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. The models exhibit moderate to low performance on these evaluation sets in all settings. Subsequent experiments show that embedding the premise under presupposition triggers or non-factives, which should exhibit opposite linguistic behavior, causes ChatGPT to predict entailment more frequently in the zero-shot and less frequently in the chain-of-thought setup, and in each case regardless of the correct label. Similar experiments with LLaMA 2 exhibit different yet equally flawed tendencies. Overall these results suggest that, despite LLMs' celebrated language understanding capacity, they have blindspots with respect to certain types of entailments, and that 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 on 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. Therefore LLMs' errors on such tasks, while making them unreliable for certain applications, do not necessarily mean that they haven't reached *human-level* performance on reasoning and understanding.

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 but nonetheless pose a challenge to LLMs. Using NLI we reveal some of the models' blind spots that indicate that they are far from a genuine *human-level* understanding. Furthermore, we show that some information-packaging structures may act as "blinds" that hinder the semantics of embedded premises, again in contrast to human-like behavior.

We focus on inference types that are solely based on common linguistic phenomena and "trival" 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 uncertainty* (*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

Condition	Section	Standalone Expected	Standalone Model	Embedded within a Pressuposition Trigger Expected	Embedded within a Pressuposition Trigger Model	Embedded within a Non-factive Clause Expected	Embedded within a Non-factive Clause 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	---	---

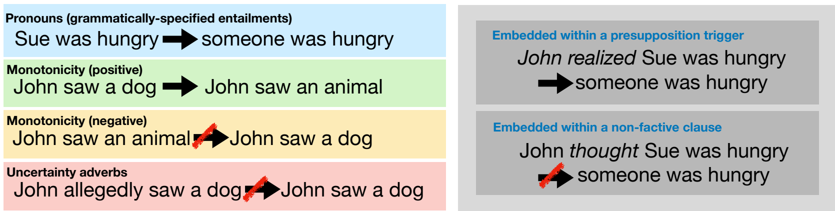

Figure 1: High-level summary of the experiments and results.

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. In particular, 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-factive verbs (e.g. *I feel that [...], They thought that [...], He imagined that [...]*) may *cancel* it. We experiment with both types of contexts and show that ChatGPT[1] has a hard time distinguishing 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 consistent across different prompts and models (such as GPT-3.5 and LLaMA 2), and in both regular and chain-of-thought prompting showing that these inference failures are robust, and merit further investigation. Figure 1 summarizes our main results.

[1] https://openai.com/chatgpt

2 Linguistic background

First, we briefly explain the linguistic phenomena that we use in the experiments described below.

Gramatically-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.

You've done something.

Someone's eaten all my apples.

Monotonicity entailments hold when less specific predicates are substituted with more specific ones, or vice versa. They can be of two types:

- Upward: more specific predicates can be substituted with less specific ones.

  Jack is a dog. $\models$ *Jack is an animal.*

- Downward: less specific predicates can be substituted with more specific ones.

  All animals need water. $\models$ *All dogs need water.*

Liu et al. (2023) show that ChatGPT's performance on monotonicity entailment datasets (HELP (Yanaka et al., 2019b) and MED (Yanaka et al., 2019a)) improves on previous models but is still far from perfect (42.13% and 55.02% respectively).

Evidential Adverbs "express degrees of certitude with respect to the speaker's subjective perception of the truth of a proposition" (Haumann, 2007). We run experiments that test LLMs' ability to understand evidential adverbs expressing *uncertainty* (*allegedly*, *purportedly*, *supposedly* etc.). Introducing such adverbs into a clause cancels the entailment of the remaining part of the clause. For example, *Mike allegedly worked all night* does not entail *Mike indeed worked all night*. The relation between the two statements is 'neutral'.

Presuppositions and Presupposition Triggers *Presupposition* (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):

a. *Jane returned to New York.*

$\models$ **b**. *Jane has been to New York before.*

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 *inform* the hearer that Jane has been to New York before: she *assumes* it; and if the hearer doesn't know it, she *accommodates* it upon hearing the utterance (Fintel, 2008).

Presuppositions are normally evoked by constructions or lexical items, called *presupposition triggers* (Karttunen, 2016). In sentence **a** above, the presupposition is triggered by the verb *returned*, from the class of *iterative verbs* which presuppose that the action has happened before. Other iterative verbs are *relearn, reread, reapply* etc. Presupposition triggers used in this work are factives, temporal, and other adverbial clauses and embedded *wh*-questions, (We detail them in Appendix A).

Non-factive Verbs and Expressions (Kiparsky and Kiparsky, 1970), such as *believe, claim, feel, hope, suspect, think*, **do not entail either truth or falsity of their complements**. For example, given:

a. *Jane thinks that Bill bought bread.*

b. *Bill bought bread.*

c. *Bill didn't buy bread.*

Sentence **a** doesn't entail either **b** or **c**. The relation between **a** and **b** is neutral, and so is the relation between **a** and **c**.

Presupposition Triggers, Non-Factives and NLI It is important to note that embedding a premise under a presupposition trigger doesn't 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:

a. *A balloon hit a light post.*

$\models$ **b**. *Something hit a light post*

Premise **a** above entails hypothesis **b**. If we embed premise **b** under a presupposition trigger as in **a'**:

a'. *<u>She realized that</u> a balloon hit a light post,*

$\models$ **b**. *Something hit a light post*

the relation does not change: the new premise **a'** still entails **b**. However, when embedding premise **a** under a non-factive verb:

a". *<u>I suspect</u> a balloon hit a light post*

$\not\models$ **b**. *Something hit a light post*

the relation becomes neutral: without additional context the new premise **a"** doesn't entail **b**.

