

Experimental Pragmatics with Machines: Testing LLM Predictions for the Inferences of Plain and Embedded Disjunctions

Polina Tsvilodub (polina.tsvilodub@uni-tuebingen.de)

Department of Linguistics, University of Tübingen

Paul Marty (paul.pierre.marty@gmail.com)

Institute of Linguistics and Language Technology, University of Malta

Sonia Ramotowska (ramotowska.or.s@gmail.com)

Institute of Logic, Language and Computation, University of Amsterdam

Jacopo Romoli (jacopo.romoli@hhu.de)

Department of Linguistics, University of Düsseldorf

Michael Franke (michael.franke@uni-tuebingen.de)

Department of Linguistics, University of Tübingen

Abstract

Human linguistic communication is based on a variety of inferences that we draw from sentences, often going beyond what is literally said. While there is wide agreement on the basic distinction between entailments, implicatures, and presuppositions, their boundaries are continuously being re-drawn and the status of many inferences remains open. In this paper, we focus on three inferences of plain and embedded disjunctions, and compare them with regular implicatures. We investigate this comparison from the novel perspective of the predictions of state-of-the-art Large Language Models, using the same experimental paradigms as recent studies investigating the same inferences with humans. The results of our best performing models mostly align with those of humans, both in the large differences we find between those inferences and regular implicatures, as well as in fine-grained distinctions among different aspects of those inferences.

Keywords: disjunction; inference; implicature; large language models; pragmatics

Introduction

Human linguistic communication is based on a variety of inferences we draw from sentences, often going beyond what is literally said (Grice, 1975, and much subsequent work). One of the main results of theoretical approaches to meaning is the discovery of a rich typology of inferences, exhibiting different conversational status and diverging behaviour in complex sentences. Over the last two decades, experimental work in pragmatics has tested the often subtle predictions of different theories of these inferences, by comparing them across a variety of tasks and measures with human participants (Noveck, 2018, and references therein). In recent years, advancements in machine learning have added the novel perspective of the investigation of these inferences in non-human language agents such as large language models (LLMs) in systematic comparison to humans (Gauthier, Hu, Wilcox, Qian, & Levy, 2020; Warstadt et al., 2020). The integration of these different perspectives has increased our understanding of the typology mentioned above. However, while there is a general consensus regarding the basic differences between entailments, implicatures, and presupposi-

tions, their boundaries are continuously being re-drawn, and the status of many inferences remains open.

In this paper, we focus on three of such inferences, triggered by *disjunction* (“or”) in different configurations, which we describe in turn. Firstly, plain disjunctions like (1) are associated with so-called ‘IGNORANCE’ inferences (II), suggesting that the speaker is uncertain as to which disjunct is true (1-a) and considers each of them possible (1-b) (Sauerland, 2004, among many others).

- (1) This box contains a blue ball or a yellow ball.
 - a. \rightsquigarrow *The speaker is not certain that the box contains a blue ball and s/he is not certain that it contains a yellow ball.* UNCERTAINTY
 - b. \rightsquigarrow *The speaker deems it possible that the box contains a blue ball and that it contains a yellow ball.* POSSIBILITY

Next, when embedded under a universal quantifier as in (2), disjunctions can give rise to so-called ‘DISTRIBUTIVE’ inferences (DI), suggesting that the property associated with each disjunct applies to some but not all individuals in the domain of the quantifier (Sauerland, 2004; Fox, 2007, among others).

- (2) Every box contains a blue ball or a yellow ball.
 - a. \rightsquigarrow *Not every box contains a blue ball and not every box contains a yellow ball.* NEGATED UNIVERSAL
 - b. \rightsquigarrow *Some box contains a blue ball and some box contains a yellow ball.* DISTRIBUTIVE

Finally, in the scope of a possibility modal as in (3), the resulting inference is called ‘FREE CHOICE’ (FC), conveying in this case that each of the disjuncts is an open possibility (Kamp, 1978; Fox, 2007, among others).

- (3) This box might contain a blue ball or a yellow ball.
 \rightsquigarrow *The box might contain a blue ball and it might contain a yellow ball.* FREE CHOICE

These inferences have all been argued to be derived as, or to result from scalar implicatures (SI) by some theories (Sauerland, 2004; Fox, 2007; Bar-Lev & Fox, 2020, among others) and thus, to rely on the same mechanisms as those deriving regular SIs like the one in (4).

