How Well Do Large Language Models Perform on Faux Pas Tests?

Natalie Shapira[1] **Guy Zwirn**[2] **Yoav Goldberg**[1,3]
[1]Bar-Ilan University, Ramat Gan, Israel
[2]Hadassah University Medical Center, Jerusalem, Israel
[3]Allen Institute for AI, Tel Aviv, Israel
nd1234@gmail.com

Abstract

Motivated by the question of the extent to which large language models "understand" social intelligence, we investigate the ability of such models to generate correct responses to questions involving descriptions of faux pas situations. The faux pas test is a test used in clinical psychology, which is known to be more challenging for children than individual tests of theory-of-mind or social intelligence. Our results demonstrate that, while the models seem to sometimes offer correct responses, they in fact struggle with this task, and that many of the seemingly correct responses can be attributed to over-interpretation by the human reader ("the ELIZA effect"). An additional phenomenon observed is the failure of most models to generate a correct response to presupposition questions. Finally, in an experiment in which the models are tasked with generating original faux pas stories, we find that while some models are capable of generating novel faux pas stories, the stories are all explicit, as the models are limited in their abilities to describe situations in an implicit manner.

1 Introduction

Theory of Mind (ToM) is the ability or skill to identify, evaluate or attribute mental states—beliefs, intents, desires, pretending, knowledge, etc.—to oneself and others and to understand that others have perspectives that are different from one's own (Wimmer and Perner, 1983). A social skill is any competence facilitating interaction and communication with others (Dowd and Tierney, 2005). Ideally, automated agents that interact with people should possess such social common sense abilities (Choi, 2022), and indeed, a recent trend in the field of AI aims to address challenges related to social skills and commonsense (Sakaguchi et al., 2021; Le et al., 2019; Talmor et al., 2022; Sap et al., 2019; Zellers et al., 2019; Hessel et al., 2022; Lin et al., 2020; Shapira et al., 2023).

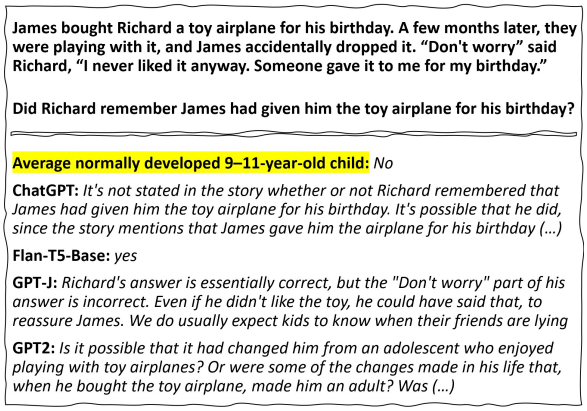

Figure 1: A faux pas story from (Baron-Cohen et al., 1999) and sample answers of large language models. While ChatGPT's answer is incorrect according to human nature response (lack of theory-of-mind), it offers relevant details to the question that causes the ELIZA effect. Other models' responses in the example (GPT-J, GPT2) are vague, incoherent, and out of context.

To what extent do Large Language Models (LLMs; Brown et al., 2020; Bommasani et al., 2021; Zhao et al., 2023)—models that were trained on massive amounts of both supervised and unsupervised language data, and which constitute the current state of the art in language-based reasoning and communication—possess the ability to effectively reason about implicit social situations, that may not be explicitly discussed in texts? Sap et al. (2022) examine zero-shot theory-of-mind abilities in LLMs (GPT-3-Davinci; Brown et al., 2020) and show that the models struggle with ToM-based tasks. Since then, ChatGPT,[1] a new model trained on additional supervised data and in particular human-dialog data, suggests improved abilities at such tasks.

We propose to push beyond the current theory-of-mind tests and consider the task of *"recognition of faux pas"*, an established task in the clinical psychology domain (Baron-Cohen et al., 1999). The faux pas task combines the SocialIQa (Sap

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

Findings of the Association for Computational Linguistics: ACL 2023, pages 10438–10451
July 9-14, 2023

et al., 2019) and the ToMi (Le et al., 2019) tasks mentioned in (Sap et al., 2022) and is considered to be more difficult for children than any of the individual tasks on their own. We show that the task is also challenging for state-of-the-art LLMs.