3 Main Experiments

Setup. In this section our experiments use the gpt-3.5-turbo-0301 ChatGPT version (while Section 4 tests more models). We access ChatGPT through OpenAI's API[2] with the default settings, and use 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). This prompt yields a 71% accuracy[3] on a 300-instance SNLI sample, consistent with previous SNLI literature (Qin et al., 2023; Wang et al., 2023; Jang and Lukasiewicz, 2023).

Probing ChatGPT for Phenomena Explaining. Before each experiment we probe ChatGPT's ability to explain the relevant linguistic phenomenon. In humans, being able to explain a concept typically implies being able to apply it. However in LLMs, explaining does not necessarily imply knowing. While not our study's main focus, this probing exemplifies a contrast between humans' and LLMs' language comprehension attested abilities.

[2] https://openai.com/product

[3] All the results reported in this paper are from a single run.

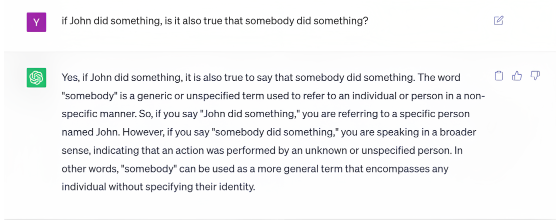

Figure 2: ChatGPT correctly explains grammatically specified entailment when prompted for it directly.

3.1 Grammatically-Specified Entailments

3.1.1 Probing ChatGPT for the phenomenon explanation

When prompted explicitly with the concept, ChatGPT correctly explains the meaning and usage of indefinite pronouns (see Figure 2).

3.1.2 Testing ChatGPT in an NLI setting

Data: we manually curated a dataset of 100 pairs with grammatically specified entailments by selecting naturally occurring sentences and, in each one, replacing a noun-phrase with an indefinite pronoun:

> **Premise:** *Crown Princess Mary of Denmark has given birth to a healthy baby boy.*
>
> **Hypothesis**: *Someone has given birth to a healthy baby boy.*

This is a seemingly very easy dataset, in which all the gold labels are “ENTAILMENT”, trivial for any human to solve.

Results: The model’s accuracy on this (seemingly simple) dataset is 53%, failing in almost half (47%) of the cases.

3.1.3 Embedding the Premises under Presupposition Triggers

Data: Next we modify this dataset by embedding the original premise (but not the hypotheses) under a presupposition trigger. For instance:

> **Original premise:** *Crown Princess Mary of Denmark has given birth to a healthy baby boy.*
>
> **Modified premise**: *We are happy that Crown Princess Mary of Denmark has given birth to a healthy baby boy.*

The hypotheses remain unmodified:

> *Someone has given birth to a healthy baby boy.*

Since presupposition triggers don’t cancel entailment, the gold label in all the cases also remains unchanged: “ENTAILMENT”.

We used 23 presupposition trigger types: 21 factives, such as *know, realize, be glad*; embedded wh-questions, such as *This explains why he came*; and adverbial clauses, e.g. *After he came, I cooked dinner*. Embedding the 100 original premises under 23 triggers resulted in 2300 examples.

Results: In the same setting as before, on this data we achieve a significant increase in accuracy: 72.30% for embedded premises vs. 53% for original, "non-embedded", ones. The model predicted entailment more frequently than it originally did.

3.1.4 Embedding the Premises under Non-Factive Verbs

Data: We modify the dataset, embedding the premises under non-factive predicates as follows:

> **Original premise:** *Crown Princess Mary of Denmark has given birth to a healthy baby boy.*
>
> **Modified premise**: *I hope Crown Princess Mary of Denmark has given birth to a healthy baby boy.*

The hypotheses remain unmodified:

> *Someone has given birth to a healthy baby boy.*

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

We embedded each of the 100 original premises under 23 different non-factives, such as *feel, hope, believe* etc., obtaining 2300 examples.

Results: We run the same NLI experiment. Surprisingly, the model still predicts “ENTAILMENT” in 56.65% of the cases - more often than on the original dataset with non-embedded premises (53%).

3.1.5 Bottomline

Our results suggest that ChatGPT struggles with grammatically-specified entailments. It tends to predict “ENTAILMENT” more often when the premise is embedded under a presupposition trigger, which creates a seeming accuracy increase. But embedding the premise under non-factives also causes the LLM to predict “ENTAILMENT” more, even though in this case it is wrong. So, the mere syntactic embedding affects NLI consistently, regardless of the semantics of the embedding clause.

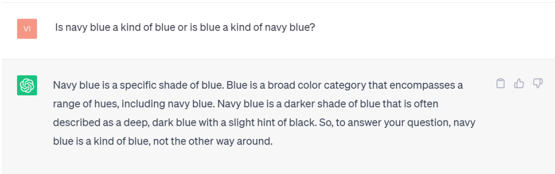

Figure 3: ChatGPT correctly explains set-membership relations.

3.2 Monotonicity Entailment

3.2.1 Probing ChatGPT for the phenomenon explanation

ChatGPT correctly explains the set-membership relations (which are typically the basis of monotonicity entailments) when prompted for it directly (see Figure 3). Since the example in the figure only covers a specific set-membership pair and thus might not be representative enough, we run a similar test with more pairs extracted from our monotonicity dataset and get similarly accurate explanations in all the cases (see Appendix C for more details).

3.2.2 Testing ChatGPT in an NLI setting

We sample two subsets of the Monotonicity Entailment Dataset (MED) (Yanaka et al., 2019a):

- 100 positive examples (the gold label is "ENTAILMENT").
- 100 negative examples (the gold label is "NEUTRAL").

