u>	0	0	31	61	8
	MP	89	11	0	25	46	29	43	51	6	21	35	44	39	61	0	70	30	0	36	59	5
	MN	5	94	1	22	28	50	37	42	21	12	37	51	6	88	6	33	56	11	46	35	19
	UA	0	100	0	77.56	4.67	17.78	79.89	9.11	11.00	34	19	47	30	70	0	61	38	1	89.89	9	1.11
PT	P	100	0	0	65.35	27.83	6.82	72.30	23.35	4.35	6	42	52	90	10	0	100	0	0	32.00	58.13	9.87
	MP	97	3	0	38.04	25.13	36.83	55.65	37.22	7.13	17	40	43	39	61	0	71	29	0	36.87	44.87	18.26
	MN	5	95	0	39.78	19.04	41.18	52.83	29.39	17.78	21	24	55	9	87	4	36	53	11	32.43	38.52	29.05
	UA	1	99	0	87.78	2.65	9.57	86.91	7.52	5.57	28	25	47	37	63	0	70	30	0	58.39	29.35	12.26
NF	P	0	100	0	51.00	41.00	8.00	56.65	40.04	3.31	4	49	47	55	45	0	97	3	0	29.74	64.74	5.52
	MP	1	99	1	28.83	27.26	43.91	60.17	30.17	9.66	13	37	50	16	84	0	44	56	0	34.35	52.35	13.30
	MN	0	100	0	30.30	17.43	52.26	52.91	22.35	24.74	15	23	62	4	94	2	23	69	8	29.30	41.96	28.74

Table 3: Experiment results across all models. Background indicates the expected label (green - "entailment", blue - "neutral"). The green and red font colors indicate increase or decrease compared to standalone premises. The numbers in **bold** also indicate *accuracy* (e.g. for uncertainty adverbs accuracy coincides with the percentage of neutrals). The **underlined bold** numbers indicate *the best accuracy for a specific inference type* across all models (human baseline excluded). **SA** - standalone premises; **PT** - premises under presupposition triggers; **NF** - premises under non-factives; **P** - pronouns (grammatically-specified-entailments); **MP** - monotonicity entailment (positive); **MN** - monotonicity entailment(negative), **UA** - uncertainty adverbs; **e** - entailment (%); **n** - neutral (%); **c** - contradiction (%).

1040 As can be seen, we 1) removed the requirement
 1041 to return a one-word answer; 2) added the words
 1042 "Let's think step by step" at the end.

1043 After the model outputs a chain of thought, an
 1044 additional step is needed to obtain a final one-word
 1045 answer. For this *answer extraction* step we use an
 1046 additional prompt:

1047 *Therefore, the one-word answer (True,*

False or Neutral) is

1048
 1049 For the CoT experiments with standalone
 1050 premises we use the same 100-example test sets
 1051 as for the original experiments (see Section 3 for
 1052 details). For experiments with embeddings we sam-
 1053 ple 100 sentence pairs out of each 2300-example
 1054 test set.

The results of the original experiments (see Sec-

	Inference type	human	LLaMA 2 70b	text-davinci-003	gpt-3.5-turbo-0301	gpt-3.5-turbo-0613	gpt-4-0314	gpt-4-0613
Standalone	pronouns	100.00	31.00	39.00	53.00	9.00	84.00	100.00
	monotonicity positive	89.00	36.00	25.00	43.00	21.00	39.00	70.00
	monotonicity negative	94.00	35.00	28.00	42.00	37.00	88.00	56.00
	uncertainty adverbs	100.00	9.00	4.67	9.11	19.00	70.00	38.00
Under presupposition triggers	pronouns	100.00	32.00	65.35	72.30	6.00	90.00	100.00
	monotonicity positive	97.00	36.87	38.04	55.65	17.00	39.00	71.00
	monotonicity negative	95.00	38.52	19.04	29.39	24.00	87.00	53.00
	uncertainty adverbs	99.00	29.35	2.65	7.52	25.00	63.00	30.00
Under non-factives	pronouns	100.00	64.74	41.00	40.04	49.00	45.00	3.00
	monotonicity positive	99.00	52.35	27.26	30.17	37.00	84.00	56.00
	monotonicity negative	100.00	41.96	17.43	22.35	23.00	94.00	69.00
	overall accuracy	97.55	36.98	27.95	36.78	24.27	71.18	58.73