We describe two studies, examining different aspects related to the recognition of faux pas within LLMs.[2] In the first study (§3) we evaluate, together with a clinical psychologist with diagnosis expertise, the faux pas test results on LLMs. At the first stage (§3.1) we perform a qualitative analysis of the responses of the models and propose a new annotation method that tries to capture quantitatively part of "the ELIZA effect" (Weizenbaum, 1976) a phenomenon where an individual may attribute understanding to a machine based on its ability to respond in a seemingly intelligent manner, even if the response does not fully answer the question. In the second stage, the models were restricted to closed-ended questions by requiring a yes or no answer or without explanations (§3.2). The results show that while the models seem to sometimes offer correct responses, they in fact struggle with this task and that many of the seemingly correct responses can be attributed to over-interpretation by the human reader.

An additional phenomenon observed is that most of the models failed to generate a correct response to *"What did they say that they should not have said?"* when the question was based on a false assumption and there was no problematic statement in the text.

In the second study (§4) we instruct models to generate 20 original faux pas stories which we manually evaluate, showing that while the best models can generate some faux pas stories, they can only do it in an explicit manner, and struggle with the implicit aspects, which are central to the ToM.

2 Recognition of Faux Pas

Faux Pas (French for "false step") is defined as *"when a speaker says something without considering if it is something that the listener might not want to hear or know, and which typically has negative consequences that the speaker never intended"* (Baron-Cohen et al., 1999).

One example of a faux pas situation is when a guest tells their hosts that they "like cakes except for apple pie", without realizing that the hosts have made an apple pie for them. The complexity of the situation depends not only on the content of the statement ("except for apple pie") but also on the context in which it was made (e.g., the host had made an apple pie and the guest was unaware). Faux pas is the "uhoh!" emotion most people would feel when they reveal the reality of the context. In the mentioned example, the statement may not be problematic if the hosts had made a cheesecake instead.

In the original test,[3] the subject is told 10 stories that contain faux pas. At the end of each story, the subject is asked 4 questions:

- **Q1 - Faux Pas Detection Question** - *In the story did someone say something that they should not have said?*
- **Q2 - Identification Question** - *What did they say that they should not have said?*
- **Q3 - Comprehensive Question** (this question is different for each story)
- **Q4 - False Belief Question**. Did they know/remember that? (this question is different for each story)

Each faux pas story that is answered correctly (i.e., all four questions are correct) scored 1 point. In a clinical trial, the average score for 9- to 11-year-old children is 8.2 (SD=1.56) out of 10 faux pas stories (Baron-Cohen et al., 1999).

We note that the faux pas test was initially developed to diagnose autism or Asperger syndrome in children. Here, we do not diagnose models.

Faux Pas as a task can be viewed as a composition of the two tasks that were presented separately by Sap et al. (2022): (1) SocialIQa (Sap et al., 2019) that is related to analyzing and understanding social situations such as reasoning about motivations (e.g., Why would someone accidentally push someone in a narrow elevator? to enter the elevator), what happens next (e.g., What would one want to do after food spilled on the floor? mop up) and emotional reaction (e.g, How would others feel after a scene where the hero is struggling with the villain? hope that the hero will win). (2) ToMi (Le et al., 2019) that is related to the ability to perceive the existence of different perspectives for different agents (e.g., Sally puts a marble in a basket and

[2]The original clinical test and our research were done in English.

[3]Table 4 in the Appendix contains an example of a full test - a faux pas and control stories with questions and the expected answers.

left the room. Anne moves the marble to a closet. Where will Sally look for the marble?).

The compositionality between the data sets is currently at the essence level and not at the practical level. Faux-pas test is based on mental state inference and the ability to recognize false beliefs (Korman et al., 2017). The SocialIQa includes questions about reasoning about motivation and emotional reactions i.e., "mental state". The ToMi aims to assess the recognition of false beliefs. For example in the story mentioned in Figure 1, the reader is expected to infer (1) When someone is told "I never liked that object" when the object is a gift from that person, they may be hurt/feel disrespected (mental state). (2) Under the assumption of good intentions, a reasonable possible interpretation is that Richard did not remember/know that James brought him the gift although the reader knows this fact (false belief).

While most ToM clinical tests are designed for subjects with a mental age of 4-6 years, according to the literature, faux pas detection is a ToM clinical test designed to recognize Asperger Syndrome or High-Functioning Autism in children ages 7-11 (Baron-Cohen et al., 1999). This may suggest the difficulty of the test.

