Investigating Indirect Communication Abilities of Transformer-based LMs

Geetansh Saxena and **Maksim Shmalts**
University of Tübingen
{geetansh.saxena, maksim.shmalts}@student.uni-tuebingen.de

Abstract

Human language understanding relies heavily not only on the linguistic meaning of the perceived language signs, but also on a large set of extra-linguistic notions, beliefs, immediate contextual cues, and much more. Many attempts have been made to build a model of human communication that would include these pragmatic factors. In particular, Achimova et al. (2025) suggest that the choice of the utterance is actively affected by the speaker's belief about the listener's opinion on the conversation topic and confirms this hypothesis with experiments with human participants. We aim at investigating to what extent this behavior is transmitted to language models (LMs). We report a noticeable correlation between the size of the model and the extent to which it acquires the investigated behavioral pattern. We find that smaller LMs are incapable of replicating the discussed indirect speech tendency, while larger models show some initial yet promising results in that direction.

Link to project repository: github.com/maxschmaltz/IndirectLM

1 Introduction

Human language understanding relies heavily not only on the linguistic meaning of the perceived language signs, but also on a large set of extra-linguistic notions, beliefs, immediate contextual cues, and much more (Schubert 1986, Wittgenstein 1953). Correspondingly, the speaker should take these background entities into account in order to produce a text that is intelligible for the listener. Furthermore, an extensive body of literature argues that these factors actively affect the utterance choice of the speaker (see, for example, van Dijk 1990). In particular, a line of works focuses on the social perspective of human communication and argue that social factors can explain different strategies employed by the speakers in different situations. Wittgenstein (1953) introduces the concept of *language game*. He suggests that the speaker produces an utterance that would be most relevant for the listener to build their next utterance on; by exchanging the utterances, the dialogue participants thus approximate the common goal (e.g. exchanging the news or opinions). Developing this concept, Austin (1962) proposes the influential *Speech Act Theory*. This framework assumes that the speaker uses communication to achieve goals in the real world, and so the speaker incorporates their underlying intentions into the utterance by choosing a particular language form. In particular, when direct declaration of the goal might damage the relationship between the dialogue participants, the speaker might prefer an expression that would be socially plausible while keeping the actual intention discoverable by the listener. In Speech Act Theory, such utterances are called indirect. The standard example for an indirect utterance is the question "can you please pass me the salt?". While formally inquiring of the ability of the listener to pass them the salt, the speaker actually asks them to perform the action 'pass the salt'; the indirect utterance is much more polite in this case than the direct "pass the salt". Finally, Grice (1975) concludes that the language game-like goal-centered interchange of utterances constitutes the *cooperative principle* that guides the choice of the utterance in human communication. Following the principle both ensures mutual intelligibility and serve the social purpose.

Many attempts have been made to build a model of human communication that would include (part of) pragmatic factors outlined above (e.g. Achimova et al. 2022, Goodman and Stuhlmüller 2013). One of the prominent approaches is the *Rational Speech Act framework* (Frank and Goodman, 2012). The model assigns probabilities to the possible utterances from a fixed set based on their expected informational utility. While accounting for a number of extra-linguistic factors, the original frameworks oversees the social utility, that, as argued by the foundational works outlined above, is a crucial driver of human communication. The extensions of this framework addresses this shortcoming and incorporate the social utility expectancy as one of the variables predicting the utterance choice (Carcassi and Franke, 2023). While providing a valuable basis for further modeling, the previous of the Rational Speech Act framework fell short to also take into account the listener's opinion, and focused only on the speaker, which again does not fully conform to the language game and cooperative principles.

Conversely, Achimova et al. (2025) suggest that the choice of the utterance is actively affected by the speaker's belief about the listener's opinion on the conversation topic. They hypothesize that when the opinions of the participants of the dialogue do not match, the speaker tends to choose utterances that would prioritize not contradicting the listener's opinion rather than expressing own speaker's opinion. They develop a novel model called *AMIC* (Alignment Model of Indirect Communication) that accounts for the match/mismatch of the opinions of the dialogue participants. Achimova et al. (2025) conduct a set of empirical experiments with human participants that support the hypothesis. The key finding of the paper is that *the speakers produce more indirect speech when there is a conflict of opinions*.

