Pragmatic Reasoning in GPT Models: Replication of a Subtle Negation Effect

Francesca Capuano (francesca.capuano@uni-tuebingen.de)
Department of Psychology, University of Tübingen
Schleichstraße 4, 72076 Tübingen, Germany

Barbara Kaup (barbara.kaup@uni-tuebingen.de)
Department of Psychology, University of Tübingen
Schleichstraße 4, 72076 Tübingen, Germany

Abstract

This study explores whether Large Language Models (LLMs) can mimic human cognitive processes, particularly pragmatic reasoning in language processing. Focusing on how humans tend to offer semantically similar alternatives in response to negated statements, the research examines if LLMs, both base and fine-tuned, exhibit this behavior. The experiment involves a cloze task, where the models provide completions to negative sentences. Findings reveal that chat models closely resemble human behavior, while completion models align worse with human responses. This indicates that mere linguistic input statistics might be inadequate for LLMs to develop behaviours consistent with pragmatic reasoning. Instead, conversational fine-tuning appears to enable these models to adopt behaviors akin to human pragmatic reasoning. This research not only sheds light on LLMs' capabilities but also prompts further inquiry into language acquisition, especially the role of conversational interactions in developing pragmatic reasoning.

Keywords: GPT; ChatGPT; LLaMA; Negation; Pragmatics; Alternatives; LLMs; Theory of Mind.

Introduction

Large language models (LLMs) like the GPT series have boosted exponential progress in the treatment of natural language. So much so that nowadays diagnostics tend to be more and more often informed by psycholinguistics: the question shifts from whether the models produce a generally acceptable (*grammatical*) performance, to whether it resembles that of humans in broader cognitive terms. In fact, the models' behaviour is by now comparable to that of humans in a variety of settings and tasks (for an overview, see Chang & Bergen, 2023). For this reason, investigating their abilities and underlying mechanisms becomes an interesting avenue from a cognitive perspective, as it can potentially inform research on human language processing. An example comes from a recent branch of research on LLMs looking at the emergence of linguistic and extra-linguistic abilities which in humans are debated to be either innate or resulting from experiences that the machines might not have access to, such as embodied or social interactions (Futrell et al., 2019; Gulordava, Bojanowski, Grave, Linzen, & Baroni, 2018; Linzen & Baroni, 2021; Sinclair, Jumelet, Zuidema, & Fernández, 2022; Trott, Jones, Chang, Michaelov, & Bergen, 2023; Wilcox, Futrell, & Levy, 2023). For example, complex syntactic representations (Linzen & Baroni, 2021), as well as behaviour compatible with belief attribution (Trott et al., 2023), are shown to emerge simply from a language modeling task: some LLMs can correctly predict long-distance agreements; some seem to encode the depth of the syntactic tree of the sentences they are fed with, even though the input does not encode any hierarchical information; they can perform above chance on a false belief task, seemingly inferring the mental states of others, even though their training objective is limited to word prediction. In principle, the emergence of such abilities would suggest that these are learnable from statistical regularities of language alone. In fact, findings on complex linguistic emergent abilities have been used to *in principle* reject Chomsky's *poverty-of-stimulus* argument (Chomsky, 1986, e.g.), according to which the linguistic input to which humans are exposed would be insufficient to infer the correct grammar of a language. More in general, this approach is apt to test the sufficiency of the exposure to the statistical regularities of language and the training regime, raising questions on innateness, as well as the necessity to rely on specific mechanisms and experiences to acquire specific abilities.