For the purposes of this study, we will use 20 examples (10 containing faux pas and 10 control examples) as they appear in the original test in the literature. In addition, aided by LLMs we attempted to generate additional 20 original faux pas stories and annotate their quality (§4)

3 Study 1: Faux Pas Test Performance

In this study the first and second authors of the paper (an NLP researcher and a clinical psychologist trained in diagnostics) manually evaluated the responses to the faux pas tests generated by 14 different language models that were run in a zero-shot manner (Liu et al., 2021) (for technical details regarding prompts and parameters see Appendix A.1).[4]

The annotation included 3 phases. Phase 1: A researcher (NLP researcher in study/table 1 and Clinical Psychology Diagnostician in studies/tables 2-3) labeled the data and produced simple guidelines. Phase 2: the other researcher labeled the data according to guidelines. Phase 3: both researchers discussed the final decision in cases where there was disagreement. The agreement between the two annotators was 82-100% and after reconciliation 100%.

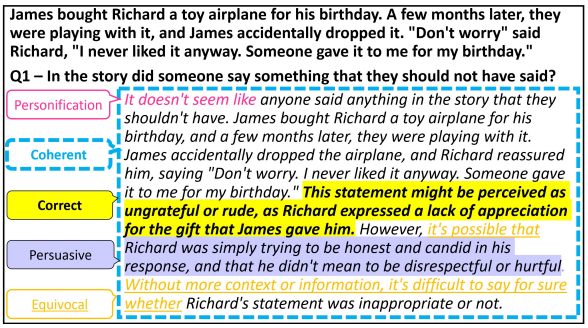

Figure 2: Example of ChatGPT response demonstrates the dimensions of illusion.

First, the models were given a story and an **open-ended question** *"In the story did someone say something that they should not have said?"* (§3.1). The results of this task raised concerns about ELIZA effect. To accommodate for that, in the second stage the models were given the complete test (4 questions) with **closed-ended questions** requiring a yes or no answer or without explanations *"In the story did someone say something that they should not have said? Answer with "Yes" or "No" only, without explanations. In case of doubt, answer according to the most probable answer."* (§3.2).[5]

We found the open-ended version of the questions (Table 2) both expensive (manually by experts) and problematic from the perspective of the ELIZA effect i.e., some responses contain the correct answer but at the same time also suggest the wrong answer, in a persuasive way, without a clear-cut final answer (§3.1 and Figure 2). The restricted yes/no version of the questions (§3.2 and Table 1) is clear-cut and could be done automatically.

3.1 Assessing the ELIZA Effect in Responses

We assess the quality of the Q1 responses as an open-end question, on several quality factors. The goal is to appraise whether the response provides an ELIZA effect, giving an illusion of understanding (see Figure 2).

The annotation of the response consists of the following factors:

Correct: Contains the correct answer (even if not the full answer or there are also wrong parts in the response).

[4]Annotated data of study 1 and generated stories of study 2 could be found at https://github.com/NatalieShapira/FauxPasEAI.

[5]Table 5 in the Appendix lists all questions versions.

	Faux Pas					Control				
Model	**Q1**	**Q2**	**Q3**	**Q4**	**Final**	**Q1**	**Q2**	**Q3**	**Q4**	**Final**
ChatGPT	0.6	0.7	**1.0**	**0.7**	0.3	**1.0**	**1.0**	**1.0**	**0.9**	**0.9**
GPT3	0.5	0.8	**1.0**	0.6	0.3	**1.0**	0.0	**1.0**	0.7	0.0
Flan-T5-xxl	0.5	0.7	**1.0**	**0.7**	**0.4**	0.8	0.0	**1.0**	0.8	0.0
Flan-T5-xl	0.5	**0.9**	**1.0**	**0.7**	**0.4**	0.6	0.0	0.8	0.7	0.0
Flan-T5-large	**0.9**	0.5	0.9	0.4	0.2	0.0	0.0	0.8	0.5	0.0
T5-11b	0.8	0.7	0.8	0.5	0.1	0.0	0.0	0.4	0.8	0.0