Example from the positive subset:

> **Premise:** *She planted blue and purple pansies in the flower bed.*
>
> **Hypothesis**: *She planted pansies in the flower bed.*
>
> **Label:** *Entailment*

Example from the negative subset:

> **Premise:** *Susan made a dress for Jill.*
>
> **Hypothesis**: *Susan made a long dress for Jill.*
>
> **Label:** *Neutral*

We first describe our experiments with the positive subset, and then with the negative one.

3.2.3 Positives: Standalone Premises

Our usual NLI experiment on the positive part of the monotonicity data, yields an accuracy of 43%.

Embedding the premises under presupposition triggers As above, we embed each premise under 23 types of triggers, resulting in 2300 examples. We modify the premises accordingly:

> **Original premise:** *She planted blue and purple pansies in the flower bed.*
>
> **New premise**: *After she planted blue and purple pansies in the flower bed, she started planting other flowers.*

The hypothesis remains unchanged:

> *She planted pansies in the flower bed.*

Since presupposition triggers don't change the relations between the premise and the hypothesis, the gold labels remain unmodified: "ENTAILMENT".

Results: After this modification the accuracy for positive examples improves significantly: 55.65% — again, the model predicts entailment more often.

Embedding the premises under non-factive verbs We use the same 23 non-factives as for the grammatically-specified entailment experiments above. We apply each non-factive to each original premise, obtaining 2300 examples. For example:

> **Original premise:** *She planted blue and purple pansies in the flower bed.*
>
> **New premise**: *I think she planted blue and purple pansies in the flower bed.*

The hypothesis remains unmodified:

> *She planted pansies in the flower bed.*

Since non-factives cancel entailment, the gold label now becomes "NEUTRAL" for all the pairs.

Results: The model predicts entailment even more often than for premises under presupposition triggers, 60.17% of the time, even though in this case it is wrong.

Bottomline: As in the case of grammatically-specified entailments, the mere embedding of the premise, whether under presupposition triggers or non-factives, causes the LLM to predict entailment much more, regardless of the correct inference.

3.2.4 Negatives: Standalone Premises

In the NLI experiment on the negative part of the monotonicity data, the model predicts entailment in 37% of the cases (which is wrong). It predicts the correct label "NEUTRAL" for 42% of the pairs and "CONTRADICTION" for the remaining 21%.

Embedding the premises under presupposition triggers We embed the premises under presupposition triggers, as before:

> **Original premise:** *Susan made a dress for Jill.*
>
> **New premise**: *They are aware that Susan made a dress for Jill.*

The hypothesis remains unchanged:

> *Susan made a long dress for Jill.*

Since presupposition triggers don't change the relations between the premise and hypothesis, the gold labels remain unmodified: “NEUTRAL”.

Applying 23 trigger types to 100 premises results in 2300 new examples.

Results: After this modification, the model (incorrectly) predicts "ENTAILMENT" even more: in 52.83% of the cases. The correct label, "NEUTRAL", is predicted for 29.39% of the cases, and the remaining 17.78% are "CONTRADICTION".

Embedding the premises under non-factive verbs We modify the premises accordingly:

> **Original premise:** *Susan made a dress for Jill.*
>
> **New premise**: *They believe Susan made a dress for Jill.*

The hypothesis remains unmodified:

> *Susan made a long dress for Jill*

The gold label remains “NEUTRAL” for all pairs.

To each premise we apply the same set of 23 non-factives as above, obtaining 2300 pairs.

Results: The model, again, predicts "ENTAILMENT" even slightly more often than for premises under presuppositon triggers: in 52.91% of the cases. The correct label "NEUTRAL" appears in 22.35% of the cases, and the remaining 24.74% of the labels are "CONTRADICTION".

Bottomline Thus, both on the positive and negative monotonicity entailment test sets the model demonstrates the same pattern we observed before: embedding the premise inside a clause, under presupposition triggers or non-factives, causes the model to predict entailment more often, regardless of the embedding clause's semantic content.

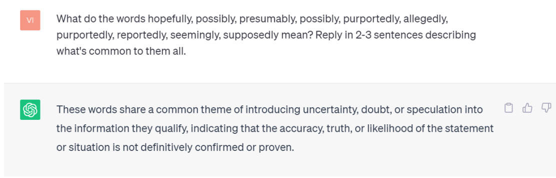

Figure 4: ChatGPT correctly explains the meaning of evidential adverbs when prompted for it directly.

3.3 Adverbs of Uncertainty

3.3.1 Probing ChatGPT for the phenomenon explanation

ChatGPT correctly explains the meaning and usage of evidential adverbs of uncertainty when prompted for it directly (see Figure 4).

3.3.2 Testing ChatGPT in an NLI setting

We manually create a dataset of 100 sentence pairs where the only difference between the premise and the hypothesis is that the premise contains an adverb of uncertainty, while the hypothesis omits it:

> **Premise:** *These persons were allegedly inhabiting the home.*
>
> **Hypothesis**: *These persons were inhabiting the home.*

We apply 9 uncertainty 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 the pairs is “NEUTRAL”.

Results: The model predicted "ENTAILMENT" in 79.9% of the cases, which is wrong. It predicts "NEUTRAL" in 9.1% of the cases (which is correct) and "CONTRADICTION" in 11% of the cases. The results are mostly consistent across all the individual adverbs (see Appendix D for details).