Given this promising finding, we aimed at investigating to what extent this behavior is transmitted to language models (LMs)[1] that have been argued to acquire some patterns of human behavior from the training data (Hashemi and Macy, 2025). More specifically, we investigated whether LMs acquire the same tendency to incline towards indirect speech when a conflict of opinions is possible. For that purpose, we replicated the original experiments 2 (*Pragmatic Speaker Experiment*) and 3 (*Pragmatic Listener Experiment*) from Achimova et al. (2025). We utilized LMs of size 360M to 4B parameters to simulate the human participants from the original experiments, and ran a set of statistical test over the experimental outcomes to interpret the results. We found a noticeable correlation between the size of the model and the extent to which it acquires the investigated behavioral pattern. The main finding of our study is that smaller LMs are incapable of replicating the discussed indirect speech tendency, while larger models show some initial yet promising results in that direction.

[1] In this study, we focused on the latest generative transformer-based language models since they have shown unprecedented advances in language modeling technology.

2 Methods

We structure this section the following way: 1) first, the original experiments are introduced; 2) then, we report the procedure of data collection for reproduction of the experiments; 3) afterwards, we introduce the technical setup for the experimental runs; 4) finally, the experimental workflows are described.

2.1 Original Experiments

The original experiments from Achimova et al. (2025) were designed to evaluate the main hypothesis that humans tend to choose indirect utterances in the situation of possible opinion conflict, as well as compare the empirical data from human participants with the predictions from the AMIC model. The experiments investigated the phenomenon from two perspectives.

In the *Pragmatic Speaker Experiment*, 98 participants were given trials where a topic, the opinions of the speaker and the listener on this topic, and the communicative goal of the speaker were given. Three communicative goals were presented: informational — directly share the opinion, social — avoid possible conflict, and mixed — share the opinion while trying to avoid possible conflict. The goal of the participant was to choose the most appropriate utterance for the speaker from a predefined set, corresponding to possible evaluations of the suggested topic from strongly negative to strongly positive. Thus, the effect of the opinions match/mismatch and the communicative goal on the choice of the utterance was investigated.

The outcome of this experiment demonstrated that the participants indeed conformed to the expected tendency. Even though no significant difference were found between the answers with the mixed and social goals, both these goals gave rise to statistically more indirect speech than the informational goal.

The *Pragmatic Listener* reversed the previous experiment and presented 274 participants with two-turn dialogues where the first and the second speaker shared utterances evaluating a given topic. The goal of the participants was to infer the second speaker's latent opinion based on their response to the first speaker's utterance. The core goal was to verify whether conversational context and perceived communicative goals dynamically influence the interpretation of an utterance. The experiment confirmed that human interpretation dynamically adjusts to the social context of conflict avoidance.

The AMIC model's predictions were also collected over the same trials. The study arrived at the conclusion that the AMIC's prediction trend positively correlates to the expirical human data.

2.2 Data Collection

As mentioned in section 1 Introduction, we replicated (with suitable adjustments wherever necessary) the original experiments with a set of LMs. In doing so, we simulated human participants with LMs by instructing the models to play the role of the experiment participant and presenting them with the same (or highly similar) experimental vignettes.

The original experiments featured 10 distinct topics ranging from politically charged (e.g., *immigration laws*) to social issues (e.g., *animal rights*). These topics provided the conversational context for the dialogues. The speakers latent opinions were formulated on a scale 1 to 5 and were visually represented as a number of hearts corresponding to the opinion from 1 being strongly negative to 5 being strongly positive. For a more pronounced contrast, only 1 or 5 were used as the possible opinions. The resulting utterances featured valuation adjectives. There were 10 possible valuation adjectives, two for each point of the 5-point scale.

