Presupposition Recognition in Chat-Optimized Language Models

Aarushi Singhal
Universität des Saarlandes
aasi00003@stud.uni-saarland.de

Abstract

Despite the impressive human-like conversational abilities of recent language models, a lot of research has pointed out issues in their linguistic comprehension and reasoning capabilities. Understanding natural language encompasses both the semantic meaning of the discourse, as well as the assumptions made during communication. Presuppositions are an integral part of this process. We evaluate the ability of ChatGPT and Llama 2 70B Chat models in predicting entailment status across 10 types of presupposition triggers. We find that these models are poor at recognising presuppositions with softer triggers, and that in general, they tend to predict entailment labels more frequently than the correct label. Some of their achievement can be attributed to flawed heuristics like lexical overlap, but it is difficult to be completely sure since we cannot isolate the effects of few-shot examples and prompting.

1 Introduction

Presuppositions are assumptions or beliefs that are considered to be true in a given context so that the speaker's utterance is meaningful and these assumptions can often be inferred by the listener using signal words called presupposition triggers. In communication, presuppositions are an essential prerequisite for understanding the content expressed by an utterance and for the coherence of the semantic relations of a discourse (Domaneschi, 2015), and therefore inferring presuppositions is likewise an important prerequisite for language models' understanding of user intents. Despite that, there have been very few systematic investigations into how well language models can infer presuppositions.

Most work on language models and presuppositions so far has focused on the ability of natural language inference (NLI) models based on pretrained language models to classify whether a statement is presupposed by another statement. Jeretic et al. (2020) generated NLI examples using templates and found that BERT-based (Devlin et al., 2019) NLI models correctly predict entailment relations between a sentence with a presupposition and the presupposed content for some, but not all, of the investigated triggers, and Asami and Sugawara (2023) replicated this finding for DeBERTa (He et al., 2020). Kabbara and Cheung (2023) recommend pre-finetuning to improve the performance of models on NLI tasks. Sieker and Zarrieß (2023) find that fine-tuned large language models (LLMs) don't follow Maximise Presuppositions! (Heim, 1991) while predicting determiners and reiterate that studying model performance on presuppositions is important.

Kim et al. (2023) found that LLMs perform poorly at answering questions with questionable assumptions but in-context learning, finetuning, and rule-based NLI hybrid approach (Kim et al., 2021) improved the performance considerably for some models. One important question that these studies raise is whether the newer chat-optimized language models encounter similar challenges while detecting presuppositions as well.

Basmov et al. (2023) have conducted experiments in zero-shot settings on linguistic phenomena that are trivial for humans such as inference. They show that ChatGPT fails to correctly establish systematic relationships, indicating an "absence of human-like text understanding". They also find that by adding a presupposition trigger to the premise, the model tends to predict entailment more strongly, regardless of the correct semantic label. Qin et al. (2023) study ChatGPT's zero-shot capabilities on NLP tasks. Among their various observations, it is evident that ChatGPT excels in handling factually consistent text, making it superior at classifying entailment rather than non-entailment. McCoy et al. (2019) hypothesise that NLI systems might score

well by adopting certain shallow heuristics. They propose an evaluation set HANS, which includes sentences where these heuristics fail, and observe that many models trained on MNLI perform poorly on HANS. Basmov et al. (2023) tested ChatGPT on a subset of HANS and observed that the model uses all three heuristics but recommended testing on more data for certainty. Plátek et al. (2023) observe that LLMs are sensitive to prompts and few-shot examples and that Llama 2 models are bridging the gap between ChatGPT and other open source models. They further show that even though few-shot examples help Llama 2 and ChatGPT generalise better to unseen data, the former does not benefit from it the same way as the latter.