- (4) It is possible that this box contains a blue ball.
 \rightsquigarrow *It is not certain that the box contains a blue ball.*
 IMPLICATURE

Experimental data with human subjects has recently challenged the standard implicature approach to these inferences. First, Marty, Romoli, Sudo, and Breheny (2023) show that all these inferences are more readily derived by speakers than regular SIs, suggesting that they should be treated differently. Second, according to traditional implicature accounts like Sauerland (2004)’s, IGNORANCE and DISTRIBUTIVE inferences are derived in two consecutive steps. For IGNORANCE, the first step involves deriving the UNCERTAINTY (UNC) part in (1-a), from which the POSSIBILITY (POS) part in (1-b) follows; for DISTRIBUTIVE inferences, the first step involves deriving the NEGATED UNIVERSAL part in (2-a), from which the DISTRIBUTIVE part in (2-b) follows. However, the results from Crnič, Chemla, and Fox (2015), Marty, Ramotowska, Romoli, Sudo, and Breheny (2023) and Degano et al. (2023) show that the inferences in (1-b) and (2-b) are accessible to speakers even in the absence of the corresponding inferences in (1-a) and (2-a), suggesting that the former may arise independently of the latter.

Taken together, these findings challenge traditional implicature approaches to IGNORANCE, DISTRIBUTIVE and FREE CHOICE inferences. They are compatible, on the other hand, with more recent implicature approaches, sometimes radically departing from initial assumptions about implicatures, as well as non-implicature approaches to these inferences, such as Aloni (2022), which are not committed to similarities between them and regular SIs and can further derive POSSIBILITY and DISTRIBUTIVE inferences without their UNCERTAINTY and NEGATED UNIVERSAL counterparts.

This Study

We seek to address whether state-of-the-art LLMs predict the fine-grained inferences arising from plain and embedded disjunctions that we just described and whether they exhibit the same pattern of results exhibited in the human data. LLMs have been tested on entailments (Wang et al., 2019, among others), implicatures (Schuster, Chen, & Degen, 2020; E. Li, Schuster, & Degen, 2021; Hu, Levy, & Schuster, 2022), and presuppositions (Jeretic, Warstadt, Bhooshan, & Williams, 2020; Sieker & Zarrieß, 2023; Sravanthi et al., 2024), but to our knowledge, the cases above have not been looked at yet.

To assess the abilities of LLMs, we address two questions. First, we compare their performance to human data from three experiments in Marty, Romoli, et al. (2023), three experiments in Degano et al. (2023) and two experiments in

Marty, Ramotowska, et al. (2023). Second, we compare if LLM predictions align to the same theoretical predictions as human results, which were compared to the traditional implicature account (TIA), revised implicature account (RIA, Bar-Lev & Fox, 2020), and non-implicature account (NIA, Aloni, 2022) in the three studies. In particular, in parallel to what Marty, Romoli, et al. (2023) do with humans, we test whether the predictions of LLMs for FREE CHOICE, IGNORANCE, and DISTRIBUTIVE inferences differ from those of regular IMPLICATURES. In addition, following Marty, Ramotowska, et al. (2023) and Degano et al. (2023), we test whether DISTRIBUTIVE and POSSIBILITY inferences can arise also in the absence of the corresponding NEGATED UNIVERSAL and UNCERTAINTY inferences. The predictions tested are summarized in Table 1.

Before moving on to the details of the study, we should emphasize that the theories above are designed to predict humans’ linguistic behavior and, as such, they are of course not directly theories of LLM mechanics. Nonetheless, we think that using the same experimental paradigms as those used with humans in experimental pragmatics allows us to systematically check whether LLMs, as powerful statistical (fine-tuned) models trained on a lot of textual data, predict the inferences above in a ‘package’ with regular implicatures, or distinguish between them in a way that aligns with humans, thus providing a novel perspective on the debate above.

Table 1: Relevant predictions of the three accounts. ‘IMP’ indicates whether the account predicts the inference in question to be an IMPLICATURE; ‘IND’ indicates whether the account predicts POSSIBILITY and DISTRIBUTIVE inferences to arise INDEPENDENTLY from the corresponding UNCERTAINTY and NEGATED UNIVERSAL ones.