Table 4: Human accuracy and accuracy (%) across LLMs compared. LLaMA 2, GPT-3.5 (text-davinci-003) and gpt-3.5-turbo-0301 were tested on full data. Human performance, gpt-3.5-turbo-0613 and both GPT-4 versions were tested on 100 examples per inference type.

		original prompt		paraphrased prompt 1		paraphrased prompt 2	
		accuracy (%)	entailment (%)	accuracy (%)	entailment (%)	accuracy (%)	entailment (%)
Standalone	pronouns	53.00	53.00	67.00	67.00	62.00	62.00
	monotonicity positives	43.00	43.00	68.00	68.00	65.00	65.00
	monotonicity negatives	42.00	37.00	8.00	53.00	37.00	42.00
	uncertainty adverbs	9.1	79.9	5.67	80.33	8.44	78.67
Under presupposition triggers	pronouns	72.30	72.30	86.70	86.70	76.61	76.61
	monotonicity positives	55.65	55.65	78.30	78.30	70.39	70.39
	monotonicity negatives	29.39	52.83	6.17	59.35	29.70	50.70
	uncertainty adverbs (sample)	7.52	86.91	1.39	88.65	9.61	82.65
Under non-factives	pronouns	40.04	56.65	20.09	72.87	40.52	56.65
	monotonicity positives	30.17	60.17	9.39	72.83	29.57	63.61
	monotonicity negatives	22.35	52.91	6.00	48.87	40.00	38.39

Table 5: The experiment results for the original prompt and its two paraphrases suggested by ChatGPT itself. The background color indicates the expected label: green - “ENTAILMENT”, blue - “NEUTRAL”. The pink cells indicate the results that do not fit the pattern exhibited by the original prompt (see Appendix B): predicting more entailment under all embedding types. The bold figures indicate the highest accuracy for a specific inference type across all 3 prompts.

tion 3) and the CoT experiments are compared in Table 6.

E.2 Manual analysis of ChatGPT’s reasoning chains

This expands on the results presented in Section 4, “Analysis of CoT Results“. We sampled 55 examples of reasoning chains produced by the ChatGPT at the time of the CoT experiments (5 from each test set) and analyzed them manually.

Some chains of thought (about 14.5%) are clear, correct, lead to a correct answer and demonstrate understanding of the task and the input premise-hypothesis pair. For example:

Premise: *It’s worth noting that all philosophers are mortal.*

Hypothesis: *All hegelians are mortal.*

Chain of thought: *Step 1: Examine Text 1. It states that “All philosophers are mortal.” This means that every philosopher who ever existed or will exist must eventually die. This statement is categorical and absolute.*

Step 2: Compare Text 2 to Text 1. Text 2 states that “All hegelians are mortal.” Hegelians are a specific subset of philosophers who follow the philosophy of Georg Wilhelm Friedrich Hegel. Since Text 1 states that all philosophers are mortal, and Hegelians fall under the category of philosophers, it follows that Text 2 is true.

Answer: True.