Over the past few years, LLMs have undergone huge developments in terms of model parameters, training data, and performance. Newer models are fine-tuned for conversation usage and undergo reinforcement learning with human feedback which might make them better for pragmatic tasks. These models are not trained on any specific NLI task but it would be interesting to see if they can generalise well enough to understand one of the basic things that make a conversation – the implicated information during the discourse. The main contributions of this paper are:

- we look at more recent models that are predominantly used in chatbots and as the foundation of other NLP systems, as compared to the older models evaluated in NOPE (Parrish et al., 2021)
- we use the NOPE data sets which are broader and more diverse than Basmov et al. (2023), and are human verified to gauge the inference capabilities of LLMs

On a preliminary glance at Srivastava et al. (2022), it seems like the models that haven't been fine-tuned on NLI tasks do not perform particularly well on a subset of the NOPE data set. One limitation they don't take into account is that the performance might be affected by less than ideal prompts.

Our work aims to provide a better idea about the current state of inference abilities of more recent language models (without substantial fine-tuning) on naturally occurring presuppositions, especially by giving the model context before the premise as well. To test if the models use the lexical overlap heuristic further, we use the NOPE adversarial data set (Parrish et al., 2021) that minimally differs from the main data in that the hypothesis is no longer entailed in the premise. This allows us to ascertain whether the models rely on the lexical similarity to achieve high accuracy for a specific trigger type or employ a more advanced inference approach. Since previous studies have shown that ChatGPT does not particularly perform well for relatively easier presuppositions (Basmov et al., 2023) (especially as compared to adversarial sentences used in our experiment), we conduct in-context learning so as to give the models a better idea of the task they are expected to do.

2 Experiments

2.1 Data sets

- **NOPE** (Parrish et al., 2021): For 10 specific presupposition trigger types, the Corpus of Naturally-Occurring Presuppositions in English has instances of non-negated and negated sentence pairs and where possible, two previous sentences for context. The original sentences were extracted from the Corpus of Contemporary American English, and then were negated to test the extent to which the presupposition projects out of negation. The data set also has manually written purported presuppositions for each sentence. The paper tests LMs' presupposition abilities using an NLI task. In addition to the main data set, the adversarial data set is also used in our experiment.

- **MNLI** (Williams et al., 2018) contains sentence pairs that are annotated with entailment information. The examples are collected from various sources of spoken and written English text.

- **ANLI** (Nie et al., 2020) is an NLI benchmark data set that is more difficult than SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018). It was collected in English over three rounds, with different target model adversaries each round.

In the experiment we have restricted the labels the models need to give to entailed "E" and not-entailed "N" to simplify the task.

2.2 Models used

ChatGPT [1] creates seemingly human-like conversational dialogue and demonstrates remarkable capabilities, such as filtering inappropriate questions, and rectifying prior mistakes through subsequent conversations (Guo et al., 2023). Simultaneously, issues like reliability concerns (Shen et al., 2023) and the generation of contradictory or irrational responses (Zhong et al., 2023) have been documented.

Llama 2 (Touvron et al., 2023) is an open source pre-trained model released by Meta that was also fine-tuned for chat use cases. It outperforms other open-source language models on many external benchmarks. However, we do not yet know the extent of the models' language "comprehension" abilities.

2.3 Procedure

We conduct in-context learning experiments on two models: ChatGPT and Llama 2 70B Chat. For each model, the examples are taken in 2 ways: *a). NOPE + MNLI* and *b). NOPE + ANLI*. In-context learning with ANLI might help the model overcome some adversaries and help it perform better. Since most examples from NOPE corpus have presuppositions as entailed "E" by the premise, we provide some examples from MNLI or ANLI R3 that are annotated as contradictory "C" to compensate for this bias. We reserve the rest of the sentences from NOPE as test data.

GPT-3.5-turbo chat completions API endpoint from OpenAI and Llama 2 70B Chat model API from Replicate [2] are used to access the respective models. We provide a system prompt, premise and hypothesis to the models and ask them to predict whether the hypothesis is entailed "E" or not entailed "N" by the premise. We provide further details on this in Appendix A.