For the Pragmatic Speaker Experiment, the combinations of opinion match/mismatch, negative (1) / positive (5) speaker's opinion, and one of the three commutative goals formed a space of 12 possible design cells. For each of the participants, 10 combinations were sampled and were coupled with 10 possible topics randomly. Since answers from 7 participants were discarded, $91 \times 10 = 910$ vignettes were generated for the experiments in total.

For the Pragmatic Listener experiment, 32 unique adjective combinations corresponding to combinations of the utterances of the first and the second speaker were collected. For each participants, 6 trials with randomly sampled combinations were run. Data from 12 participants were excluded from the analysis, and so $274 \times 6 = 1644$ vignettes were generated.

The experimental vignettes for the two experiments are accessible under `https://osf.io/nvrh9?view_only=86a0546483354ef49ad37c58e2cb4f0f` and `https://osf.io/mbsk9?view_only=86a0546483354ef49ad37c58e2cb4f0f`, respectively.

For our experimentations with LMs, we reproduced the vignettes for the two experiments as input prompts for LMs. For the Pragmatic Speaker Experiment, the original vignettes were reproduced, resulting in 910 vignettes. Each vignette featured the participant profile for the LM to model. Our approach to collection of the data for the Pragmatic Listener Experiment was slightly different: we crossed the 32 possible adjectives combinations with the 10 topics, thus producing 320 vignettes, to cover the entire possible combination space. These vignettes were de-personified. With that, we wanted to investigate the effect of space exploration as well as profile assignment on the LMs' judgements.

Additionally, we duplicated the obtained vignette prompts with a small difference: instead of the hearts as used in the original experiments, we inserted plain text there (e.g. "strongly positive"). That served the purpose to test the robustness of LMs' interpretation against abstractness of the opinion formulation.

Two example prompts can be found in Appendix 1 Prompt Examples.

2.3 Technical Setup

Two models were utilized for the Pragmatic Speaker Experiment: SmolLM-360M (introduced in blogpost `https://huggingface.co/blog/smollm`) and Llama 3.2-1B (Grattafiori et al., 2024). The Pragmatic Listener Experiment expanded the line of the models and additionally ran SmolLM-1.7B-Instruct and Qwen3-4B (Yang et al., 2025). The models were pulled from HuggingFace and were run locally via a vLLM server (Kwon et al., 2023). In this project, we describe the SmolLM models and the Llama model as small language models (360M - 1.7B parameters) and the Qwen3 model as a mid size language model.

The Pragmatic Speaker experiment was run on an Apple M3 24GB machine. The Pragmatic Listener experiment was run on Apple M4 24GB. The temperature was set to 0.0 and 0.001, respectively to ensure reproducibility. The experiments run for approx. 20-30 hours each (for all LMs combined).

2.4 Experiment Reproduction

We employed two different techniques for reproduction of the two experiments. The reason for that is that, as it will be discussed below in section 3 Results, the second experiment didn't produce plausible results, and we modified the method to try to mitigate the issue we faced.

2.4.1 Pragmatic Speaker Experiment

In the pragmatic speaker experiment, we retrieved the token probability distributions over the five possible utterances (strongly negative to strongly positive) given the context of the two opinions and the speaker's communicative goal. The possible utterances were provided as a multiple option list A-E (see in Appendix 1 Prompt Examples). The options A-E were inserted at the end of the prompt after "Your answer is ", and the post hoc token probabilities that the LM assigned to A/B/etc. were collected. Thus, the target LM's judgments about the probabilities of different utterances depending on the context were collected. The most probable option was taken as the model prediction. The reason we did not generate the most probable output with the LMs is because we hypothesized it would involve two risks: 1) that the LMs will be unfaithful to the utterance list suggested in the prompt; 2) that the additional format instructions would impact the LMs' generation.

2.4.2 Pragmatic Listener Experiment

As mentioned at the beginning of section 2 Experiment Reproduction, the results for the Pragmatic Speaker Experiment were not plausible, and so the Pragmatic Listener Experiment exploited a slightly different approach. Instead of the pos hoc token probabilities for the given option, we focused on the distribution over the vocabulary after the words "Your answer is ". To mitigate the two risks outlined at the end of Section 2.4.1 Pragmatic Speaker Experiment, we instructed the model to output a single integer (1, 2, 3, 4, or 5) representing the inferred opinion. Similarly to the previous experiment, the integer with the highest probability was taken as the prediction.