Prediction: *Entailment.*

In some other cases (about 9.1%) the prediction is correct, but the reasoning chain is too vague, the logical transitions between the steps are unclear, so that it’s impossible to tell if the model reasoned correctly or not. For example:

Premise: *Sam violated the school’s jewelry policy.*

Hypothesis: *Sam violated something.*

Chain of thought: *1. Start by analyzing Text 1. It states that “Sam violated the school’s jewelry policy.” This implies*

		original prompt		chain-of-thought prompt		
		accuracy (%)	entailment (%)	accuracy (%)	entailment (%)	neutral (%)
Standalone	pronouns	53.00	53.00	7.00	7.00	90.00
	monotonicity positives	43.00	43.00	44.00	44.00	53.00
	monotonicity negatives	42.00	37.00	53.00	39.00	53.00
	uncertainty adverbs	9.1	79.9	46.56	46.44	46.56
Under presupposition triggers	pronouns	72.30	72.30	8.00	8.00	91.00
	monotonicity positives	55.65	55.65	26.00	26.00	70.00
	monotonicity negatives	29.39	52.83	56.00	31.00	56.00
	uncertainty adverbs (sample)	7.52	86.91	52.00	43.00	52.00
Under non-factives	pronouns	40.04	56.65	99.00	1.00	99.00
	monotonicity positives	30.17	60.17	78.00	21.00	78.00
	monotonicity negatives	22.35	52.91	58.00	30.00	58.00

Table 6: The experiment results for the original prompt and the CoT prompt. The background color indicates the expected labels: green – “ENTAILMENT”; blue – “NEUTRAL”. The bold figures indicate which prompt scored higher on a specific inference type.

1100	<i>that there is a school policy regarding</i>	<i>statement seems to suggest that all aliens</i>	1134
1101	<i>wearing jewelry and Sam broke this rule.</i>	<i>used coins, but it does not specify which</i>	1135
1102	<i>2. Next, move on to Text 2. It states that</i>	<i>type of coins they used.</i>	1136
1103	<i>“Sam violated something.” It is not speci-</i>	<i>Step 3: Since Text 2 does not specify</i>	1137
1104	<i>fied what Sam violated, but it is implied</i>	<i>which types of coins alien used, we can-</i>	1138
1105	<i>that Sam did break a rule or a policy of</i>	<i>not say for sure whether it is true or</i>	1139
1106	<i>some sort.</i>	<i>false. However, we can say that the state-</i>	1140
1107	<i>3. Based on Text 1 and Text 2, it is safe to</i>	<i>ment in Text 2 is neutral since it does</i>	1141
1108	<i>say that Text 2 is true. The fact that Text 1</i>	<i>not contradict Text 1, but also does not</i>	1142
1109	<i>explicitly mentions a specific policy that</i>	<i>provide us with any new information that</i>	1143
1110	<i>Sam violated supports the idea that a vi-</i>	<i>supports Text 1.</i>	1144
1111	<i>olation did occur. Text 2 confirms this by</i>	Prediction: Neutral.	1145
1112	<i>stating that Sam violated something, thus</i>		
1113	<i>aligning with the information provided</i>	The correct label here is indeed “neutral”, how-	1146
1114	<i>in Text 1.</i>	ever the reason cited in the chain of thought (“Since	1147
1115	<i>Therefore, the answer is: Text 2 is true.</i>	<i>Text 2 does not specify which types of coins alien</i>	1148
1116	Prediction: Entailment.	<i>used, we cannot say for sure whether it is true or</i>	1149
		<i>false”) is wrong. In reality, the relation is neu-</i>	1150
1117	Yet in other cases (about 23.6%) the model is	<i>tral because of the non-factive expression “It was</i>	1151
1118	“right for the wrong reasons” arriving at a correct	<i>reported”.</i>	1152
1119	prediction through a wrong reasoning chain. For	Remarkably, in one case, the model does not	1153
1120	example:	arrive at the gold label, but seems to be - at least	1154
		pragmatically - correct in its reasoning, and the	1155
1121	Premise: <i>It was reported that every alien</i>	CoT clearly makes sense:	1156
1122	<i>used some quarters.</i>		
1123	Hypothesis: <i>Every alien used coins.</i>	Premise: <i>it’s worth noting that I don’t</i>	1157
1124	Chain of thought: <i>Step 1: Let’s ana-</i>	<i>eat chicken skin.</i>	1158
1125	<i>lyze Text 1 to extract the key informa-</i>	Hypothesis: <i>I don’t eat chicken skin at</i>	1159
1126	<i>tion. The statement says “Every alien</i>	<i>any cost.</i>	1160
1127	<i>used some quarters.” This means that</i>	Chain of thought: <i>Step 1: Analyze Text</i>	1161
1128	<i>all aliens used quarters, but it is possi-</i>	<i>1 The first text states that the person</i>	1162
1129	<i>ble they also used other types of coins or</i>	<i>speaking does not eat chicken skin. This</i>	1163
1130	<i>currency.</i>	<i>seems pretty clear and straightforward.</i>	1164
1131	<i>Step 2: Now, let’s examine Text 2 in</i>	<i>Step 2: Analyze Text 2 The second</i>	1165
1132	<i>light of Text 1. The statement in Text</i>	<i>text states that the person does not eat</i>	1166
1133	<i>2 says “Every alien used coins.” This</i>	<i>chicken skin at any cost. This seems to</i>	1167
		<i>be an exaggerated statement compared</i>	1168