The models are tested on *a). NOPE Main* and *b). NOPE Adversarial* data sets. The premise and hypothesis, i.e. the purported presupposition that is triggered, often have very high lexical similarity. Giving the models hypothesis that is not entailed but still shares vocabulary with the premise allows us to distinguish whether high accuracy on a trigger type is the result of a lexical overlap heuristic or a more sophisticated inference.

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

[2] https://replicate.com/meta/llama-2-70b-chat/api?tab=python

We calculate accuracy, precision, recall and F1 score for all the cases and report them in Table 1. We also report trigger specific accuracy and 95% bootstrapped confidence interval in each case (Appendix A) and visualise them in Figure 1. Since the number of sentences per trigger type differs, we normalised the counts before computing the accuracy.

3 Interpretation of results

Looking at the metrics reported in Table 1, we observe a drop in precision in all models for the adversarial data, which indicates that the models might just be predicting the majority label, i.e. entailment "E". Furthermore, the drop in recall, which is not as much as the drop in precision, tells us that the models are using shallow heuristic to some extent, for instance lexical overlap, to predict entailment status. The accuracy for the main data set in all experiments is comparable to or better than the accuracy of ChatGPT in zero-shot setting for similar NLI prompting task by Basmov et al. (2023) (70.75% for SNLI, 78.75% and 52.5% for embedded and non embedded premise in their manually curated data set), suggesting that while the model still has a long way to go with respect to presupposition identification, its performance can improve with in-context learning and better prompting.

Similar to what was observed by Parrish et al. (2021), the models achieve lower accuracy on the adversarial data than on the main data for majority of cases, suggesting that both models exhibit a bias for predicting entailment "E" when lexical overlap is high. The confidence intervals for each model when trained by MNLI and ANLI are quite similar, hence we can say that the difference in few-shot examples does not make a lot of impact on predictions by ChatGPT and Llama 2 70B Chat. This is in contrast to fine-tuned models in Parrish et al. (2021) where RoBERTa-L had approximately 90% accuracy when validated on MNLI and 49.2% on ANLI R3, and DeBERTa-XL had approximately 91% accuracy when validated on MNLI and 60.5% on ANLI R3.

Llama 2 70B Chat performs quite comparable to ChatGPT and shows better results on adversarial data for some trigger types. This could mean that it employs certain advanced heuristics. Since we know a lot more about its weights and pre-training process, further studies to improve conversational LLMs' NLI abilities can have more transparency.

Few-shot Examples	Model	NOPE Main				NOPE Adversarial			
		Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1
NOPE + MNLI	ChatGPT	79.6	86.3	88.7	87.5	64.5	23.3	60.8	33.7
	Llama 2 70B Chat	68.7	85.4	73.9	79.2	66.5	21.1	43.4	28.4
NOPE + ANLI	ChatGPT	79.0	85.8	88.6	87.2	59.3	21.3	64.7	32.0
	Llama 2 70B Chat	73.8	85.6	81.2	83.3	63.3	19.7	45.3	27.4

Table 1: Aggregate results on the NOPE Main and NOPE Adversarial data sets.