Theories	FREE CHOICE	IGNORANCE		DISTRIBUTIVE	
	IMP	IMP	IND	IMP	IND
TIA	yes	yes	no	yes	no
RIA	yes	yes	yes	yes	yes
NIA	no	no	yes	no	yes

Experiments and Data

To assess LLM predictions, we closely replicated the human experiments’ set-up of the *mystery box* paradigm from Degano et al. (2023), Marty, Romoli, et al. (2023) and Marty, Ramotowska, et al. (2023). In this way, we were able to directly compare the humans’ performance in these experiments with the LLMs’ performance. Therefore, we first describe the three human studies.

Human Studies

In all experiments, participants were presented with pictures of three boxes whose contents were visible and one box whose content was not visible, the so-called *mystery box*.

The visible boxes contained one or two colored balls and the fourth mystery box had a question mark on it. Participants were instructed that the mystery box always has the same contents as one of the visible boxes. They were also introduced with two child characters (Sam and Mia). The characters were familiarized with the rule about the mystery box and, therefore, they could make certain inferences about its contents. In each trial, the four boxes were presented and a sentence was uttered by one of the characters. The truth value of the sentences was manipulated by varying the contents of the visible boxes. Participants’ task was a two-alternative forced choice task wherein they had to judge the sentences as “good” or “bad” given the context and the rule about the mystery box.

In Marty, Romoli, et al. (2023) (abbreviated “MRo”), all target trigger sentences were about the mystery box and tested the robustness of FREE CHOICE (FC), DISTRIBUTIVE (DI) and IGNORANCE (II) inferences against the robustness of regular SIs. For our purposes, we selected materials from Experiments 4–6 of Marty, Romoli, et al. (2023) which only contained positive polarity trigger sentences. For FC, the trigger sentence was of the form “It is possible that the mystery box contains either a [A] ball or a [B] ball”, where [A] and [B] are placeholders for different color adjectives (e.g., yellow and blue); for DI, it was of the form “It is certain that the mystery box contains either a [A] ball or a [B] ball”; for II, it was of the form “The mystery box contains a [A] ball or a [B] ball”; finally, for SI, it was of the form “It is possible that the mystery box contains a [A] ball”. In the critical trials, trigger sentences were presented in a TARGET context like the one in Table 2 (first row), where every visible box contained an [A] ball and no [B] ball. TARGET contexts were designed so that the trigger sentence was false if the inference of interest is present, but true if it is absent. Thus, the more robust a given type of inference, the more participants should select the “bad” response option in these trials and, consequently, the lower the acceptance rate should be.

In Degano et al. (2023) (abbreviated “D”), all target trigger sentences were about the mystery box and were of the form: “The mystery box contains a [A] ball or a [B] ball”. There were two target contexts, TARGET-1 and TARGET-2. TARGET-1 contexts made POSSIBILITY inferences true and UNCERTAINTY inferences false; TARGET-2 contexts made both these inference types false (see Table 2, second and third row). As in MRo study, acceptance rates in these contexts were used as a proxy measure for the robustness of the inferences of interest: the lower the acceptance rate, the more robust the inference(s). Degano et al. (2023) found that TARGET-1 contexts yield higher acceptance rates than TARGET-2 contexts, suggesting that POSSIBILITY inferences are accessible independently of UNCERTAINTY inferences.

In Marty, Ramotowska, et al. (2023) (abbreviated “MRa”), target trigger sentences were either about the mystery box or about the visible boxes. The disjunction in them was embedded either under a NOMINAL or a MODAL universal quantifier.

For the NOMINAL cases, the trigger sentence was of the form “Every visible box contains a [A] ball or a [B] ball”. For the MODAL cases, the trigger sentence was of one of two forms: “The mystery box must contain a A ball or a B ball” (epistemic *must*, Exp. 1) or “[Name] must pick a [A] ball or a [B] ball” (deontic *must*, Exp. 2), where [Name] is a placeholder for a character’s name. Target contexts in this study followed the same logic as in D study above and applied it to DISTRIBUTIVE inferences: TARGET-1 contexts made NEGATED UNIVERSAL inferences true and DISTRIBUTIVE inferences false whereas TARGET-2 contexts made both these inference types false.