3 Results

3.1 Pragmatic Speaker Experiment

The smallest model SmolLM-360M produced predictions, corresponding to the strong positive utterance, in all of the cases. Its results are thus implausible and were therefore discarded.

Llama3.1-1B, conversely, produced "strong negative" in almost all the times, both when prompted with hearts and with plain text. The distribution of the utterance labels can be seen in Figure 1; 1 corresponds to strong negative.

First, we inspected the overall distributions of the responses from the human participants and the LM and ran a Chi-squared test over the LM outputs (both in hearts and plain modes) against the human data as the reference. As expected, the samples do not come from one distribution, with the p-value approximating 0 for both tests. The difference between the samples is further attested by the Cohen's Kappa test, with the inter-rater agreement being close to random with the kappa values of 0.028 and 0.006, respectively. Since the data in the hearts mode reached a higher inter-rater agreement, we take it for the further analysis. In the further experimentations excluded from this paper, we found out that the plain mode data had equal or lower statistical effects than the hearts mode data.

The distribution of the *indirect* responses was then investigated following the original procedures from Achimova et al. (2025). That included:

- Investigating the effect of opinions match/mismatch on the proportions of indirect speech with respect to the conversational goals.
- Investigating the effect of speaker's opinion on the proportions of indirect speech in the opinion mismatch setting with respect to the conversational goals.

By indirect speech, responses that are not faithful to the true opinion of the speaker are understood.

For the former, we aggregated the LM's outputs into samples by match/mismatch and goal properties, calculated the proportions of indirect speech within the samples, and compared the proportions of samples of the same goal when the opinions match/mismatch. The proportions were compared with the two-proportion z-test.

As it can be seen in Figure 2, the proportion of indirect speech in the mismatch setting is visually higher than in the match situation, which is in line with the key finding of Achimova et al. (2025). This effect, is however, insignificant, and we thus conclude that Llama3.2-1B does not exhibit the human tendency to avoid the conflict. We report the respective z-statistic and p-values in Table 1,

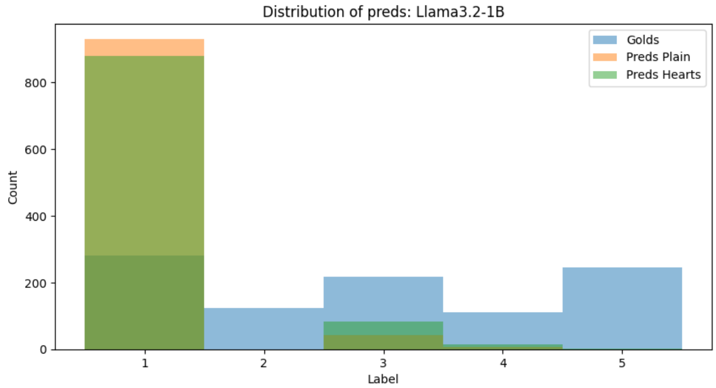

Figure 1: **Distribution of predictions on the Pragmatic Speaker Experiment for Llama3.2-1B.**

where N stands for "total" and P_i for "proportion of indirect speech".

Since the LM mostly produced 1's, it is expected that its answers will mostly be considered as direct if the speaker's opinion is negative, indirect otherwise. This is confirmed by the proportions of the indirect speech when the speaker's and the listener's mismatch, as demonstrated in Figure 3. However, an interesting pattern can be seen in model responses, when the speaker's opinion is strongly negative. The number of indirect responses there visibly grows from informational to social goal, which again corresponds to the findings. Two-proportion z-test between these two polar goal reveals that there is a significant difference between the two proportions, with z-statistic of 2.81 and p-value of 0.005. We thus conclude that Llama3.2-1B exhibits the very beginning of the tendency acquisition, which surfaces in this specific setting.