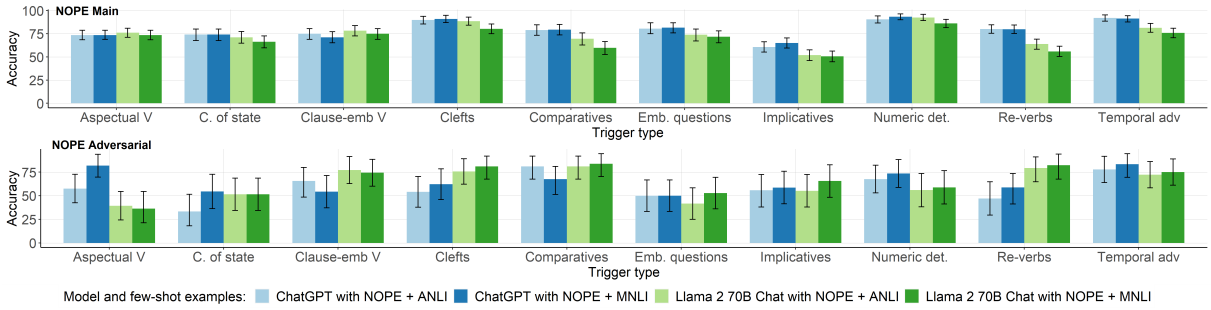

Figure 1: Accuracy and 95% CI for each trigger type.

From looking at the model output data, we observe that their accuracy is better for "hard" triggers for which is presupposition is more difficult to be suspended (e.g. numerical and temporal presuppositions, as well as *it*-clefts). This echoes the findings in Parrish et al. (2021) and suggests that language models in general find it more difficult to detect "soft" triggers which are also more context dependent and commonsense based. In a lot of sentences, especially re-verbs, when the trigger sentence is negated, the entailment status changes to neutral "N", but ChatGPT is ubable to recognise that and still infers entailment relation "E" between the premise and the hypothesis. This might mean that the model is poor at detecting when presupposition projects out of negation and when it does not.

4 Conclusion

We evaluated ChatGPT and Llama 2 70B Chat on predicting the entailment status between the premise and the hypothesis based on 10 different presupposition trigger types for NOPE main and adversarial data sets. We find that the models predict entailment "E" much more often, even when that is not the gold label. This is in line with the findings of Basmov et al. (2023). Furthermore, we also find that the higher accuracy of most presupposition triggers is, to some extent, a result of error-prone heuristic like lexical overlap. We conclude that even though recent LLMs are far from achieving an *"understanding"* of natural language presupposition tasks, we cannot definitively say that the models only use fallible heuristics to accomplish the given NLI task. It is possible that the models are bad at inferring what the task is from the instructions and in-context learning examples. It could also be that many of the in-context learning examples can be solved through a lexical overlap heuristic, and therefore, the models apply the same heuristic for the NOPE test data. In a more naturalistic conversation, on the other hand, it may be that the model is able to draw the right inferences.

Limitations

Our work has some limitations which could be resolved by further studies. In our experiments, we only test English presuppositions. The model performances might vary if we use different languages or test code-switched sentences. We are considering just two models and it remains an open question whether the results would generalise to other conversational models. If the models are tested in a more naturalistic setting with conversational data for other NLI tasks, then they may perform better. We only run the tasks with one prompt, and since LMs are sensitive to the prompts they are given (Plátek et al., 2023), they may perform differently when tested with some variations. We use API endpoints in our experiments which are subject to

change over time without notice. During in-context learning, we give examples from ANLI R3. Since each round captures different target model adversaries, sentences from other rounds might have made the models perform differently.

References

Daiki Asami and Saku Sugawara. 2023. PROPRES: Investigating the projectivity of presupposition with various triggers and environments. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 122–137, Singapore. Association for Computational Linguistics.

Victoria Basmov, Yoav Goldberg, and Reut Tsarfaty. 2023. Chatgpt and simple linguistic inferences: Blind spots and blinds.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Filippo Domaneschi. 2015. Introduction: Presuppositions philosophy, linguistics and psychology. *Topoi*, 35(1):5–8.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention.

Irene Heim. 1991. *Artikel und Definitheit*, pages 487–535. De Gruyter Mouton, Berlin • New York.

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*. Association for Computational Linguistics.

Jad Kabbara and Jackie Cheung. 2023. Investigating the effect of pre-finetuning BERT models on NLI involving presuppositions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10482–10494, Singapore. Association for Computational Linguistics.