A fertile testing ground for the emergence of behaviours compatible with the use of extra-linguistic knowledge is negation. For a long time now, an automatic treatment of linguistic negation has been challenging (Dobreva & Keller, 2021; Ettinger, 2020; Hosseini et al., 2021; Jang, Ye, & Seo, 2023; Kassner & Schütze, 2019; Truong, Baldwin, Verspoor, & Cohn, 2023). One of the underlying issues is that logical approaches to the meaning of negation do not fully grasp the actual meaning of negation in natural language (Horn, 1989). This is clearly the case for alternatives to a negated entity: logically speaking, any member of the complement set of a negated entity (e.g. *This is not a dog*) should be an equally valid alternative; effectively though, speakers find alternatives that are very similar to the negated entity particularly likely (e.g. a wolf is a more likely alternative to a dog than a screwdriver and a segment fragment such as *I see no dog but I see a*___ is more likely completed with *wolf* than with *screwdriver*) (Capuano, Dudschig, Günther, & Kaup, 2021; Kruszewski, Paperno, Bernardi, & Baroni, 2016). This preference has been argued to be grounded in pragmatics: everything else being equal, speakers tend to maximise the informativity of their statements while avoiding effort and over-informativity (Grice, 1975). Negation is often optimally informative in those cases where it corrects a false presupposition (Clark & Clark, 1977; Givón, 1978; Glenberg, Robertson, Jansen, & Johnson-Glenberg, 1999; Horn, 1989; Norde-

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

Experiment	Sentence Type	Task	Polarity	Sentence
1	there	1a	negative	There is no X here, but there is Y.
		1b	affirmative	There is X here, and there is Y there.
	this	1a	negative	This is not X, it is Y.
		1b	affirmative	This is X, and that is Y.
2	there	2	negative	There is no X here, but there is Y.
			affirmative	There is X here, and there is Y there.
3	this	3	negative	This is not X, it is Y.
			affirmative	This is X, and that is Y.
4	this	4	negative	This is not X here, but it is Y.
			affirmative	This is X here, and that is Y there.
5	this	5	negative	This is not X here, it is Y.
			affirmative	This is X here, and that is Y there.
6	see	6	negative	(Pron) see(s) no X, but (Pron) see(s) Y.
			affirmative	(Pron) see(s) X and (Pron) see(s) Y.
7	want	7	negative	(Pron) want(s) no X, but (Pron) want(s) Y.
			affirmative	(Pron) want(s) X and (Pron) want(s) Y.

Table 1: Overview of the experiments in Capuano et al. (2021).

meyer & Frank, 2014; Nordmeyer & Frank, 2015; Wason, 1965, 1972). Therefore, speakers are likely to infer a corrective reading of negation when they prefer similar alternatives to a negated entity: two similar entities (e.g. a dog and a wolf) are more likely to be confused with one another than two less similar entities (e.g. a dog and a screwdriver) (Capuano et al., 2021). For such reasons, negation has been argued to require pragmatic reasoning in order to be fully understood. In fact, in order to infer the communicative intent behind the use of negation, a recipient needs to be able to reason on the mental state of their interlocutor and the process which led to that choice (Frank & Goodman, 2012). In the context of cooperative conversation, this process is assumed to follow the application of the Gricean maxims of cooperation (Grice, 1975). An indication that pragmatic reasoning is particularly relevant in the case of negation comes from the observation that people seem to correctly understand and use negative sentences only when they have acquired Theory of Mind (Cuccio, 2011; Schindele, Lüdtke, & Kaup, 2008).

Large Language Models are not explicitly trained to infer the pragmatic reasoning of their interlocutor. Therefore, if LLMs exhibit a pragmatics-informed understanding of negation, then, in principle, the correct usage of negation could be acquired through statistical regularities of the linguistic input alone, assuming that the training data reflects the usage of negation displayed by the humans in the experiments. The question becomes even more relevant when we compare the performance of simple autoregressive models with models fine-tuned on chat data, which could be argued to provide the model with some form of indirect social interaction, and with models trained with RLHF (Ouyang et al., 2022), as the Reinforcement Learning framework has been paralleled to *Theory of Mind* (e.g. Jara-Ettinger, 2019). A better alignment of the latter types of models with human data would be in line with the centrality for negation of pragmatic reasoning arising in conversational settings.