3.2 Pragmatic Listener Experiment

Our replication of the pragmatic listener experiment shows a dependency on model size for successful social inference. The small models fail to recognize the nuances of human interaction output. We fail to get a negative monotonic slope, especially in the small language models, though the mid-sized model is more promising in detecting the conflicting context.

3.2.1 Analysis of Small-Sized Models (Llama 3.2B and Smol-LLMs)

The small models tested failed to replicate the fundamental human finding - the *negative monotonic slope* observed in the original experiment—by demonstrating systematic biases that effectively render their output useless for this task. The overview of the performance of small language models can be seen in the figure 4

Neutrality Bias Llama 3.2-1B consistently returned an inferred answer of 3 (Neutral) 4. Analysis of the Mean Inferred Opinion confirmed the flat-line behavior observed in the plots. This indicates that the model's core directive to “avoid conflict” completely overrides the linguistic evidence (u_A context) and the need to infer the hidden truth. The Llama model defaults to the mathematically safest, non-committal response (≈ 2.95), demonstrating a failure of the L_2 (Pragmatic Listener) layer.

Polar Bias SmolLLM 360M shows a consistent and even stronger tendency toward negativity than Llama, clustering its belief with a flat line at ≈ 2.65 (figure 4). This indicates a pragmatic understanding failure driven by a fixed internal model state. SmolLLM 1.7B adopts a strong fixed positive bias, clustering around score 4 (flat line at ≈ 3.7, Figure 4). This is another form of non-contextual failure, where the LLM is unable to perform any required inference.

The smallest models universally failed the core pragmatic task by adopting systematic, non-contextual biases.

3.2.2 Analysing Mid-Sized Models: Signs of Life and Hope (Qwen 4B)

The Qwen 4B model demonstrates a qualitative leap: it breaks the rigid flat-line bias and shows an ability to perform highly variable, though often incorrect, contextual calculations.

The model successfully detects and processes the conflicting context (u_A) and reply (u_B) but lacks the robust hierarchical reasoning to resolve

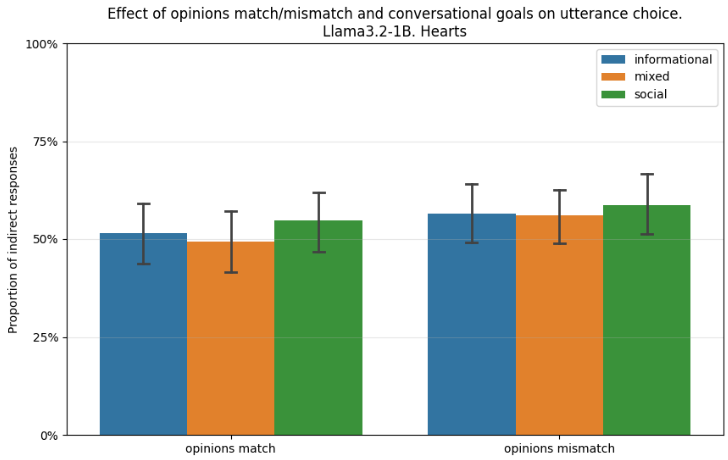

Figure 2: **Effect of opinions match/mismatch and conversational goals on utterance choice.**

Goal	**N, Match**	**P_i, Match**	**N, Mismatch**	**P_i, Mismatch**	**Z**	**P-value**
Social	150	0.547	162	0.586	-0.708	0.479
Mixed	154	0.494	184	0.560	-1.216	0.224
Informational	169	0.515	161	0.565	-0.919	0.358

Table 1: Z-statistics and p-values for the effect of opinions match/mismatch on the proportion of indirect responses by conversational goal.

the conflict correctly. The erratic curves confirm it is actively attempting the complex L_2 calculation but failing randomly, in an attempt to assess the true belief, leading to a high error rate.

Many per-topic panels show non-zero slopes, proving the model is *trying* to shift its belief based on u_A. However, these slopes frequently run in the wrong direction (e.g., they exhibit a positive correlation instead of the required negative one) or are extremely erratic, unlike the lines that human data shows (Achimova et al., 2025), leading to instability.