Najoung Kim, Phu Mon Htut, Samuel R. Bowman, and Jackson Petty. 2023. (qa)2: Question answering with questionable assumptions: Question answering with questionable assumptions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Najoung Kim, Ellie Pavlick, Burcu Karagol Ayan, and Deepak Ramachandran. 2021. Which linguist invented the lightbulb? presupposition verification for question-answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics.

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

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

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. pages 349–366.

Ondřej Plátek, Vojtěch Hudeček, Patrícia Schmidtová, Mateusz Lango, and Ondřej Dušek. 2023. Three ways of using large language models to evaluate chat. *arXiv (Cornell University)*.

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

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

Judith Sieker and Sina Zarrieß. 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*, pages 180–198, Singapore. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya

Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto,

Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.

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.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. pages 1112–1122.

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.

A Appendix

Statistics for the data: We report the trigger specific accuracies and 95% CI for all experiments in Tables 2, 3, 4 and 5. NOPE Main data set consists of 2386 sentences while the Adversarial data set contains 346 sentences. We use 6 sentences from Main data set (including one common sentence from Adversarial data) for in-context learning, and the rest as test data. We conduct single runs for all the experiments.

Few shot examples: We give the model eleven examples to illustrate how the input messages are structured and how the model should react. Six of the examples are taken from NOPE and the other five are taken from either MNLI or ANLI. The first three pairs of sentences from NOPE are intended to help the model understand that in successive iterations, it would be given a sentence and its negated version. Table 6 shows some examples of the same.

Giving Prompts to Models: The users can communicate with the model by structuring conversations in the form of messages. The system message acts as the prompt template. Table 7 shows one iteration of conversation with the model.

- **ChatGPT API:** The messages consist of roles ("system," "user," or "assistant") and content. The system message is optional and is used to influence the assistant's behaviour. The user messages contain requests or comments, which in our case are used to send the model the premise (sentences along with context) and hypothesis (purported presupposition) from the NOPE corpus. The assistant messages can store prior responses or be used to demonstrate desired behaviour.

- **Replicate Llama 2 70B Chat API:** The concept of communicating with the model is very similar to ChatGPT. The only difference is that the API needs [INST][\INST] tokens to recognise that the message is given by the user and not generated by the model.

Model Parameters: The parameters for Llama 2 70B Chat (the ones we can modify through Replicate API) are mentioned in Table 8. We used the default values in our experiments. The endpoint we used for accessing ChatGPT (ChatCompletion) has changed and hence, we cannot report the parameters of the same.

Trigger Type	Main	95% CI	Adversarial	95% CI
Numeric Determiners	93.277	[89.916, 96.218]	73.529	[58.824, 88.235]
Temporal Adverbs	91.270	[87.698, 94.444]	83.333	[69.444, 94.444]
clefts	90.821	[86.957, 94.686]	62.162	[45.946, 78.378]
Embedded Question	81.538	[75.897, 86.667]	50.000	[33.333, 66.667]
Re-Verbs	79.739	[75.163, 84.314]	58.824	[41.176, 73.529]
Comparatives	79.381	[73.711, 85.052]	67.568	[51.351, 81.081]
Change of State	74.020	[67.647, 79.902]	54.545	[36.364, 72.727]
Aspectual Verbs	73.529	[68.832, 78.676]	81.818	[69.697, 93.939]
Clause Embedding Predicates	71.163	[65.116, 77.209]	54.286	[37.143, 71.429]
Implicative Predicates	64.983	[59.596, 70.370]	58.621	[41.379, 75.862]

Table 2: Accuracies for ChatGPT (with NOPE + MNLI)