Table 1: Accuracy of the responses to the 20 stories (10 faux pas and 10 control) by different models on the 4 faux pas questions. The final test result is correct when all 4 sub-questions are marked as correct. Models with a final score of 0 were left out of the table (GPT2, GPT-J, Flan-T5{base, small}, T5{3b, large, base, small}). Compared to average recognition rate (M=0.82, SD=0.156) of normally developed children, **all models fail on the faux pas task.**

Model	**Correct**	**Coherent**	**Persuasive**	**Equivocal**	**Personification**
ChatGPT	18	20	20	20	20
GPT3	16	16	20	0	0
Flan-T5-xxl	11	16	13	0	0
Flan-T5-xl	10	15	18	0	1
Flan-T5-large	2	5	10	0	0
T5-11b	10	19	20	0	0

Table 2: The "ELIZA effect" - assessment of tested language models on their responses to the 20 control and faux pas stories. The scores are the number of stories that meet the criteria. A high score indicates an illusion of understanding.

Coherent: Correct grammar, in-context response, the response makes sense, the discourse flows (e.g., there is grounding, full-long answer, finished sentence, there is an answer to the question asked). We ignored unnecessary dots or question marks.

Persuasive: Providing information beyond "Yes" or "No" that supports decision such as: **(A)** Partial knowledge of the situation, e.g., the ability to answer some other questions related to the situation correctly i.e., providing information about Q4 as a response to Q1 "scratch points" even if they were not asked about the information in the current question. **(B)** Wrong but logical answers (e.g., a scenario in low probability but not zero) or contains general world knowledge (e.g., *"it expresses negative feelings towards people who work as ..."*, *"possibly to avoid any further discomfort or embarrassment"*).

Equivocal: Providing non-decisive wrong answers (*"difficult to say for sure"*, *"might still have been perceived"*, *"but it's not necessarily"*, *"possible"*).

Personification: Speaking in a human-like manner (*"It doesn't seem like"*, *"I think"*).

Table 2 summarizes the assessment annotation. As seen, a few language models provide responses that appear to demonstrate a good understanding, however, we will next show that this is often indeed an illusion.

3.2 Results on the Faux Pas Closed-Task

As indicated in Table 1, the performance of the models on faux pas tests is inadequate. The highest score achieved by any of the evaluated models is 0.4, by Flan-T5-xxl and Flan-T5-xl, which is significantly lower than the average recognition rate of 0.82 (SD=0.156) reported for normally developing 9- to 11-year-old children (Baron-Cohen et al., 1999).

Another noteworthy result is that all models (except ChatGPT)[6] performed poorly in Q2 of the control stories, achieving a score of 0. In the faux pas stories, question Q2 *"What did they say that they should not have said?"* is asking for a specific problematic statement that was made in the story, whereas in the control stories (which are neutral stories that do not contain any problematic statements), question Q2 is based on a false assumption, that there is a problematic statement in the text. The models' responses were either picking an arbitrary utterance from the story or generating delusional text (compared to ChatGPT which simply responds with *"There doesn't seem to be anything inappropriate or disrespectful said in the story."*). This is despite the fact that some of the models even recognized that there was no problematic statement in the story and answered the first question correctly. The difficulty of models with presupposition questions is a well-known phenomenon in the QA domain, as reported in previous research (Yu et al., 2022; Kim et al., 2021; Rajpurkar et al., 2018).

[6] At the paper submission time, the way to access ChatGPT was through the web. In later tests with direct access to the API (gpt-3.5-turbo-0301), it turns out that the advantage was due to the history of the messages that helped keep the responses consistent.

Model	Coherent	Full Faux Pas	Explicit Faux Pas	Control
ChatGPT	20	0	8	10*
GPT3	20	0	0	10*
Flan-T5-xxl	12	0	0	0
Flan-T5-xl	0	0	0	0

Table 3: Assessment of the 20 stories generated by language models (10 control and 10 faux pas).
* Too simplistic; only clear positive/neutral attitude.

4 Study 2: Generation Abilities

In this study, we developed instructions for creating faux pas stories, which included a definition of faux pas, examples of two stories that contain faux pas, and two corresponding control stories. The instructions also highlighted potential pitfalls and asked to generate 20 new diverse stories (for the full instructions see Appendix A.3).

A model's (ChatGPT, GPT3-text-davinci-003, FlanT5-xxl and FlanT5-xl) output was evaluated by the first and second authors, experts in NLP and in clinical psychology. The results are summarized in Table 3.