We tried to replicate a behavioural finding that is argued to result from a pragmatic understanding of negation with a series of GPT models. In particular, we looked at the production of alternatives to negated entities in the context of a cloze task: the preference for similar alternatives is so peculiar to negation, that alternatives produced to negated entities are even more similar than completions to conjuncts produced in a similar affirmative context (Capuano et al., 2021). A range of models with different training objectives and regimes were tested, in order to provide a first overview of the impact of these differences on a pragmatically-informed treatment of negation.

Method

The study with humans

In a series of seven experiments, Capuano et al. (2021) presented subjects with minimalistic negative and affirmative sentences to complete in a natural way (e.g. *This is not a goat, it is*___ and *This is a goat, and that is*___) (for a similar approach used in the study of scalar inferences, see Ronai & Xiang, 2023; Hu, Levy, Degen, & Schuster, 2023). 50 common nouns were selected to construct the sentences (see their Appendix for the full list). Each experiment corresponds to a different sentential context (e.g. *This is not a dog, it is* ___ vs. *There is no dog there, but there is* ___). In every experiment, the participants were presented with all 50 items once, each item randomly assigned to either the negative or the affirmative condition. An overview of the experiments can be seen in Table 1.

The hypothesis was that, if similar alternatives are peculiar to negation, subjects should produce continuations to the negative sentences that are semantically more similar to the negated entity than the continuations to the affirmative sentences. In other words, the continuation to *This is not a dog, it is* ___ (e.g. *a wolf*) was expected to be more semantically similar to *dog* than the continuation to *This is a dog, and that*

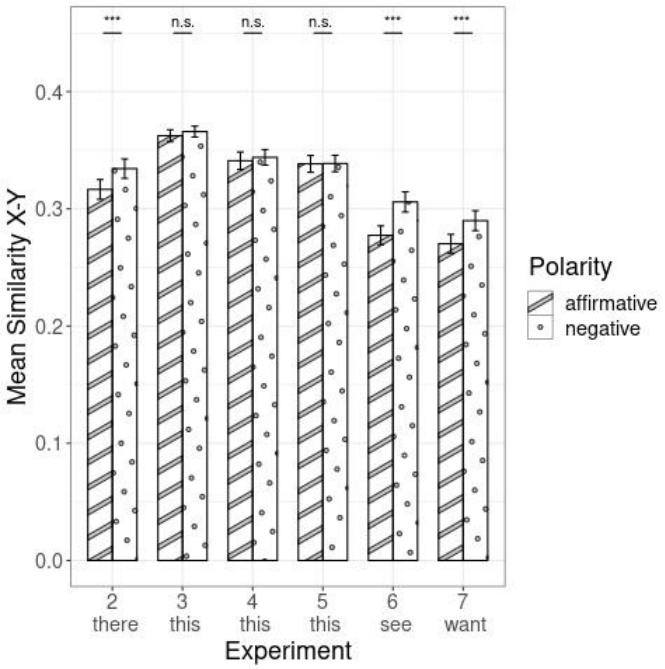

Figure 1: Results of the experiments with humans (Capuano et al., 2021).

is ___ (e.g. *a leash*). Semantic similarity was quantified in the cosine similarity scores derived from the best-performing vector-space in Baroni, Dinu, and Kruszewski (2014) with the help of the LSAfun package (Günther, Dudschig, & Kaup, 2015).

Experiment 1 is not considered here because the design was not optimal in the original study (a between-subject design which resulted in insufficient power). We will focus on the results of Experiments 2 to 7. In these experiments, three out of four sentential contexts produced the expected main effect of Polarity, with an advantage for negation. Three variations of the *This* sentential context produced a null effect, which was argued to be due to a ceiling effect. Therefore, from the human data we can conclude that continuations to negative sentences are at least as similar to continuations to affirmative sentences, if not more, confirming that very similar alternatives seem to be specific to negation. The results of Experiments 2-7 from Capuano et al. (2021) are summarised in Figure 1.