3.3.3 Embedding the premises under presupposition triggers

We randomly sample 100 examples from our dataset of 900 sentence pairs, and apply each of the 23 triggers to each sampled example:

> **Original premise:** *These persons were allegedly inhabiting the home.*
>
> **New premise**: *The owner was aware that these persons were allegedly inhabiting the home.*

The hypothesis remains the same:

These persons were inhabiting the home.

Since presupposition triggers don't affect the original relation, the gold label also remains the same: “NEUTRAL”.

Results: The model predicts "ENTAILMENT" even more often: in 86.91% of the cases (which is wrong). It predicts "NEUTRAL" (the correct label) only for 7.52% of the pairs, and "CONTRADICTION" - for the remaining 5.57%.

3.3.4 Embedding the premises under non-factive verbs

We omit this experiment for adverbs of uncertainty for semantic reasons: including both an uncertainty adverb and a non-factive into the premise (*I guess he allegedly worked all night.*) results in double expression of uncertainty, which creates a tautology.

3.3.5 Bottomline

These experiments exhibit the same pattern we observed earlier: when the premise is embedded under a presupposition trigger, the model tends to predict entailment. As a result, in presupposed content ChatGPT appears to overlook evidential adverbs even more than in unembedded premises.

See Figure 1 for a summary of the experimental results described in this section.

4 Model and Prompt Variations

To assess further the LLMs' performance on these phenomena, we use model and prompt variations.

Prompt variation. We ask ChatGPT to rephrase our prompt template (see Appendix B), obtaining two templates (see Appendix G) that we verify to be semantically equivalent w.r.t. the task. The first prompt shows the same pattern as our original prompt; the second one yields higher accuracy, and even shows a more reasonable trend of predicting entailment *less* often in non-factive contexts (which indeed cancel entailment). But even this, better, prompt is very far from solving the task.

These differences in the behavior of the three prompts again stress ChatGPT's inconsistency: according to ChatGPT itself they all have the same meaning and describe the same task; however, they produce different outputs. See Appendix G for full experiment details, results and their comparison.

Model variation. We repeated the experiments from Section 3 with additional models.

1. GPT-3.5. We run the same experiments over the text-davinci-003 model,[4] with a temperature of 0. The results are overall lower than those of ChatGPT, while exhibiting the same trends. At the same time ChatGPT tends to overpredict entailment more than text-davinci-003. See Appendix E for full GPT-3.5 results and their comparison with ChatGPT results.

2. LLaMA 2. To verify that the problem is not model-specific, we run the same set of experiments using the LLaMA-2 Chat model with 70 billion parameters[5] (Touvron et al., 2023).[6] We first assess its performance on the NLI task using the same SNLI sample as for ChatGPT (see Section 3). With our original prompt LLaMA 2 only achieves a 39% accuracy on this test. Therefore we search for another prompt and find that prompt (1) (see Appendix G) from our prompt variation experiments (see above) yields a 61% accuracy. Using this prompt, the temperature of 0.01 (lowest possible) and top-k=1, we run the set of experiments described in Section 3.

LLaMA 2 also exhibits moderate to low accuracy (mostly far below its own SNLI results) on all the test sets. For unembedded premises, the LLM tends to predict the opposite of the expected label ("entailment" for uncertainty adverbs where "neutral" is expected; "neutral" for grammatically-specified entailments where "entailment" is expected, etc.)

Like for GPT models, embedding premises in larger grammatical contexts affects the predictions, but the pattern is different. For all inference types (grammatically-specified entailments, monotonicity entailments, uncertainty adverbs), under presupposition triggers and, even more, under non-factives the accuracy increases relative to the standalone premises. Looking just at the numbers, it seems that such embeddings in larger grammatical contexts make the model more sensitive to simple linguistic inferences, that is, to some extent more accurate.

However, a closer examination of the results suggests that the increase in accuracy might be due to the wrong reasons. For example, for grammatically-specified and positive monotonicity entailments, the "neutral" predictions prevail with or without the embedding clauses, regardless of the correct label. Hence the results are seemningly better under non-factives, where the more-frequent neutral

[4] https://platform.openai.com/docs/models/gpt-3-5

[5] The largest currently available LLaMA 2 version.

[6] We access LLaMA 2 through the Replicate API: https://replicate.com

label happens to be correct. See Appendix F for full results.

Chain-of-Thought Prompting (CoT). Using the gpt-3.5-turbo-0301 model as in our main experiments (see Section 3), we investigate CoT prompting (Kojima et al., 2023), and find that the results exhibit a different pattern that is equally incorrect. CoT *reverses* ChatGPT's trends for embedded contexts observed in Section 3: the number of entailment predictions *decreases* when the premises are embedded under presupposition triggers or non-factives, while the number of neutral predictions increases - again, regardless of the correct label. CoT prompting improves the accuracy only in a one-sided way: scorring much higher on all the "neutral" test sets, but much 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 undelying linguistic phenomena explicitly, but only in half of those (23.6% of the total) it reflected a correct understanding of the phenomena and only in 14.5% of the cases used them as a basis for the final prediction.

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

5 Conclusions and discussion

Recent studies have highlighted the need for improvement in LLMs' inferencing and reasoning abilities, since they impact the LLM's consistency, reliability, practical applicability and performance on downstream tasks (e.g., Liu et al., 2023; Jang and Lukasiewicz, 2023; Shen et al., 2023; Zheng et al., 2023; Gao et al., 2023; Jin et al., 2023; Plevris et al., 2023; Luo et al., 2023; Lin and Zhang, 2023).