All three studies further contained TRUE and FALSE control contexts for each target trigger sentence. These contexts were similar in composition to the target ones but were designed so as to make the trigger sentences either true or false independent of the target inference. Control contexts served to provide clear baselines for acceptance and rejection of the target sentences under investigation. Additionally, all three studies included control trials involving non-target trigger sentences associated with true and false contexts. Responses to these trials were used to assess participants’ general performance in the task independent of the critical items.

Experiments with LLMs

Materials for the LLM studies were constructed by converting all selected vignettes from the human studies described in the previous section to text-based prompts for the LLMs. That is, the visual stimuli of the boxes were converted to textual descriptions. For instance, the TARGET-1 context from Table 2 was represented as shown in Figure 1. General instructions, the cover story, the mystery box rule and examples were prepended to the critical context. An instruction presenting the answer options “good” and “bad” in randomized order was added to each prompt (see Fig. 1).¹

Table 2: Example TARGET context in Marty, Romoli, et al. (2023) and example TARGET-1 and TARGET-1 contexts in Degano et al. (2023) and Marty, Ramotowska, et al. (2023). In these examples, [A] is yellow and [B] is blue.

Context	Example			
TARGET				
	A	A	A	?
TARGET-1				
	A	AB	A	?
TARGET-2				
	A	AA	A	?

¹Full materials can be found under: <http://tinyurl.com/4asyhzv>

{Instructions}

Sam and Mia will be presented with quadruplets of boxes containing balls of various colours. In each quadruplet, Box 1, Box 2 and Box 3 are always open whereas Box 4 is always closed so that Sam and Mia can only ever see what's inside the first three boxes.

However, they have been taught the rule that Box 4, which we will call the mystery box, always has the same contents as one of the three open boxes. Thanks to this rule, Sam and Mia can make certain inferences about what the mystery box contains and does not contain. To illustrate, imagine that Sam is looking at the following quadruplet:

{Examples 1, 2 with example reasoning}

Your task is to decide if this utterance is right given the information available to the characters and the rule that they have learned about how the mystery box works. You will answer 'Good' if you consider that they got it right; otherwise you will answer 'Bad'.

{Few-shot examples}

Sam and Mia will produce utterances about what the mystery box contains and you will decide whether these utterances are appropriate descriptions of the pictures you see.

Currently {Mia/Sam} sees the following quadruplet of boxes:

Box 1 contains a yellow ball. Box 2 contains a yellow ball and a blue ball. Box 3 contains a yellow ball. Box 4 is the mystery box.	Context
--	---------

{Sam / Mia} says: **The mystery box contains a blue ball or a yellow ball.**

How would you judge {Mia/Sam}'s utterance in this situation? Here are your answer options:

good

bad

Your answer: I would judge this utterance as *{good / bad}*.

Figure 1: Example prompt for the LLM experiments. Some parts of the prompt are omitted for brevity (in gray). Boldface trigger sentence is an example from Degano et al. (2023). Underlined sentence is the mystery box rule. Expressions in curly braces in gray vary by study. The character name is sampled at random by-trial. The likelihood for the last word (one of “good” / “bad”, italicized) is retrieved for scoring the trigger, given the context.

Methods

We tested a broad range of LLMs which are freely accessible and where retrieval of log probabilities of strings is possible, selecting different state-of-the-art models so as to cover models with different architectures, fine-tuning, training data composition and scale. Specifically, we used the following LLMs: text-davinci-003 version of GPT-3.5 (Brown et al., 2020), Llama-2 (7B, 13B, 70B parameters, base and chat versions) (Touvron et al., 2023), Mistral-7B (v0.1) and Mistral-7B-Instruct (v0.2) (Jiang et al., 2023), Mixtral-8x7B (v0.1) and Mixtral-8x7B-Instruct (v0.1) (MistralAI, 2023), Pythia (2.8B, 6.9B and 12B parameters, Biderman et al. (2023)), Phi-2 (Y. Li et al., 2023) and Falcon-7B (Almazrouei et al., 2023).²

We used the same prompting and scoring strategies for retrieving predictions from all models for all experiments. Specifically, we follow the design of human studies previously described. Human subjects completed training trials where they learned the experimental task by seeing control conditions and receiving feedback about correctness of their responses. We use all control trials with correct answers that were used in the human training phase as a few-shot prompt for each main trial for the LLMs. Therefore, each experimental item i consisted of a prompt C_i (including the instructions and the cover story description, the few-shot prompt, the critical item context and the task) and the two answer options $o_{ij} = t_{ij1} \dots t_{ijn}$, $j \in \{\text{“good”}, \text{“bad”}\}$, $n \geq 1$, each consisting of n tokens (depending on the model’s tokenizer).