The study with the models

In order to reproduce the open cloze tasks administered to the human participants, we relied on the Completion and Chat Completion API from OpenAI to sample responses from the models. We prompted the models with the incomplete sentences (e.g. *There is no dog here, but there is*) and generated completions until a full stop was encountered, for a maximum of 20 tokens. Every model was queried with 1000 sentences, corresponding to the amount of data collected from 20 participants (each of the 50 items was presented 20 times). Similarly to the human study, the sentences were randomly assigned to either the affirmative or the negative Polarity condition.

The analysis was restricted to those experiments where the sentential contexts end with the completion and no further linguistic material. Therefore, we only reproduced Experiments 3, 6 and 7 from Capuano et al. (2021).

We interrogated four models: `davinci-002` (from now "GPT-3"), `gpt-3.5-turbo-instruct` (from now "InstructGPT"), `gpt-3.5-turbo` ("GPT-3.5") and `gpt-4` ("GPT-4"). The first two are trained on next token prediction, with InstructGPT being additionally fine-tuned on instructions and through RLHF (Ouyang et al., 2022). The latter two are chat models, fine-tuned for the purpose of multi-turn interactions. The first two models could be interrogated through the Completion API (Legacy), whereas the two latter only with the Chat Completion API. Although GPT-3.5 and GPT-4 are chat models designed for multi-turn conversations, they can perform single-turn tasks when queries are limited to the user role. We limited the queries in this way to keep the task akin across models. The completions were generated on the 22nd January 2024.

In order to address the lack of transparency of the OpenAI models, we repeated the experiments with four models from the LLaMA-2 series: `Llama-7b`, `Llama-7b-chat`, `Llama-70b` and `Llama-70b-chat` (Touvron et al., 2023). These allow us to run more controlled comparisons, since the base models and the corresponding chat-fine-tuned versions use the same number of parameters and are pretrained on the same data, which rules out that differences in performance are attributable to whether the pretraining corpora include code (e.g. Kim & Schuster, 2023).