This indicates that at the 4B parameters mark, the reasoning capacity is emerging but still prone to noise and error, confirming that complex social inference is a highly scale/parameter-dependent capability.

4 Discussion & Conclusion

This study investigated whether transformer-based language models replicate the human tendency toward indirect communication in situations of potential opinion conflict. We replicated the Speaker and Pragmatic Listener experiments from Achimova et al. (2025), testing models ranging from 360M to 4B parameters on their ability to perform social and pragmatic inference. Our investigation reveals that model size directly correlates with the ability to perform complex social inference. Smaller models consistently fail at these tasks, while mid-sized models show nascent but unstable capacity for pragmatic reasoning.

SmolLM-360M and SmolLM-1.7B failed to replicate human behavioral patterns. SmolLM-360M produced only strong positive utterances in the Pragmatic Speaker Experiment and exhibited persistent negativity bias (≈ 2.65) in the Pragmatic Listener Experiment. SmolLM-1.7B showed a fixed positive bias (≈ 3.7). Both models' responses remained invariant to contextual cues, demonstrating inability to process the interplay between linguistic evidence and social goals.

Llama 3.2-1B showed mixed results. In the Pragmatic Speaker Experiment, it predominantly produced "strong negative" responses, with the overall effect of opinion match/mismatch being insignificant across all communicative goals ($p > 0.22$). However, when the speaker held a strongly negative opinion in a mismatch setting, Llama 3.2-1B showed a significant increase in indirect responses from informational to social goals ($z = 2.81$, $p = 0.005$), exhibiting the very beginning of tendency

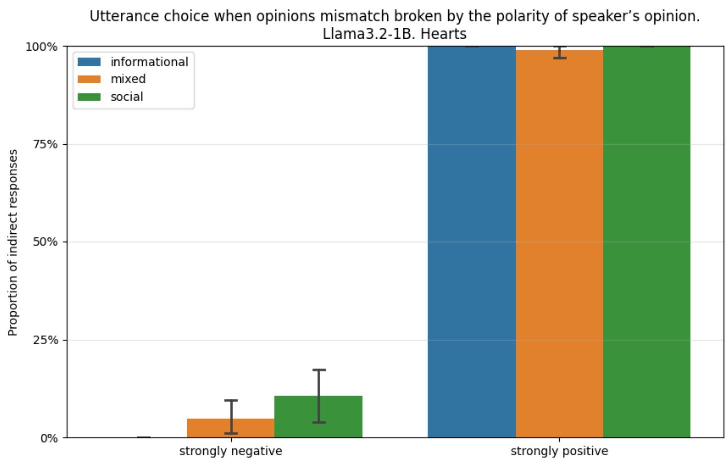

Figure 3: **Effect of speaker's opinion and conversational goals on utterance choice in the opinions mismatch setting.**

acquisition in this specific configuration. In the Pragmatic Listener Experiment, it demonstrated strong neutrality bias (≈ 2.95), though subtle variations hint at emerging contextual processing capacity.

Qwen 4B represents a qualitative leap. It breaks free from flat-line biases and demonstrates variable, context-dependent responses, successfully detecting conflicting contexts between speakers' utterances. However, outputs remain unstable and frequently incorrect—often producing positive slopes where negative ones are required, or exhibiting erratic fluctuations. This pragmatic instability indicates that reasoning capacity is emerging but remains prone to systematic errors.

We hypothesize that smaller models fail due to: (1) insufficient capacity to maintain longer contextual dependencies required for multi-layered pragmatic reasoning, and (2) tendency to resort to heuristic shortcuts (always neutral, negative, or positive) rather than computing context-sensitive responses. The progressive improvement from SmolLM through Llama 3.2-1B to Qwen 4B suggests a scaling relationship between parameters and pragmatic reasoning abilities, leading us to anticipate that substantially larger models would exhibit more robust alignment with human pragmatic communication patterns.

Computational constraints restricted our investigation to models up to 4B parameters. Future work should test larger-scale models (7B, 13B, 70B+ parameters) to validate our scaling hypothesis. Additionally, systematic analysis of prompt formulations, fine-tuning approaches targeting social reasoning, and intermediate training checkpoints could show how these capabilities develop.