Flawed heuristics and biases in earlier NLI systems have been extensively studied, forming an important research direction (e.g., Poliak et al., 2018; Nie et al., 2018; Glockner et al., 2018; Sanchez et al., 2018; Gururangan et al., 2018; McCoy et al., 2019; Ross and Pavlick, 2019; Zhou and Bansal, 2020; Asael et al., 2022; Gubelmann and Handschuh, 2022). State-of-the-art LLMs' limitations in NLI indicate the presence of fallible tendencies and blind spots in recent foundation models as well, necessitating further investigation. Our study is an attempt to pinpoint some of these weaknesses.

Our experiments focus on simple inference tasks that are typically trivial for humans. We conduct experiments testing LLMs' ability to handle grammatically specified entailments, evidential adverbs of uncertainty, and monotonicity entailment. The findings suggest that LLMs struggle with these types of inferences, exhibiting a gap between their performance and human-level text understanding.

We observe a trend in ChatGPT to predict entailment more often (regardless of the correct label) when premises are embedded in main clauses, containing presupposition triggers or non-factives. Our experiments with text-davinci-003 suggest that this erroneous trend was inherited by ChatGPT from the earlier model despite the upgrading process. Interestingly, in the CoT setting the trend is reversed: clause-embedding contexts cause the LLM to predict entailment *less* - again, regardless of the correct label. Using paraphrases of the same prompt suggested by ChatGPT itself yields other (but similarly inaccurate) results, stressing again the LLM's lack of self-consistency and oversensitivity to prompt phrasing. The LLaMA 2 experiments exhibit low accuracy and other, but also flawed, trends, confirming that the issue is not limited to GPT models, but is likely to affect up-to-date LLMs in general.

Our work shows 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 only for three specific models, we do expect them to hold more generally.

We share the evaluation sets and methodology we created, facilitating study with other models. These results also call for further research to uncover other fallible tendencies and blind spots that might hinder LLMs' ability to make accurate textual inferences, which critically affect their reasoning abilities and their potential for real-world applications.

6 Limitations

This work has some limitations which we list herein.

First, as a consequence of LLMs' sensitivity to prompts, there may exist prompts that can potentially modify the reported results. Moreover, we only tested the LLMs in a regular zero-shot and zero-shot chain-of-thought setting and did not try other approaches like in-context learning or few-shot chain-of-thought prompting. 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, not all possible types of non-factive verbs have been considered. 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 an uncertain, I disbelieve*). It's possible that implicit or explicit negation in the embedding predicates may change the LLMs' behavior. Also, we didn't try other types of clause-embedding predicates (e.g., implicative verbs).

Finally, since ChatGPT undergoes continuous updates, the test results presented here may vary over time. The same is true for LLaMA and future versions of newer models that may become available. That said, our data and methodology for benchmarking these capabilities is model-agnostic and remains intact.

References

Dimion Asael, Zachary Ziegler, and Yonatan Belinkov. 2022. A generative approach for mitigating structural biases in natural language inference. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 186–199, Seattle, Washington. Association for Computational Linguistics.

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.

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

Kai Fintel. 2008. What is presupposition accommodation, again? *Philosophical Perspectives*, 22:137–170.

Jinglong Gao, Xiao Ding, Bing Qin, and Ting Liu. 2023. Is chatgpt a good causal reasoner? a comprehensive evaluation.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

Reto Gubelmann and Siegfried Handschuh. 2022. Uncovering more shallow heuristics: Probing the natural language inference capacities of transformer-based pre-trained language models using syllogistic patterns.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Dagmar Haumann. 2007. *Adverb Licensing and Clause Structure in English*.

Yan Huang. 2011. *14. Types of inference: entailment, presupposition, and implicature*, pages 397–422. De Gruyter Mouton, Berlin, New York.

Myeongjun Jang and Thomas Lukasiewicz. 2023. Consistency analysis of chatgpt.

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

Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona Diab, and Bernhard Schölkopf. 2023. Can large language models infer causation from correlation?

Lauri Karttunen. 2016. Presupposition: What went wrong? *Semantics and Linguistic Theory*, 26:705–731.

Paul Kiparsky and Carol Kiparsky. 1970. *FACT*, pages 143–173. De Gruyter Mouton, Berlin, Boston.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners.

Guo Lin and Yongfeng Zhang. 2023. Sparks of artificial general recommender (agr): Early experiments with 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.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for text summarization.

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

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

Yixin Nie, Yicheng Wang, and Mohit Bansal. 2018. Analyzing compositionality-sensitivity of nli models.

Alicia Parrish, Sebastian Schuster, Alex Warstadt, Omar Agha, Soo-Hwan Lee, Zhuoye Zhao, Samuel R. Bowman, and Tal Linzen. 2021. NOPE: A corpus of naturally-occurring presuppositions in English. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 349–366, Online. Association for Computational Linguistics.

Vagelis Plevris, George Papazafeiropoulos, and Alejandro Jiménez Rios. 2023. Chatbots put to the test in math and logic problems: A preliminary comparison and assessment of chatgpt-3.5, chatgpt-4, and google bard.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver?

Alexis Ross and Ellie Pavlick. 2019. How well do NLI models capture verb veridicality? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2230–2240, Hong Kong, China. Association for Computational Linguistics.

Ivan Sanchez, Jeff Mitchell, and Sebastian Riedel. 2018. Behavior analysis of NLI models: Uncovering the influence of three factors on robustness. 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 1975–1985, New Orleans, Louisiana. Association for Computational Linguistics.

Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2023. In chatgpt we trust? measuring and characterizing the reliability of chatgpt.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Haojun Huang, Wei Ye, Xiubo Geng, Binxin Jiao, Yue Zhang, and Xing Xie. 2023. On the robustness of chatgpt: An adversarial and out-of-distribution perspective.