Llama chat models would not provide usable completions

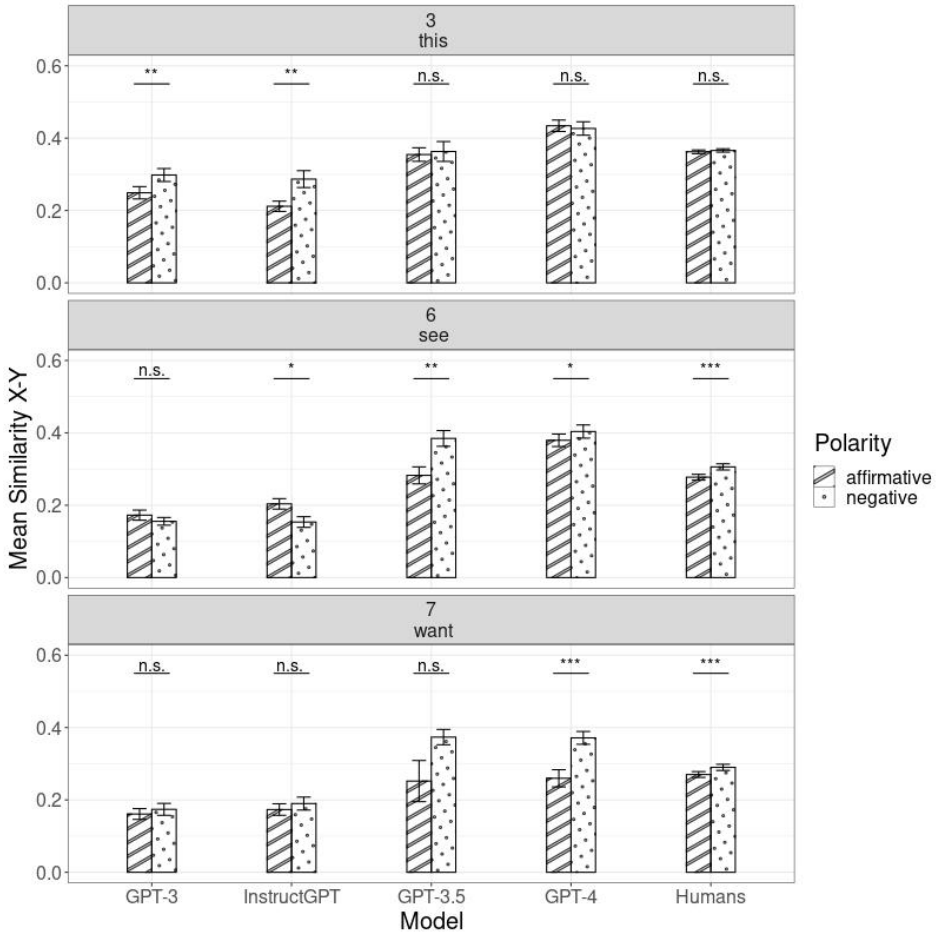

Figure 2: Results of Experiments 3, 6 and 7 with the OpenAI models, side to side to the human results. Mean similarity scores ± standard error are plotted.

	GPT-3	InstructGPT	GPT-3.5 (Chat)	GPT-4 (Chat)	Humans
Experiment 3 (this)	neg >aff**	neg >aff**	n.s.	n.s.	n.s.
Experiment 6 (see)	n.s.	neg <aff*	neg >aff**	neg > aff*	neg >aff***
Experiment 7 (want)	n.s.	n.s.	n.s.	neg > aff***	neg >aff***

Table 2: Summary of the results of the OpenAI models.

to the sentences when prompted without instructions. Therefore, we prompted the systems with the instruction *Complete the given sentences with one noun or one adjective and one noun*. Base models were not provided with instructions. [1]

The code used to generate the completions and the completions are available at `https://github.com/FrancescaCapuano/pragmatic-reasoning-in-llms.git`.

Results

Data cleaning across all experiments followed a very similar procedure to that in Capuano et al. (2021): the completions were lowercased, stripped, and deprived of punctuation (except for "-"). Stopwords were removed. Only answers consisting of one noun were considered, except if the noun was identical to the negated noun, in which case the trial was discarded. The cosine similarity score was calculated for each item noun-completion noun pair using the LSAfun package Günther et al. (2015) with the best performing vector space

[1] Base models were also prompted with the same instructions as the chat models to ensure that the results would not change qualitatively.