1169	<i>to the first text. The phrase “at any cost”</i>	Here the model not only demonstrates misunder-	1215
1170	<i>implies that the person is willing to go</i>	standing of the task trying to determine the veracity	1216
1171	<i>to great lengths or make great sacrifices</i>	of Text 1 based on Text 2, while clearly instructed	1217
1172	<i>to avoid eating chicken skin. This level</i>	to do the opposite, but also makes a moral judge-	1218
1173	<i>of emphasis is not present in Text 1, so</i>	ment (see step 4) which is totally out of place in	1219
1174	<i>it’s possible that Text 2 is an exaggerated</i>	this task.	1220
1175	<i>version of the truth.</i>	Below we list some more observations regarding	1221
1176	<i>Step 3: Decide if Text 2 is true, false,</i>	the analyzed reasoning chains:	1222
1177	<i>or neutral Based on the information we</i>	1. The embedding context (presupposition trig-	1223
1178	<i>have, we can’t definitively say whether</i>	gers or non-factives) are sometimes men-	1224
1179	<i>Text 2 is true or false. It’s possible that</i>	tioned in the chain of thought, but are never	1225
1180	<i>the person speaking really does avoid</i>	used as a basis for the final decision. (One	1226
1181	<i>eating chicken skin at all costs, but it’s</i>	consequence of this is that for the test sets	1227
1182	<i>also possible that they are exaggerating</i>	with non-factives all the correct answers re-	1228
1183	<i>and simply avoid it when possible. There-</i>	sult from wrong reasoning chains.)	1229
1184	<i>fore, the answer is neutral.</i>	2. For premises with uncertainty adverbs, in 8	1230
1185	Model’s prediction: <i>Neutral.</i>	out of the 10 analyzed cases the adverb is	1231
1186	Gold label: <i>Entailment.</i>	mentioned and its meaning explained, but only	1232
1187	Finally, in the remaining cases both the chain of	in 3 out of the 10 cases the adverb serves as a	1233
1188	thought and the resulting prediction are wrong. For	basis for the final answer.	1234
1189	example:	3. The model constantly misinterprets indefi-	1235
1190	Premise: <i>The school’s deputy headmas-</i>	nite pronouns as referring to a specific entity.	1236
1191	<i>ter, Howard Jones, told her to remove the</i>	Hence the incorrect “neutral” labels for most	1237
1192	<i>crucifix around her neck.</i>	cases of grammatically-specified entailment.	1238
1193	Hypothesis: <i>The school’s deputy head-</i>	For example ChatGPT decides that “ <i>Mary</i>	1239
1194	<i>master, Howard Jones, told her to do</i>	<i>lent him money</i> ” does not entail “ <i>Someone</i>	1240
1195	<i>something.</i>	<i>lent him money</i> ” because “someone” in Text	1241
1196	Chain of thought: <i>1. Text 1 states that</i>	2 “ <i>could be referring to someone other than</i>	1242
1197	<i>Howard Jones, the deputy headmaster of</i>	<i>Mary</i> ”. (The correct answer is, of course,	1243
1198	<i>the school, asked a student to remove her</i>	“entailment” because “someone” is a generic	1244
1199	<i>crucifix.</i>	term encompassing <i>any</i> individual including	1245
1200	<i>2. Text 2 only confirms that Howard</i>	Mary.)	1246
1201	<i>Jones did tell the student to do something,</i>	4. The model often gets confused about the	1247
1202	<i>but it doesn’t specify what that something</i>	monotonicity entailment directions (upward	1248
1203	<i>was.</i>	vs. downward), stating, for example, that “ <i>No</i>	1249
1204	<i>3. Therefore, Text 2 cannot be used to</i>	<i>alien ate pork</i> ” entails “ <i>No alien ate meat</i> ”	1250
1205	<i>determine whether or not Text 1 is true</i>	since “ <i>pork is a type of meat</i> ”, but “ <i>Every</i>	1251
1206	<i>or false.</i>	<i>alien used some quarters</i> ” does not entail	1252
1207	<i>4. However, it is important to note that</i>	“ <i>Every alien used some coins</i> ” because Text	1253
1208	<i>if Text 1 is true, and Howard Jones did</i>	2 “ <i>does not specify which type of coins they</i>	1254
1209	<i>ask a student to remove their crucifix,</i>	<i>used</i> ”.	1255
1210	<i>this would be a violation of the student’s</i>	5. More generally, the model usually predicts en-	1256
1211	<i>religious freedom and may constitute dis-</i>	tailment when Text 2 contains a more specific	1257
1212	<i>crimination.</i>	mention than Text 1 (which is, in fact, only	1258
1213	Model’s prediction: <i>Neutral.</i>	correct for cases of downward entailment),	1259
1214	Correct prediction: <i>Entailment.</i>	and vice versa.	1260
		6. The reasoning chains are mostly vague, exces-	1261
		sively wordy, with unclear logical relations	1262