Deirdre Wilson and Dan Sperber. 1979. *Ordered Entailments: An Alternative to Presuppositional Theories*, volume 11, pages 299–323.

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

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and

Johan Bos. 2019b. HELP: A dataset for identifying shortcomings of neural models in monotonicity reasoning. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 250–255, Minneapolis, Minnesota. Association for Computational Linguistics.

Shen Zheng, Jie Huang, and Kevin Chen-Chuan Chang. 2023. Why does chatgpt fall short in providing truthful answers?

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert.

Xiang Zhou and Mohit Bansal. 2020. Towards robustifying nli models against lexical dataset biases.

Guido Zuccon and Bevan Koopman. 2023. Dr chatgpt, tell me what i want to hear: How prompt knowledge impacts health answer correctness.

A Presupposition triggers

Presupposition triggers are constructions or lexical items that give rise to *presuppositions* (see Section 2). Presupposition trigger types used in this work are factive predicates, such as *know, realize, regret*, which presuppose the truth of their complement clause (*I regret leaving the party* presupposes that *I left the party*), temporal and other adverbial clauses (*Paul worked/didn't work as a driver before he started his company* presupposes that *Paul started his company*), and embedded *wh*-questions (*This explains/doesn't explain why Jane likes him* presupposes that *Jane likes him*).

There are many other types of presupposition triggers, for example, definite descriptions (*The king of France is bold* presupposes that a king of France exists), aspectual/change of state predicates (*Kate quit smoking* presupposes that *Kate smoked before*), implicative predicates like *manage, dare* (*Joe managed to book a table* presupposes that *Joe tried to book a table*), cleft sentences (*It was/wasn't my brother who brought wine* presupposes that *Someone brought wine*), counterfactual conditionals (*If Mike were a politician, I would vote for him* presupposes that *Mike isn't a politician*) and others.

B Main entailment experiments prompt

The prompt below is used throughout the experiments described in Section 3, as well as in the experiments with the text-davinci-003 model (see Section 4 and Appendix E).

> *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"*

The model outputs one of three possible labels: *"true"* (corresponding to *"entailment"*), *"false"* (corresponding to *"contradiction"*) or *"neutral"*.[7]

C Probing ChatGPT for explanations of set-membership relations

To probe ChatGPT for explanations of set-membership relations (which are typically the basis of monotonicity entailments) more systematically, we ran a test on 30 set-membership pairs manually extracted from our monotonicity data, asking the model "Is X a kind of Y or is Y a kind of Y?" In each case the model returned a reasonable answer, correctly explaining the concept, similar to the one shown in Figure 3.

Since monotonicity entailments can also be based on pairs of less specific and more specific verbal phrases (e.g. *drinking coffee - drinking coffee at night*) we also extracted 30 VP pairs of this type from our monotonicity entailment data snd asked the model about each pair: "Which description is more specific: X or Y?" Similarly to set-membership pairs, in each case the model provided an accurate explanation, e.g. ""Drinking coffee at night" is more specific because it specifies the time of day when the activity is taking place."

These results show that the model is able to correctly explain relations that form the basis of monotonicity entailments.

D Adverbs: detailed experiment results

In Table 1 we present the results of the experiments with unembedded premises with uncertainty adverbs using the original prompt (see Appendix B) and the gpt-3.5-turbo model (see Subsection 3.3.2 for details). The table shows the accuracy and the percentage of entailment predictions by individual adverb, as well as the overall results.

adverb	**accuracy (%)**	**entailment(%)**
allegedly	12	74
hopefully	10	81
possibly	10	62
presumably	10	82
probably	5	82
purportedly	7	85
reportedly	15	81
seemingly	4	92
supposedly	9	80
overall	9.1	79.9

Table 1: The results of the experiment with unembedded premises with uncertainty adverbs, using the original prompt and the gpt-3.5-turbo model.

E GPT-3.5 experiment results

The results of our experiments with ChatGPT (see Section 3) and GPT-3.5 (see Section 4) are compared in Table 2.

[7] In 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".

		ChatGPT		GPT-3.5	
		accuracy(%)	entailment (%)	accuracy(%)	entailment (%)
Standalone	pronouns	53.00	53.00	39.00	39.00
	monotonicity positives	43.00	43.00	25.00	25.00
	monotonicity negatives	42.00	37.00	28.00	22.00
	uncertainty adverbs	9.1	79.9	4.67	77.56
Under presupposition triggers	pronouns	72.30	72.30	65.35	65.35
	monotonicity positives	55.65	55.65	38.04	38.04
	monotonicity negatives	29.39	52.83	19.04	39.78
	uncertainty adverbs (sample)	7.52	86.91	2.65	87.78
Under non-factives	pronouns	40.04	56.65	41.00	51.00
	monotonicity positives	30.17	60.17	27.26	28.83
	monotonicity negatives	22.35	52.91	17.43	30.30

Table 2: Comparison between ChatGPT's and GPT3.5's experiment results

Looking only at the accuracy columns, we see that ChatGPT constantly beats its predecessor. However, the entailment columns show that it also more readily predicts entailment in almost all the experiments regardless of whether or not it is correct. It is also clear that both models share the same erroneous tendencies. Therefore the higher accuracy scores as such do not necessarily mean that the newer model is more consistent and reliable.

F LLaMA 2 experiment results

The results of our experiments with the LLaMA 2 model with 70 bilion parameters are detailed in Table 3.