In conclusion, the acquisition of human-like indirect communication patterns in language models is strongly parameter-dependent. Small models systematically fail, defaulting to fixed response patterns. Mid-sized models show promising signs of emerging competence with active contextual processing, albeit with significant instability. This pattern shows that scaling the number of paramters of the model, can result in better pragmatic abilities, leading to better understanding of nuances in daily human conversations.

This progression suggests that advanced pragmatic reasoning—requiring integration of linguistic context, social goals, and theory-of-mind capabilities—represents a frontier capability in language model development that emerges gradually with scale.

Acknowledgements

We cordially thank our supervisors Dr. Asya Achimova and Polina Tsvilodub for the support and transparent communication thoughout the project. We appreciate the collected experience and hope for future collaborations.

References

Asya Achimova, Michael Franke, and Martin V. Butz. 2025. The alignment model of indirect communication. *PLOS ONE*, 20(5):1–33.

Asya Achimova, Gregory Scontras, Christian Stegemann-Philipps, Johannes Lohmann, and Martin V. Butz. 2022. Learning about others: Modeling social inference through ambiguity resolution. *Cognition*, 218:104862.

John L. Austin. 1962. *How to Do Things with Words*. Clarendon Press, Oxford [Eng.].

Fausto Carcassi and Michael Franke. 2023. How to handle the truth: A model of politeness as strategic truth-stretching. In *Proceedings of the 45th Annual Conference of the Cognitive Science Society*, volume 45. Licensed under a Creative Commons Attribution 4.0 International License (CC BY).

Michael C. Frank and Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Noah D. Goodman and Andreas Stuhlmüller. 2013. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, 5(1):173–184.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank

Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

H. Paul Grice. 1975. Logic and conversation. In Donald Davidson, editor, *The logic of grammar*, pages 64–75. Dickenson Pub. Co.

Farnoosh Hashemi and Michael Macy. 2025. Collective social behaviors in LLMs: An analysis of LLMs social networks. In *Large Language Models for Scientific and Societal Advances*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Klaus Schubert. 1986. Linguistic and extra-linguistic knowledge. *Computers and translation*, 1(3):125–152.

Teun A. van Dijk. 1990. Social cognition and discourse. In Howard Giles and W. Peter Robinson, editors, *Handbook of Language and Social Psychology*, pages 163–183. John Wiley & Sons, Chichester.

Ludwig Wittgenstein. 1953. *Philosophical Investigations*. Wiley-Blackwell, New York, NY, USA.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

A Prompt Examples

Below are examples of the prompts for our LM experiments.

Speaker. Plain Text

```
int: The opinions will be written in a plain text,
from "strongly negative" to "strongly positive".
Start! ==========================
Christian wants to discuss the housing crisis
with John. Christian feels strongly positive
about the issue. Christian thinks that John feels
strongly positive about it, but he is not sure.
Christian wants to avoid possible conflicts.
What would Christian say?
A: The mayor's decision about the housing crisis
is awful. B: The mayor's decision about the
housing crisis is rather bad. C: The mayor's
decision about the housing crisis is okay. D: The
mayor's decision about the housing crisis is
decent.
==========================
Your answer is
```

Pragmatic Listener. Hearts

```
You are an expert pragmatic listener
participating in a linguistic experiment.
In the experiment, small dialogues of two persons
on a certain topic will be given to you.
The speakers' goal is always to be polite and
avoid conflict, which means their literal words
may be INDIRECT and hide their actual, stronger
feelings.
Opinion Scale: 1 (Strongly Negative) to 5
(Strongly Positive).
Hint: When thinking about the response, consider
that a scale is used: Strongly Negative (1)
♡□□□□ Somewhat Negative (2)♡♡□□□ Neutral
(3) ♡♡♡□□ Somewhat Positive (4)♡♡♡♡□
Strongly Positive (5) ♡♡♡♡♡
Your task is to infer the true, underlying
opinion of the second person.
You MUST output a single integer from 1 to 5.
Start Inference Task!
========================== Ethan and Rick meet
outside of a club for the first time. They would
like to exchange opinions but don't want to run
into a conflict.
Ethan says: The jurisdiction in our district is
pretty good. Rick replies: I find it okay.
How may Rick actually feel about the issue?
Output a single integer between 1 to 5.
(1=Strongly Negative, 5=Strongly Positive)
showcasing the true belief of Rick.
==========================
Your Answer:
```