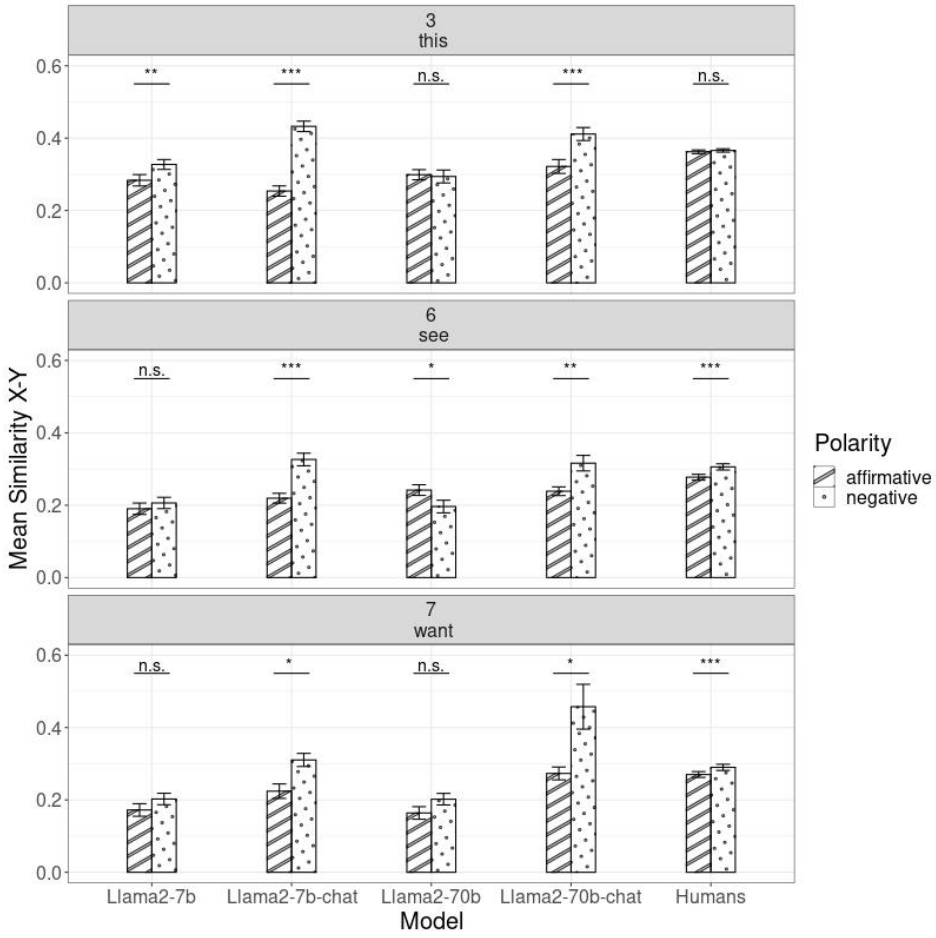

Figure 3: Results of Experiments 3, 6 and 7 with the LLama models, side to side to the human results. Mean similarity scores $\pm$ standard error are plotted.

	Llama-7b	Llama-7b-chat	Llama-70b	Llama-70b-chat	Humans
Experiment 3 (this)	neg >aff**	neg >aff***	n.s.	neg >aff***	n.s.
Experiment 6 (see)	n.s.	neg >aff***	neg <aff*	neg >aff**	neg >aff***
Experiment 7 (want)	n.s.	neg >aff*	n.s.	neg >aff*	neg >aff***

Table 3: Summary of the results of the Llama-2 models

.

from Baroni et al. (2014). Completion nouns that were not found in the semantic space were not considered.

We did not provide instructions to the models to keep the results comparable across completion and chat models (except for the Llama chat models). This procedure was different from the human study, where participants were explicitly instructed to complete the sentences with either one noun or a determiner and a noun. The discrepancy resulted in considerably larger data loss than in the study with the humans (on average 63% of responses per experiment were excluded), as the models often produced longer completions than the humans.

Analogously to Capuano et al. (2021), the data were analysed with a linear mixed effect model. We always tried to fit the following model:

$$Cosine \sim Polarity + (1 + Polarity|Item) \quad (1)$$

unless singular fit warnings were issued, in which case we dropped the random slope. The results of the OpenAI models are plotted in Figure 2. The results of the chat models best align with the human results. In particular, GPT-4 replicates the results of all three human experiments. On the other hand, completion models do not replicate any of the human patterns. An overview of how the results of models align with the human results can be found in Table 2.

The results of the Llama models are plotted in Figure 3 and summarised in Table 3. Again, chat models align best to the results of the human experiments. Table 4 reports the

percentage of overlap of the analysed models' responses with the analysed human responses, confirming that the qualitative alignment of the effects between models and humans reflects a larger proportions of shared answers.

Discussion

We queried a series of LLMs to replicate some behavioural findings on negation that are argued to be motivated by the pragmatic reasoning behind negative sentences. Specifically, we looked at the finding that negative sentences tend to be completed with alternatives that are particularly similar to the material in the scope of negation (Capuano et al., 2021). This finding suggests that, even in very minimalistic and unconstrained sentential contexts, recipients tend to infer a corrective intent of negation. In humans, this inference can be justified as a consequence of pragmatic reasoning, which includes assuming an informative interlocutor, therefore a context that maximises the informativity of the chosen statement (Horn, 1989; Grice, 1975; Frank & Goodman, 2012; Nordmeyer & Frank, 2015).