G Consistency across prompts

It is well known that GPT models are very sensitive to the phrasing of prompts: even slight modifications of the prompt can produce considerably different results. Researchers point out that ChatGPT "even produces inconsistent outputs for paraphrased inputs generated by itself" (Jang and Lukasiewicz, 2023). To find out if this is the case also for the inference types explored in this work, we ask ChatGPT twice to generate a paraphrase of our original prompt (see Appendix B). We obtain the following two paraphrases.

(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"

(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"

For both prompts we run the same set of experiments as described in Section 3, with the same data and setup (changing only the prompt).

The first paraphrased prompt demonstrates the same pattern as the original prompt (see Section 3): embedding the premises under either a presupposition trigger or a non-factive predicate makes the model more likely to predict entailment (with only one exception - see Table 4).

Interestingly, the second paraphrased prompt behaves differently: while presupposition triggers again increase entailment predictions, non-factives have the opposite effect: they cause the model to predict entailment *less* often.

The comparison between the three prompts is shown in Table 4.

While the behavior of the second paraphrased prompt might seem somewhat more logical (it predicts less entailment under non-factives which indeed cancel entailment), it is not always more accurate: the original prompt has the best accuracy for the "neutral" test sets with standalone prompts; the first paraphrased prompt yields the best accuracy for the "entailing" test sets, the second paraphrased prompt - for the "neutral" test sets with embeddings, and in general they all struggle with the inference types discussed in this work.

		accuracy(%)	entailment (%)	neutral (%)
Standalone	pronouns	31.00	31.00	61.00
	monotonicity positives	36.00	36.00	59.00
	monotonicity negatives	35.00	46.00	35.00
	uncertainty adverbs	9.00	89.89	9.00
Under presupposition triggers (the initial labels aren't expected to change)	pronouns	32.00	32.00	58.13
	monotonicity positives	36.87	36.87	44.87
	monotonicity negatives	38.52	32.43	38.52
	uncertainty adverbs (sample)	29.35	58.39	29.35
Under non-factives (all the labels should change to "NEUTRAL")	pronouns	64.74	29.74	64.74
	monotonicity positives	52.35	34.35	52.35
	monotonicity negatives	41.96	29.30	41.96

Table 3: LLaMA 2 70b results. The green background color means that the expected label is "ENTAILMENT"; the blue background color means that the expected label is "NEUTRAL".

		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 4: The experiment results for the original prompt and its two paraphrases suggested by ChatGPT itself. The green rows indicate the datasets where the expected label is always "ENTAILMENT". The blue rows indicate the datasets where the expected lable is always "NEUTRAL". The pink cells indicate the results that don't fit the pattern exhibited by the original prompt (see Appendix B). The bold figures indicate the highest accuracy for a specific test set across all 3 prompts.

H Chain-of-thought prompting

H.1 Chain-of-thought experiments

As part of our research, we ran the set of experiments described in Section 3, using zero-shot chain-of-thought (CoT) prompting, which, basically, consists in adding the phrase *"Let's think step by step"* to the end of the prompt (Kojima et al., 2023). Here we present the experiment details.

We slightly modified our original prompt (see Appendix B) 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.*

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

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

> *Therefore, the one-word answer (True, False or Neutral) is*

For the CoT experiments with standalone premises we use the same 100-example test sets as for the original experiments (see Section 3 for details). For experiments with embeddings we sample 100 sentence pairs out of each 2300-example test set.

The comparison between the original experiments described in section 3 and the CoT experiments is shown in Table 5.

As can be seen in the table, the CoT approach hardly proves helpful for the inferences discussed

		original prompt		chain-of-though 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 5: The experiment results for the original prompt and the CoT prompt. The green rows indicate the datasets where the expected label is always "ENTAILMENT". The blue rows indicate the datasets where the expected label is always "NEUTRAL". The bold figures indicate which prompt scored higher on a specific test set.

in this work. The CoT method scores much higher on all the "neutral" test sets, but much lower (with one exception) on all the "entailment" test sets.

Another important observation is that the results of the CoT experiments *do not share the pattern* observed in the previous experiments (more entailment predictions for embedded premises). Rather - surprisingly - it exhibits quite the opposite pattern! Namely, the number of entailment predictions *decreases* when we embed the premises under presupposition triggers or non-factives. At the same time, the number of neutral predictions rises when the premises are embedded (to show this, Table 5 includes an additional column reflecting the percentage of neutral predictions in each CoT experiment). In other words, the pattern we observed earlier is, in a sense, preserved, but now it affects the "neutral", rather than the "entailment" label (again regardless of the correct inference). As for the entailment predictions, now under embeddings they become *less likely*, rather than *more likely* as we observed before. In our experiments the regular prompt and the CoT prompt expose mirror-opposite tendencies, but both of these trends are equally flawed.

H.2 Manual analysis of ChatGPT's reasoning chains

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.

Here we present the results of this analysis. 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 that there is a school policy regarding wearing jewelry and Sam broke this rule.

2. Next, move on to Text 2. It states that "Sam violated something." It is not specified what Sam violated, but it is implied that Sam did break a rule or a policy of some sort.

3. Based on Text 1 and Text 2, it is safe to say that Text 2 is true. The fact that Text 1 explicitly mentions a specific policy that Sam violated supports the idea that a violation did occur. Text 2 confirms this by stating that Sam violated something, thus aligning with the information provided in Text 1.

Therefore, the answer is: Text 2 is true.