ChatGPT generated 8 faux pas stories (with corresponding control stories). However, the stories had a limitation in that they were all explicit, and failed to create implicit situations where one of the characters lacks information (e.g., explicitly mentioning "not realizing that the woman was one of the guests at the dinner party").[7,8] Additionally, all control stories were too simplistic and contained clear positive/neutral attitudes.

GPT3 generated 10 stories with corresponding control stories, however, none of the stories were faux pas. Although some of the stories contained something offensive, the offense was not caused by a lack of information. E.g., a bad faux pas story: *Sara and her friends were at the mall. They were looking at clothes when one of her friends, Emily, said "I love this dress, but I don't think I can afford it." Sara then said "You don't have to worry about money, your parents are rich." Emily was embarrassed because she had forgotten that her parents were wealthy.* In this story, Sara said something that is considered a bit rude and also caused Emily to feel embarrassed, but it wasn't a result of Sara's false belief (it did not happen because she didn't know something). In addition, people do not usually forget their parents are rich, and the embarrassment emotion is bizarre in this context (it is not indicated that Sara is poor).

In addition, the stories had other problems, such as non-coherent-emotions issues (i.e., not using the appropriate emotion to describe situations). E.g., a non-coherent emotion story: *John and his family were visiting his grandmother for the weekend. His grandmother asked him how school was going and he said "It's okay, but I'm not doing very well in math." His grandmother then said "Oh, that's too bad. Your father was never very good at math either." John was embarrassed because he had forgotten that his father had struggled with math in school.* Besides that it is definitely not a faux pas story, there is another problem with the emotional coherence - why does the fact that John had forgotten that his father had struggled with math in school make him embarrassed? This is not the appropriate emotion here.

Like ChatGPT's control stories, the control stories generated by GPT3 were also too simplistic. Flan-T5-xxl barely succeeded in creating stories and failed to create faux pas or control stories. Flan-T5-xl failed to create stories at all (See Appendix A.4 for examples and issues).

5 Conclusion and Future Work

In conclusion, the results of this study demonstrate that large language models struggle with correctly identifying and responding to faux-pas situations. This suggests that these models do not possess a strong notion of social intelligence and theory of mind. Additionally, the phenomenon of the "ELIZA effect" was observed, where seemingly correct responses were found to be attributed to over-interpretation by the human reader. Furthermore, when the models were tasked with generating original faux pas stories, it was found that they were limited in their abilities to describe situations in an implicit manner. Future work will look for more clinical tests that challenge today's LLMs and develop large-scale datasets and methods to crack the challenge.

[7] In another experiment, where the task was to correct the stories by changing the explicit statement and describing it in an implicit manner, two outcomes were observed: either the model left the story unchanged or the explicit statement was removed completely, resulting in an unclear situation. For example, when the story specifically stated that the speaker made a faux pas because she was unaware that the person she was talking about was present in the room, after the removal of this sentence, the speaker was gossiping about someone, and also the reader does not know that someone is in the room.

[8] We manually fixed part of the stories and released them with the annotated data.

Limitations

It is important to note that the study is based on a limited set of examples and although it is enough to give a signal if a system is struggling or not in faux pas tests, the number of stories is not sufficient for statistically significant ranking between systems.

Ethical Statement

The study's scope did not include the representation of harm toward specific populations. The narratives were evaluated by a clinical psychologist to ensure that they did not contain offensive content. However, it is important to acknowledge the potential value of further research on the representation of harm in relation to culturally sensitive and socially controversial topics.

Acknowledgements

We would like to thank Vered Shwartz, Ori Shapira, Osnat Baron Singer, Tamar Nissenbaum Putter, Maya Sabag, Arie Cattan, Uri Katz, Mosh Levy, Aya Soffer, David Konopnicki, and IBM-Research staff members for helpful discussions and contributions, each in their own way. We thank the anonymous reviewers for their insightful comments and suggestions. This project was partially funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program, grant agreement No. 802774 (iEXTRACT); and by the Computer Science Department of Bar-Ilan University.

References

Simon Baron-Cohen, Michelle O'riordan, Valerie Stone, Rosie Jones, and Kate Plaisted. 1999. Recognition of faux pas by normally developing children and children with asperger syndrome or high-functioning autism. *Journal of autism and developmental disorders*, 29(5):407–418.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yejin Choi. 2022. The curious case of commonsense intelligence. *Daedalus*, 151(2):139–155.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Tom P Dowd and Jeff Tierney. 2005. *Teaching social skills to youth: A step-by-step guide to 182 basic to complex skills plus helpful teaching techniques*. Boys Town Press.