We computed an LLM prediction S_i for item i via retriev-

ing the log probability $\log P_{LLM}(o_{ij} | C_i)$ of each o_{ij} following the item prompt, under each LLM, respectively. We used the average token log probability if an answer option consisted of multiple tokens:

$$\log P_{LLM}(o_{ij} | C_i) = \frac{1}{n} \sum_{l=1}^n P_{LLM}(o_{ijl} | C_i, o_{ij<l}) \quad (1)$$

Given the scores for the two answer options, we identified the LLM prediction for a given item i as the answer option with the maximal log probability $S_i = \arg \max_j \log P(o_{ij} | C_i)$.

For control trials where a correct answer existed, we computed *accuracy* of LLMs by identifying the proportion of items where the chosen response option was correct. For target conditions, we computed the *acceptance rate* predicted by LLMs as the proportion of trigger sentences for which the option “good” was chosen (i.e., the LLM assigned higher log likelihood to this option, given the context). The predicted acceptance rate was used to assess the closeness of LLM predictions and human results. Specifically, we computed the proportion of variance in human data explained by the model predictions by calculating R^2 for the simple linear model regressing human acceptance rates against model predictions, for each condition of each experiment. Finally, to identify overall model fit to human results across experiments, we calculated the adjusted R^2 (Miles, 2005).³

Results

We first compute the accuracy of all models on the control trials of each study. We then take the top five models with

²Except for GPT-3.5, all models are open-access. At the time of submission, the specific model endpoint was discontinued by the provider OpenAI.

³The linear model was $\text{human_accept.rate} \sim \text{model_accept.rate} + \text{condition}$, with seven distinct conditions from the three replicated studies.

Table 3: Average accuracy on control items by source. “L” refers to the Llama-2 model and “Inst.” to Instruct versions of the models. Boldface indicates highest accuracy among LLMs.

Study	Human	GPT-3.5	L-70b	Mixtral-Inst.	Mixtral	Mistral-Inst.
MRo	0.89	0.94	0.69	0.74	0.65	0.54
D	0.95	0.72	0.9	0.75	0.53	0.64
MRa	0.94	0.76	0.86	0.71	0.82	0.75

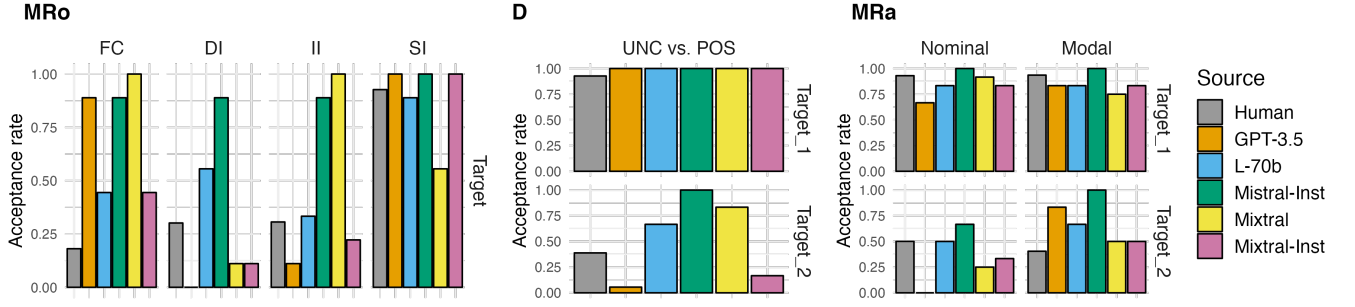


Figure 2: Mean acceptance rate in the target conditions of each study by test case and source (LLMs or humans).