Prediction: *Entailment.*

Yet in other cases (about 23.6%) the model is "right for the wrong reasons" ariving at a correct prediction through a wrong reasoning chain. For example:

Premise: *It was reported that every alien used some quarters.*

Hypothesis: *Every alien used coins.*

Chain of thought: *Step 1: Let's analyze Text 1 to extract the key information. The statement says "Every alien used some quarters." This means that all aliens used quarters, but it is possible they also used other types of coins or currency.*

Step 2: Now, let's examine Text 2 in light of Text 1. The statement in Text 2 says "Every alien used coins." This statement seems to suggest that all aliens used coins, but it does not specify which type of coins they used.

Step 3: ***Since Text 2 does not specify which types of coins alien used, we cannot say for sure whether it is true or false****. However, we can say that the statement in Text 2 is neutral since it does not contradict Text 1, but also does not provide us with any new information that supports Text 1.*

Prediction: *Neutral.*

The correct label here is indeed "neutral", however the reason cited in the chain of thought (*"Since Text 2 does not specify which types of coins alien used, we cannot say for sure whether it is true or false"*) is wrong. In reality, the relation is neutral because of the non-factive expression *"It was reported"*.

Remarkably, in one case, the model does not arrive at the gold label, but seems to be - at least pragmatically - correct in its reasoning, and the CoT clearly makes sense:

Premise: *it's worth noting that I don't eat chicken skin.*

Hypothesis: *I don't eat chicken skin at any cost.*

Chain of thought: *Step 1: Analyze Text 1 The first text states that the person speaking does not eat chicken skin. This seems pretty clear and straightforward.*

Step 2: Analyze Text 2 The second text states that the person does not eat chicken skin at any cost. This seems to be an exaggerated statement compared to the first text. The phrase "at any cost" implies that the person is willing to go to great lengths or make great sacrifices to avoid eating chicken skin. ***This level of emphasis is not present in Text 1****, so it's possible that Text 2 is an exaggerated version of the truth.*

Step 3: Decide if Text 2 is true, false, or neutral Based on the information we have, we can't definitively say whether Text 2 is true or false. It's possible that the person speaking really does avoid eating chicken skin at all costs, but it's also possible that they are exaggerating and simply avoid it when possible. Therefore, the answer is neutral.

Model's prediction: *Neutral.*

Gold label: *Entailment.*

Finally, in the remaining cases both the chain of thought and the resulting prediction are wrong. For example:

Premise: *The school's deputy headmaster, Howard Jones, told her to remove the crucifix around her neck.*

Hypothesis: *The school's deputy headmaster, Howard Jones, told her to do something.*

Chain of thought: *1. Text 1 states that Howard Jones, the deputy headmaster of the school, asked a student to remove her crucifix.*

2. Text 2 only confirms that Howard Jones did tell the student to do something, but it doesn't specify what that something was.

3. Therefore, Text 2 cannot be used to determine whether or not Text 1 is true or false.

4. However, it is important to note that if Text 1 is true, and Howard Jones did ask a student to remove their crucifix, this would be a violation of the student's religious freedom and may constitute discrimination.

Model's prediction: *Neutral.*

Correct prediction: *Entailment.*

Here the model not only demonstrates misunderstanding of the task trying to determine the veracity of Text 1 based on Text 2, while clearly instructed to do the opposite, but also makes a moral judgement (see step 4) which is totally out of place in this task.

Below we list some more observations regarding the analyzed reasoning chains:

1. The embedding context (presupposition triggers or non-factives) are sometimes mentioned in the chain of thought, but are never used as a basis for the final decision. (One consequence of this is that for the test sets with non-factives all the correct answers result from wrong reasoning chains.)

2. For premises with uncertainty adverbs, in 8 out of the 10 analyzed cases the adverb is mentioned and its meaning explained, but only in 3 out of the 10 cases the adverb serves as a basis for the final answer.

3. The model constantly misinterprets indefinite pronouns as referring to a specific entity (even though it "knows" that indefinite pronouns are "generic or underspecified" terms encompassing *any* entity or individual - see Figure 2). Hence the incorrect "neutral" labels for most cases of grammatically-specified entailment. For example ChatGPT decides that *"Mary lent him money"* does not entail *"Someone lent him money"* because "someone" in Text 2 *"could be referring to someone other than Mary"*. (The correct answer is, of course, "entailment" because "someone" is a generic term encompassing *any* individual including Mary.)

4. The model often gets confused about the monotonicity entailment directions (upward vs. downward), stating, for example, that *"No alien ate pork"* entails *"No alien ate meat"* since *"pork is a type of meat"*, but *"Every alien used some quarters"* does *not* entail *"Every alien used some coins"* because Text 2 *"does not specify which type of coins they used"*.

5. More generally, the model usually predicts entailment when Text 2 contains a more specific mention than Text 1 (which is, in fact, only correct for cases of downward entailment), and vice versa.

6. The reasoning chains are mostly vague, excessively wordy, with unclear logical relations between steps, which makes them hard to understand and analyse, and often contain obvious logical errors (e.g. *"Text 2 is likely true, as it directly contradicts the assumption made in Text 1"*).

7. The CoT can sometimes misrepresent the contents of the input sentences. For example the model claims that the text *"I love something outside the city"* doesn't mention "love".

8. Different chains of thought exhibit contradictory logics. For example, one CoT says *"There is no contradiction between the two texts... Therefore, Text 2 can be determined as true"*, while another reasoning chain states: *"Text 2 does not contradict Text 1, so it is neutral."*

Quantitatively, the results of this analysis are represented in Table 6.

The analysis shows that zero-shot CoT prompting fails to improve ChatGPT's performance on the task because of various flaws in the generated reasoning chains.

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 6: 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.