GPT-3, GPT-3.5 and GPT-4 replicate the finding that completions to negation are at least as similar if not more similar than completions to affirmative sentences. The same goes for Llama2-7b, Llama2-7b-chat and Llama2-70b-chat. Strikingly though, the patterns of results of the human participants and those of the chat models are extremely similar. In particular, GPT-4 replicates all of the behavioural results. Same as for humans, Experiment 3 does not display an effect of Polarity in the OpenAI chat models, which could similarly be attributed to a ceiling effect. Overall, chat models seem to best align to the behavioural findings. Nevertheless, further work should address ways to minimize data loss in the data cleaning procedure and confirm the reliability of these findings.

The results are suggestive that base models simply trained on next-token prediction might not be capable of capturing subtle effects of negation attributed to pragmatic reasoning. Therefore, a behaviour compatible with pragmatic reasoning might possibly not be acquired via the statistical regularities of the linguistic input alone. Other phenomena attributable to pragmatic reasoning should be investigated in this light to bring more evidence forward. Interestingly, a model fine-tuned with Reinforcement Learning like InstructGPT also fails at replicating the findings, and even shows an effect in the opposite direction. Since these models are trained to enhance their "sensitivity to the intent of the interlocutor", one might have expected the results to be closer to the behavioural data when it comes to inferring the communicative (corrective) intent of the interlocutor. Instead, models fine-tuned as chatbots seem to be the only ones that align to the human behaviour. Being trained on conversational data and with the objective of predicting conversational turns might contribute to the emergence of behaviour that is in line with human pragmatic reasoning.

Our results are in principle compatible with the idea that simple language modeling might be insufficient to exhibit behaviour consistent with pragmatic reasoning. Further effort should go into testing this idea in other settings where pragmatically informed behaviour is expected. Relatedly, the fact that chat models pick up some behaviours compatible with pragmatic reasoning does not grant that they actually possess pragmatic reasoning. As discussed elsewhere (e.g. Trott et al., 2023), the overlap in observable behaviour between humans and machines does not entail the deployment of the same underlying processes and representations. As a matter of fact, statistical regularities that result from pragmatic reasoning might be simply more apparent in conversational settings (i.e. in the training data of chat models). Investigating the inner workings of the models more closely is another avenue for future research.

Whereas these are some fundamental but still open questions, the current finding is part of a strand whose currently limited aim is to test the sufficiency of the learning mechanisms of Transformer architectures in acquiring behaviours and representations that were unlikely to be displayed by earlier generations of models. In this sense, the present paper provides valuable insights into the potential of different LLMs to replicate human-like pragmatic reasoning in the domain of negation. The results give a first overview on how different training regimes might contribute to the emergence of pragmatic behaviour but also support the relevance of a pragmatic approach to negation.

Model	Exp 3 (this)	Exp 6 (see)	Exp 7 (want)
GPT-3	59	51	45
InstructGPT	55	55	45
GPT-3.5 (Chat)	92	91	90
GPT-4 (Chat)	99	97	89
Llama2-7b	71	61	56
Llama2-7b-chat	86	88	81
Llama2-70b	77	70	58
Llama2-70b-chat	81	86	81

Table 4: Overlap of model responses with human responses in percentage.

References

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 238–247).

Capuano, F., Dudschig, C., Günther, F., & Kaup, B. (2021). Semantic similarity of alternatives fostered by conversational negation. *Cognitive Science*, *45*(7), e13015.

Chang, T. A., & Bergen, B. K. (2023). Language model behavior: A comprehensive survey. *arXiv preprint arXiv:2303.11504*.

Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. Greenwood Publishing Group.

Clark, H. H., & Clark, E. V. (1977). *Psychology and language: An introduction to psycholinguistics*. Harcourt Brace Jovanovich Ltd.

Cuccio, V. (2011). On negation. what do we need to "say no"? *Rivista Italiana di Filosofia del Linguaggio*, *4*, 47–55.

Dobreva, R., & Keller, F. (2021). Investigating negation in pre-trained vision-and-language models. In *Proceedings of the fourth blackboxnlp workshop on analyzing and interpreting neural networks for nlp* (pp. 350–362).