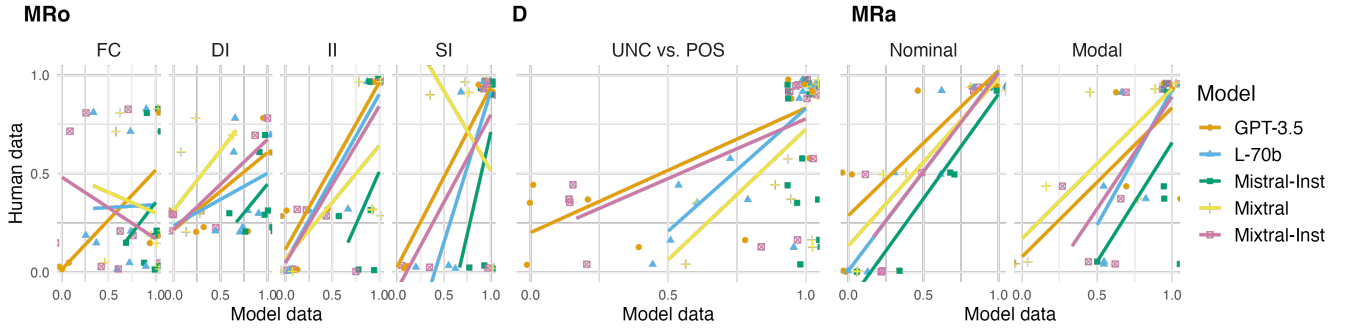


Figure 3: Human acceptance data averaged over all items in each data condition, by trigger type, plotted against model predictions (points). Lines indicate best linear model fit regressing human data against model predictions.

Table 4: Goodness of fit of model predictions to human results in each study and for each test case, reported as the R^2 value for $\text{human_data} \sim \text{model_data}$. Overall model fit reports adjusted R^2 . Higher values are better. Boldface indicates the best model(s) in each condition.

Study	Case	GPT-3.5	L-70b	Mistral-Inst	Mixtral	Mixtral-Inst
MRo	FC	0.53	0	0.1	0.1	0.75
	DI	0.65	0.97	0.97	0.97	0.97
	II	0.95	0.99	0.9	0.44	0.9
	SI	0.99	0.95	0.99	0.22	0.99
D	UNC vs. POS	0.71	0.91	0	0.9	0.49
MRa	Nominal	0.73	0.99	0.99	0.92	0.88
	Modal	0.73	0.9	0.66	0.93	0.7
<i>Overall model fit:</i>		0.68	0.66	0.23	0.18	0.42

the highest accuracy and further investigate the performance of only those models on critical trials. The control accuracy scores of these selected models are reported in Table 3. Full results for all models can be found in the repository. This selection strategy of best models deviated from the design of

human studies where the exclusion criterion for participants was an accuracy of 0.8 in Degano et al. (2023) and Marty, Ramotowska, et al. (2023) and 0.7 in Marty, Romoli, et al. (2023). Table 3 indicates that the performance of LLMs was above chance (0.5 for a single trial), but it was not consistent

across the studies. Furthermore, with one exception, LLMs performed worse than humans.

Second, we investigate the robustness of inferences predicted by the LLMs based on triggers with disjunctions in different configurations, and whether the predicted acceptance rates were similar to human inference rates. To this end, we compared the acceptance rates of the target triggers in the FREE CHOICE, IGNORANCE, DISTRIBUTIVE inference conditions to the acceptance rate of the IMPLICATURE condition (Fig. 2, left facet, first three facets vs. last facet). Visual inspection suggests that some models robustly predicted FC, DI, II inferences but not SIS (L-70b, Mixtral-Inst.), while other models were inconsistent with respect to which inferences they predicted. Next, we assess the LLMs’ fit to human responses via R^2 . To this end, we considered LLM predictions on both target trigger conditions and control conditions (acceptance rates for controls are not shown). Comparing LLM predictions to human inferences, we found that for different inferences, different models best captured human data in terms of explained variance (see Table 4, MRo results). Figure 3 (left) indicates that LLMs were less consistent across triggers with respect to human data in the FC and DI conditions, than in the other two.

Third, we turn to the question of whether LLMs predict that POSSIBILITY and DISTRIBUTIVE inferences occur together with their corresponding UNCERTAINTY and NEGATED UNIVERSAL inferences. Figure 2 (middle) shows the mean acceptance rates for the experiment wherein the triggers contained plain disjunctions, investigating IGNORANCE. Visually, the crucial comparison between the TARGET-1 and TARGET-2 conditions was borne out in prediction of GPT-3.5, Llama-70b and both Mixtral models, albeit less for the base version. Comparing the overall fit of model predictions to human responses across triggers, we found that Llama-70b best captured human responses (Table 4, D). Figure 3 indicates that, both Llama-70b and Mixtral performed close to human data, while Mistral-Instruct did poorly by always accepting both Target conditions.