Trigger Type	Main	95% CI	Adversarial	95% CI
Numeric Determiners	90.336	[86.555, 94.118]	67.647	[52.941, 82.353]
Temporal Adverbs	92.063	[88.492, 95.238]	77.778	[63.889, 91.667]
Clefts	89.855	[85.507, 93.720]	54.054	[37.838, 70.270]
Embedded Question	80.513	[74.872, 86.641]	50.000	[33.333, 66.667]
Re-Verbs	80.065	[75.490, 84.641]	47.059	[29.412, 64.706]
Comparatives	78.866	[73.196, 84.536]	81.081	[67.568, 91.892]
Change of State	74.020	[67.647, 79.902]	33.333	[18.182, 51.515]
Aspectual Verbs	73.529	[68.382, 78.676]	57.576	[42.424, 72.727]
Clause Embedding Predicates	74.884	[68.837, 80.465]	65.714	[48.571, 80.000]
Implicative Predicates	60.606	[55.219, 66.330]	55.712	[37.931, 72.414]

Table 3: Accuracies for ChatGPT (with NOPE + ANLI)

Trigger Type	Main	95% CI	Adversarial	95% CI
Numeric Determiners	86.134	[81.513, 90.336]	58.824	[41.176, 76.471]
Temporal Adverbs	75.794	[70.635, 80.952]	75.000	[61.111, 88.889]
Clefts	80.193	[74.879, 85.507]	81.081	[67.568, 91.892]
Embedded Question	71.795	[65.128, 77.949]	52.778	[36.111, 69.444]
Re-Verbs	55.921	[50.329, 61.513]	82.353	[67.647, 94.118]
Comparatives	59.794	[52.577, 66.495]	83.784	[70.270, 94.594]
Change of State	66.176	[59.804, 72.549]	51.429	[34.286, 68.571]
Aspectual Verbs	73.529	[68.382, 78.676]	36.364	[21.212, 54.545]
Clause Embedding Predicates	74.884	[68.837, 80.465]	74.286	[60.000, 88.571]
Implicative Predicates	50.505	[44.781, 56.229]	65.517	[48.276, 82.759]

Table 4: Accuracies for Llama 2 70B Chat (with NOPE + MNLI)

Trigger Type	**Main**	**95% CI**	**Adversarial**	**95% CI**
Numeric Determiners	92.437	[89.076, 95.798]	55.882	[38.235, 73.529]
Temporal Adverbs	81.349	[76.587, 86.111]	72.222	[58.333, 86.111]
Clefts	88.406	[84.058, 92.754]	75.676	[62.162, 89.189]
Embedded Question	73.846	[67.179, 80.000]	41.667	[25.000, 58.333]
Re-Verbs	63.816	[58.224, 69.079]	79.412	[64.706, 91.176]
Comparatives	69.588	[62.887, 75.773]	81.081	[67.568, 91.892]
Change of State	71.078	[64.706, 77.451]	51.429	[34.286, 68.571]
Aspectual Verbs	76.102	[70.956, 80.882]	39.393	[24.242, 54.545]
Clause Embedding Predicates	78.140	[72.558, 83.721]	77.143	[62.857, 91.429]
Implicative Predicates	51.851	[46.128, 57.576]	55.172	[37.931, 72.414]

Table 5: Accuracies for Llama 2 70B Chat (with NOPE + ANLI)

Premise	Hypothesis	Label
For three nights a comet flared through the desert sky. The winds hooted like owls. A red smudge didn't appear on the moon.	A red smudge couldn't be seen on the moon before.	E
For three nights a comet flared through the desert sky. The winds hooted like owls. A red smudge appeared on the moon.	A red smudge couldn't be seen on the moon before.	E
Fun for adults and children.	Fun for only children.	C

Table 6: In-context learning examples given to models

Role	Content
System	Assume the statement after "Premise" is true. If the statement after "Hypothesis" is also true, output "E", otherwise output "N".
User	Premise: Vrenna and I both fought him and he nearly took us. Hypothesis: Neither Vrenna nor myself have ever fought him.
Assistant	C

Table 7: Messages given to models based on role

Parameter	Default
max_new_tokens	128
min_new_tokens	-1
temperature	0.75
top_p	0.9
top_k	50
seed	randomised
debug	–

Table 8: Default parameters for Llama 2 70B Chat