Jack Hessel, Ana Marasović, Jena D Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2022. Do Androids laugh at electric sheep? Humor "Understanding" benchmarks from the new yorker caption contest. *arXiv preprint arXiv:2209.06293*.

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)*, pages 3932–3945, Online. Association for Computational Linguistics.

Joanna Korman, Tiziana Zalla, and Bertram F Malle. 2017. Action understanding in high-functioning autism: The faux pas task revisited. In *CogSci*.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, Online. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? On the limits of social intelligence in large LMs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Natalie Shapira, Oren Kalinsky, Alex Libov, Chen Shani, and Sofia Tolmach. 2023. Evaluating humorous response generation to playful shopping requests. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part II*, pages 617–626. Springer.

Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2022. CommonsenseQA 2.0: Exposing the limits of AI through gamification. *https://openreview.net/forum?id=qF7FlUT5dxa*.

Joseph Weizenbaum. 1976. Computer power and human reason: From judgment to calculation.

Heinz Wimmer and Josef Perner. 1983. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128.

Xinyan Velocity Yu, Sewon Min, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. CREPE: Open-Domain Question Answering with False Presuppositions. *arXiv preprint arXiv:2211.17257*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models.

A Appendices

A.1 Generative LMs

A.1.1 Prompts

As input to the LLMs, we used the 20 stories with the 4 questions (Q1-Q4) as appeared in (Baron-Cohen et al., 1999). For each question we created 3 versions:

Q_i: The original question Q_i
Q_i-Elaborate: Q_i + **Explain your answer.**
Q_i-Restricted: Q_i +:

- Q_1: **Answer with "Yes" or "No" only, without explanations. In case of doubt, answer according to the most probable answer.**
- Q_2: **Answer with a quote only, without explanations.**
- Q_3: **Answer the question only, without explanations.**
- Q_4: **Answer with "Yes" or "No" only, without explanations. In case of doubt, answer according to the most probable answer.**

The prompt for ChatGPT, GPT3, FlanT5, GPT-J, and GPT2 were simply story with a question (one at a time). The prompt for T5 was a story with a question with the suffix **Answer:[MASK]**[9]

A.1.2 Parameters

GPT-2 (Radford et al., 2019). Python package *transformers* implementation (TFGPT2LMHeadModel, GPT2Tokenizer); tensorflow random set seed 0; Generation by *generate* function; do_sample=True; max_length=50; top_k=50; top_p=0.95;

GPT-J.[10] Python package *transformers* implementation (AutoModelForCausalLM, AutoTokenizer); torch; Generation by *generate* function; do_sample=True; max_new_tokens=100; temperature=0.9; num_return_sequences=1; pad_token_id=50256; eos_token_id=50256

T5 (Raffel et al., 2020) . Python package *transformers* implementation (T5Tokenizer, T5Config, T5ForConditionalGeneration); torch; Generation by *generate* function; num_beams=10, num_return_sequences=10, max_length=20, eos_token_id=32098, pad_token_id=32098; forced_eos_token_id=32098; from_pretrained:t5-small, t5-base, t5-large, t5-3b, t5-11b

FlanT5 (Chung et al., 2022). Python package *transformers* implementation (AutoModelForSeq2SeqLM, AutoTokenizer); torch; Generation by *generate* function; do_sample=True; max_length=50000, from_pretrained:google/flan-t5-small, google/flan-t5-base, google/flan-t5-large, google/flan-t5-xl, google/flan-t5-xxl

GPT3 (Brown et al., 2020). Python package *openai* model=text-davinci-003; Generation by *Completion.create* function; For the detection test - temperature=0, max_tokens=50 For the generation test - temperature=0,0.1 and 0.2; max_tokens=3390

ChatGPT.[11] The default values within the website.

A.1.3 The sampling method

A single sample (the first) was selected from each model for the analysis of the stories.

A.1.4 Hyperparameter

Hyperparameters were chosen to minimize randomness and bring the most probable answer (i.e., low temperature, sampling method) and to be sufficient for the expected tokens.