Figure 2 (right) shows the mean acceptance rates for the experiment wherein the triggers contained a disjunction embedded under a nominal quantifier or a modal, investigating DISTRIBUTIVE inferences. While human inferences remained largely unaffected across modal and nominal contexts (left-most bars, left vs. right facets; cf. Marty, Ramotowska, et al. (2023)), at least visually, LLM predictions varied more across contexts (left vs. right facets). Focusing on the critical contrast between the TARGET-1 and TARGET-2 conditions, we found that it was borne out in predictions of all models in the nominal context, but not for GPT-3.5 and Mistral-Instruct in the modal context. For other models, the contrast was also smaller in the modal context. This discrepancy was reflected in the fit to human data (see Table 4, MRa rows, lower for modal than for nominal cases for some models). The modal context also appears to be more noisy than the nominal context in Figure 3.

To summarize the comparison of LLMs to human predictions across studies and conditions, we found that, overall, GPT-3.5 explained the most variance in human data, closely followed by Llama-2-70b (Table 4, last row). However, the fit to human predictions is not perfect and varies depending on the test conditions, even for the best-performing models. While Llama-70b did poorly on FC, it was otherwise better than GPT-3.5 for many other conditions.

Discussion

We tested the predictions of state-of-the-art LLMs with the main inferences of plain and embedded disjunctions, FREE CHOICE, DISTRIBUTIVE and IGNORANCE inferences, in comparison with regular IMPLICATURES. In particular, we systematically compared the LLM predictions with the human results of Marty, Romoli, et al. (2023); Marty, Ramotowska, et al. (2023) and Degano et al. (2023). In our results, we find a clear reflection of the large difference found in humans between regular SIs and FREE CHOICE, IGNORANCE, and DISTRIBUTIVE inferences in the best performing models we tested. The predictions of these models are also in line with the results from human studies showing that POSSIBILITY and DISTRIBUTIVE inferences can be present in the absence of their UNCERTAINTY and NEGATED UNIVERSAL counterparts.

Overall, we found that LLMs produced different predictions for various types of pragmatic inferences. The results from the LLM experiments show that the best performing models mostly align with humans in the fine-grained distinctions between the inferences above and regular implicatures, as well as on different aspects of those inferences. The results, like those for humans, are not in line with traditional implicature approaches to those inferences. However, we found that LLMs still exhibit inconsistencies across various conditions in non human-like ways. There remain open questions for future explorations, e.g., about the ability of these models to consistently perform on sentences with modals, as well as with the counterparts of the sentences above involving negation, tested in Marty, Romoli, et al. (2023), as it has been found that LLMs might have issues with negation (Truong, Baldwin, Verspoor, & Cohn, 2023). Furthermore, it has been shown that LLMs might be sensitive to superficial aspects of the prompts, which we would not expect to affect human performance (Liu et al., 2023); therefore, testing the robustness of such predictions under different prompting strategies might offer an interesting avenue for future work.

References

- Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocar, R., Debbah, M., . . . Penedo, G. (2023). Falcon-40B: an open large language model with state-of-the-art performance.
- Aloni, M. (2022). Logic and conversation: the case of free choice. *Semantics and Pragmatics*, 15, 5–EA.