correct CoT/correct label	23.6%
wrong CoT/correct label	23.6%
wrong CoT/wrong label	50.9%
correct CoT /wrong label	1.82%
CoT coherent and clear	16.4%
underlying LP mentioned in CoT	49.1%
correct understanding of the underlying LP reflected in CoT	23.6%
underlying LP explicitly used in prediction	14.5%
CoT demonstrates correct understanding of the task	81.8%
CoT reflects correct understanding of the input sentences	80.0%

Table 7: Manual CoT analysis results. LP stands for "linguistic phenomena". Some numbers are approximate, since not all the cases are clear-cut, and some reasoning chains are unclear and difficult to analyze.

1263 between steps, which makes them hard to un-
1264 derstand and analyze, and often contain obvi-
1265 ous logical errors (e.g. *“Text 2 is likely true,*
1266 *as it directly contradicts the assumption made*
1267 *in Text 1”*).

1268 7. The CoT can sometimes misrepresent the con-
1269 tents of the input sentences. For example the
1270 model claims that the text *“I love something*
1271 *outside the city”* does not mention “love”.

1272 8. Different chains of thought exhibit contra-
1273 dictory logics. For example, one CoT says
1274 *“There is no contradiction between the two*
1275 *texts... Therefore, Text 2 can be determined as*
1276 *true”*, while another reasoning chain states:
1277 *“Text 2 does not contradict Text 1, so it is*
1278 *neutral.”*

1279 Quantitatively, the results of this analysis are repre-
1280 sented in Table 7.

1281 The analysis shows that zero-shot CoT prompt-
1282 ing fails to improve ChatGPT’s performance on
1283 the task because of various flaws in the generated
1284 reasoning chains.