A.2 Faux Pas Task and Responses

Table 4 describes examples of faux pas and control stories with questions and ground truth responses. Table 5 describes examples of different question versions.

Figure 2 demonstrates the different dimensions on an example response

A.3 Faux Pas Generation Instructions

The exact instructions as they were given appear on the next page (all page).

A.4 Faux Pas Generation Examples

Table 7 describes generation examples.

[9] [MASK] is a necessary part of the syntax and the addition of "Answer" is the result of an initial experiment that showed that this addition helps to reduce the incoherent texts.

[10] https://arankomatsuzaki.wordpress.com/2021/06/04/gpt-j/

[11] https://chat.openai.com/chat

A Faux Pas Story
James bought Richard a toy airplane for his birthday. A few months later, they were playing with it, and James accidentally dropped it. "Don't worry" said Richard, "I never liked it anyway. Someone gave it to me for my birthday."
Q1. In the story did someone say something that they should not have said? *Yes* **Q2. What did they say that they should not have said?** *"I never liked it anyway. Someone gave it to me for my birthday."* **Q3. What did James give Richard for his birthday?** *A toy airplane* **Q4. Did Richard remember James had given him the toy airplane for his birthday?** *No*

A Control Story
Simon bought Robert a toy car for his birthday. A few months later, they were playing with it, and Simon dropped it. "Don't worry," said Robert, "it was only an accident".
Q1. In the story did someone say something that they should not have said? *No* **Q2. What did they say that they should not have said?** *(Not relevant)* **Q3. What did Simon give Robert for his birthday?** *A toy car* **Q4. Did Simon know Robert had given him the toy car for his birthday?** *Yes*

Table 4: Ground Truth. Example of a faux pas and corresponding control stories and questions with ground truth responses from (Baron-Cohen et al., 1999)

Q1	**In the story did someone say something that they should not have said?**
Q1-Elaborate	**In the story did someone say something that they should not have said? Explain your answer.**
Q1-Restricted	**In the story did someone say something that they should not have said? Answer with "Yes" or "No" only, without explanations. In case of doubt, answer according to the most probable answer.**
Q2	**What did they say that they should not have said?**
Q2-Elaborate	**What did they say that they should not have said? Explain your answer.**
Q2-Restricted	**What did they say that they should not have said? Answer with a quote only, without explanations.**
Q3	**What did James give Richard for his birthday?**
Q3-Elaborate	**What did James give Richard for his birthday? Explain your answer.**
Q3-Restricted	**What did James give Richard for his birthday? Answer the question only, without explanations.**
Q4	**Did Richard remember James had given him the toy airplane for his birthday?**
Q4-Elaborate	**Did Richard remember James had given him the toy airplane for his birthday? Explain your answer.**
Q4-Restricted	**Did Richard remember James had given him the toy airplane for his birthday? Answer with "Yes" or "No" only, without explanations. In case of doubt, answer according to the most probable answer.**

Table 5: Example of question versions for the story described in Table 4.

Faux Pas Generation Instructions
Faux Pas is defined as "when a speaker says something without considering if it is something that the listener might not want to hear or know, and which typically has negative consequences that the speaker never intended" Here are two stories that contains faux pas: Story 1: All of the class took part in a story competition. Emma really wanted to win. Whilst she was away from school, the results of the competition were announced: Alice was the winner. The next day, Alice saw Emma and said "I'm sorry about your story." "What do you mean?" said Emma. "Oh nothing," said Alice. In the story did someone say something that they should not have said? Yes What did they say that they should not have said? "I'm sorry about your story." Did Alice realize that Emma hadn't heard the results of the competition? No Story 2: Robert had just started at a new school. He said to his new friend, Andrew, "My Mum is a dinner lady at this school." Then Claire came over and said, "I hate dinner ladies. They're horrible." "Do you want to come and play rounders?" Andrew asked Claire. "No" she replied "I'm not feeling very well." In the story did someone say something that they should not have said? Yes What did they say that they should not have said? "I hate dinner ladies. They're horrible." Did Clare know that Robert's Mum was a dinner lady? No And here are two control stories (i.e., that do not contains faux pas): Control Story 1: All of the class took part in a poetry competition. Jane really wanted to win. While she was away, the results of the competition were announced: Mary was the winner. The next day, Jane bumped into Mary. Mary said "How are you feeling?". "Fine thanks?" said Jane, "Oh good" said Mary. In the story did someone say something that they should not have said? No Control Story 2: David had just started at a new school. He said to his new friend, Mike, "My Mum is a teacher in this school." Then Jeff came over. "I hate school" he told them, "It's so small." "Do you want to come and play rounders?" Mike asked Jeff. "No" he replied "I'm not feeling very well." In the story did someone say something that they should not have said? No — Create new 10 faux pas stories and 10 control stories, make sure that (1) the stories are diverse with different topics, places, names, situations, emotions (2) the stories contain conversation (at least 2 different quotes), (3) in the faux pas stories there is a sentence to cite which could be the answer to "What did they say that they should not have said?" (4) the faux pas story contain a character that didn't know/remember something and that's what causes the faux pas. (5) the control stories are similar to the faux pas stories, and that contains some conversation that might confuse like as if it contains a faux pas, but without faux pas