- Bar-Lev, M. E., & Fox, D. (2020). Free choice, simplification, and innocent inclusion. *Natural Language Semantics*, 28(3), 175–223.
- Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O’Brien, K., Hallahan, E., ... others (2023). Pythia: A suite for analyzing large language models across training and scaling. In *International conference on machine learning* (pp. 2397–2430).
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hassel, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 1877–1901). Curran Associates, Inc.
- Crnič, L., Chemla, E., & Fox, D. (2015). Scalar implicatures of embedded disjunction. *Natural Language Semantics*, 23(4), 271–305. doi: 10.1007/s11050-015-9116-x
- Degano, M., Ramotowska, S., Marty, P., Aloni, M., Breheny, R., Romoli, J., & Sudo, Y. (2023). *The ups and downs of ignorance*. Retrieved from <https://ling.auf.net/lingbuzz/007389> (under review)
- Fox, D. (2007). Free choice and the theory of scalar implicatures. In *Presupposition and implicature in compositional semantics* (pp. 71–120). Springer.
- Gauthier, J., Hu, J., Wilcox, E., Qian, P., & Levy, R. (2020, July). SyntaxGym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th annual meeting of the association for computational linguistics: System demonstrations* (pp. 70–76). Online: Association for Computational Linguistics.
- Grice, P. H. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics, vol. 3, speech acts* (pp. 41–58). New York: Academic Press.
- Hu, J., Levy, R., & Schuster, S. (2022, May). Predicting scalar diversity with context-driven uncertainty over alternatives. In E. Chersoni, N. Hollenstein, C. Jacobs, Y. Oseki, L. Prévot, & E. Santus (Eds.), *Proceedings of the workshop on cognitive modeling and computational linguistics* (pp. 68–74). Dublin, Ireland: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2022.cmcl-1.8> doi: 10.18653/v1/2022.cmcl-1.8
- Jeretic, P., Warstadt, A., Bhooshan, S., & Williams, A. (2020, July). Are natural language inference models IMPPRES- sive? Learning IMPLICature and PRESupposition. In D. Jurafsky, J. Chai, N. Schlueter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 8690–8705). Online: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2020.acl-main.768> doi: 10.18653/v1/2020.acl-main.768
- Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., ... Sayed, W. E. (2023). *Mistral 7b*.
- Kamp, H. (1978). *Semantics versus pragmatics*. Springer.
- Li, E., Schuster, S., & Degen, J. (2021). Predicting scalar inferences from “or” to “not both” using neural sentence encoders. In *Proceedings of the society for computation in linguistics 2021* (pp. 446–450).
- Li, Y., Bubeck, S., Eldan, R., Del Giorno, A., Gunasekar, S., & Lee, Y. T. (2023). Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*.
- Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2023). *Lost in the middle: How language models use long contexts*.
- Marty, P., Ramotowska, S., Romoli, J., Sudo, Y., & Breheny, R. (2023). *On the source of distributive inferences*. Retrieved from <https://lingbuzz.net/lingbuzz/007623?s=6Vr1IH6Shz7nms3wk=-kuDvMqnl4piFlbG> (under review)
- Marty, P., Romoli, J., Sudo, Y., & Breheny, R. (2023). What makes an inference robust? *Journal of Semantics*. Retrieved from <https://lingbuzz.net/lingbuzz/006205?s=pEQAHAfzjamqK7zoK=607AfYta3cP962OR>
- Miles, J. (2005). R-squared, adjusted r-squared. *Encyclopedia of statistics in behavioral science*.
- MistralAI. (2023). *Mistral of experts. a high quality sparse mixture-of-experts*. Retrieved from <https://mistral.ai/news/mistral-of-experts/>
- Noveck, I. (2018). *Experimental pragmatics: The making of a cognitive science*. Cambridge University Press.
- Sauerland, U. (2004). Scalar implicatures in complex sentences. *Linguistics and philosophy*, 27, 367–391.
- Schuster, S., Chen, Y., & Degen, J. (2020, July). Harnessing the linguistic signal to predict scalar inferences. In D. Jurafsky, J. Chai, N. Schlueter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 5387–5403). Online: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2020.acl-main.479> doi: 10.18653/v1/2020.acl-main.479
- Sieker, J., & Zarrieß, S. (2023). When your language model cannot even do determiners right: Probing for anti-presuppositions and the maximize presupposition! principle. In *Proceedings of the 6th blackboxnlp workshop: Analyzing and interpreting neural networks for nlp* (pp. 180–198).
- Pravanthi, S. L., Doshi, M., Kalyan, T. P., Murthy, R., Bhat- tacharyya, P., & Dabre, R. (2024). *Pub: A pragmatics understanding benchmark for assessing llms’ pragmatics capabilities*.
- 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*.
- Truong, T. H., Baldwin, T., Verspoor, K., & Cohn, T. (2023). Language models are not naysayers: An analysis of lan-

guage models on negation benchmarks. *arXiv preprint arXiv:2306.08189*.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., ... Bowman, S. (2019). Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., & Bowman, S. R. (2020). *Blimp: The benchmark of linguistic minimal pairs for english*.