Ettinger, A. (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.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., & Levy, R. (2019). Neural language models as psycholinguistic subjects: Representations of syntactic state. *arXiv preprint arXiv:1903.03260*.

Givón, T. (1978). Negation in language: Pragmatics, function, ontology. In *Pragmatics* (pp. 69–112). Brill.

Glenberg, A. M., Robertson, D. A., Jansen, J. L., & Johnson-Glenberg, M. C. (1999). Not propositions. *Cognitive Systems Research*, *1*(1), 19–33.

Grice, H. P. (1975). Logic and conversation. In *Speech acts* (pp. 41–58). Brill.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. *arXiv preprint arXiv:1803.11138*.

Günther, F., Dudschig, C., & Kaup, B. (2015). Lsafun-an r package for computations based on latent semantic analysis. *Behavior Research Methods*, *47*(4), 930–944.

Horn, L. (1989). *A natural history of negation*. The University of Chicago Press.

Hosseini, A., Reddy, S., Bahdanau, D., Hjelm, R. D., Sordoni, A., & Courville, A. (2021). Understanding by understanding not: Modeling negation in language models. *arXiv preprint arXiv:2105.03519*.

Hu, J., Levy, R., Degen, J., & Schuster, S. (2023). Expectations over unspoken alternatives predict pragmatic inferences. *Transactions of the Association for Computational Linguistics*, *11*, 885–901.

Jang, J., Ye, S., & Seo, M. (2023). Can large language models truly understand prompts? a case study with negated prompts. In *Transfer learning for natural language processing workshop* (pp. 52–62).

Jara-Ettinger, J. (2019). Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, *29*, 105–110.

Kassner, N., & Schütze, H. (2019). Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. *arXiv preprint arXiv:1911.03343*.

Kim, N., & Schuster, S. (2023). Entity tracking in language models. *arXiv preprint arXiv:2305.02363*.

Kruszewski, G., Paperno, D., Bernardi, R., & Baroni, M. (2016). There is no logical negation here, but there are alternatives: Modeling conversational negation with distributional semantics. *Computational Linguistics*, *42*(4), 637–660.

Linzen, T., & Baroni, M. (2021). Syntactic structure from deep learning. *Annual Review of Linguistics*, *7*, 195–212.

Nordemeyer, A., & Frank, M. (2014). A pragmatic account of the processing of negative sentences. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 36).

Nordmeyer, A., & Frank, M. C. (2015). *Negation is only hard to process when it is pragmatically infelicitous*. (Unpublished manuscript)

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... others (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Ronai, E., & Xiang, M. (2023). Tracking the activation of scalar alternatives with semantic priming. *Experiments in Linguistic Meaning*, *2*, 229–240.

Schindele, R., Lüdtke, J., & Kaup, B. (2008). Comprehending negation: A study with adults diagnosed with high functioning autism or asperger's syndrome. *Intercultural Pragmatics*, *5*(4), 421-444.

Sinclair, A., Jumelet, J., Zuidema, W., & Fernández, R. (2022). Structural persistence in language models: Priming as a window into abstract language representations. *Transactions of the Association for Computational Linguistics*, *10*, 1031–1050.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... others (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Trott, S., Jones, C., Chang, T., Michaelov, J., & Bergen, B. (2023). Do large language models know what humans know? *Cognitive Science*, *47*(7), e13309.

Truong, T. H., Baldwin, T., Verspoor, K., & Cohn, T. (2023). Language models are not naysayers: An analysis of language models on negation benchmarks. *arXiv preprint arXiv:2306.08189*.

Wason, P. C. (1965). The contexts of plausible denial. *Journal of verbal learning and verbal behavior*, *4*(1), 7–11.

Wason, P. C. (1972). In real life negatives are false. *Logique et analyse*, 17–38.

Wilcox, E. G., Futrell, R., & Levy, R. (2023). Using computational models to test syntactic learnability. *Linguistic Inquiry*, 1–44.