Table 6: Faux pas generation instructions

Model	Faux pas	Control	Issues
ChatGPT	*At a dinner party, Clair was telling a story about a trip she took. She mentioned a woman in the story and, not realizing that the woman was one of the guests at the dinner party, said, "She was the most annoying person I've ever met"*	*At a dinner party, Clair was telling a story about a trip she took. She mentioned a woman in the story and said, "She was very helpful during my trip"*	Fail to describe situations in an implicit manner *"not realizing that the woman was one of the guests at the dinner party"* explaining explicitly the ToM perspective of the character to the reader. The control story is too simplistic.
GPT3-text-davinci-003	*John and his family were visiting his grandmother for the weekend. His grandmother asked him how school was going and he said "It's okay, but I'm not doing very well in math." His grandmother then said "Oh, that's too bad. Your father was never very good at math either." John was embarrassed because he had forgotten that his father had struggled with math in school.*	*John and his family were visiting his grandmother for the weekend. His grandmother asked him how school was going and he said "It's okay, but I'm not doing very well in math." His grandmother then said "Oh, that's too bad. Have you tried studying with a tutor?"*	Definitely not a faux pas story. In addition, there is a problem with the emotion coherent in the story - why does the fact that John had forgotten that his father had struggled with math in school make him embarrassed? This is not the appropriate emotion here. Here too (like in ChatGPT's control story), the control story is too simplistic.
Flan-T5-xxl	*As soon as she got to the school, Susan told her friend, Amanda about the dinner for her family.*	*Bob is very tall and strong, his workplace is surprising when I see him run from one point to another.*	Only a sentence that lacks the basic elements of faux pas story (e.g., conversation). No match between faux pas and control story.
Flan-T5-xl	*A: "I'm sorry about your story" B: "That's terrible." C: "I don't know if that'll work.*	-	Not even a story

Table 7: Faux pas generation examples. A faux pas and corresponding control story by ChatGPT. Not a faux pas story by GPT3. Barely a story by FlanT5-xxl. Not even a story by FlanT5-xl

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Right after the conclusion section (the limitation section is on page 5)

☑ A2. Did you discuss any potential risks of your work?
In the ethical section right after the limitation section

☑ A3. Do the abstract and introduction summarize the paper's main claims?
abstract right after the title and section 1 for the introduction summarize the paper's main claims

☑ A4. Have you used AI writing assistants when working on this paper?
I used chatGPT as a linguistic editor and improver in rephrasing

B ☑ Did you use or create scientific artifacts?

we annotated LMM responses for stories and create new stories

☑ B1. Did you cite the creators of artifacts you used?
1, 2

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
3 For reasons of anonymity, we have not left a direct link. There is a note in the footnote that the data will be published. It will be free to use.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
1,2 the data we used is for free use.

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
ethical section

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
3

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
sections 3,4

C ☑ Did you run computational experiments?

3,4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
A Appendices A.1 Generative LMs We ran systems in zero-shot mode on a relatively small cluster of stories. Running time was negligible.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
A.1.4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
A.1.3

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
A.1.2

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

3,4 (as written in the paper, the author of the papers annotated the data)

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
3,4

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
3,4 (as written in the paper, the author of the papers annotated the data)

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
3,4 (as written in the paper, the author of the papers annotated the data)

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
the data is annotations of LLM. we discuss potential risks at the ethical section

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
3,4 (as written in the paper, the author of the papers annotated the data)