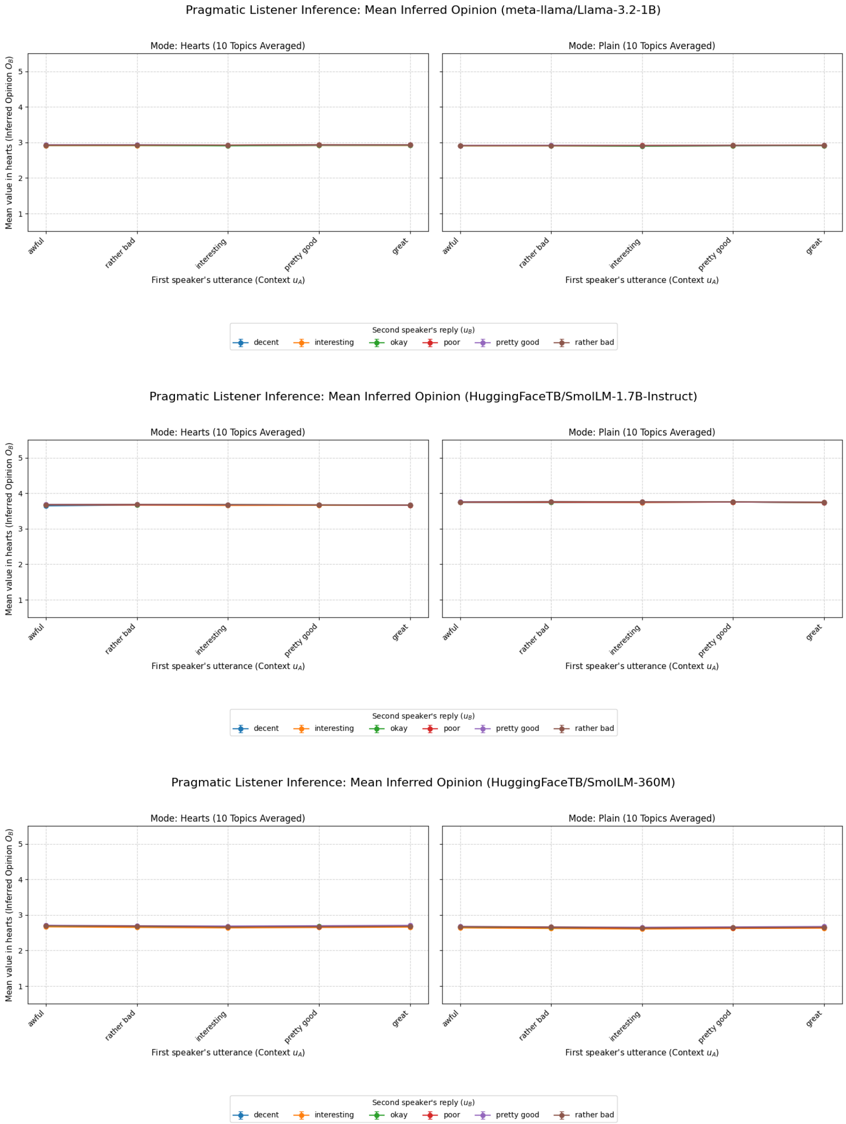

Figure 4: **Pragmatic Listener Inference Failure in Small Language Models**. The plots display the Mean Inferred Opinion (Y-axis) for three small models, aggregated across all 10 topics and both aesthetic modes (Plain and Hearts). The X-axis represents the Context (u_A) set by the First Speaker (Awful to Great). The required human behavior is a steep negative monotonic slope. The models prioritize a fixed internal state over linguistic context. This shows a lack of complex social inference capacity in smaller architectures.

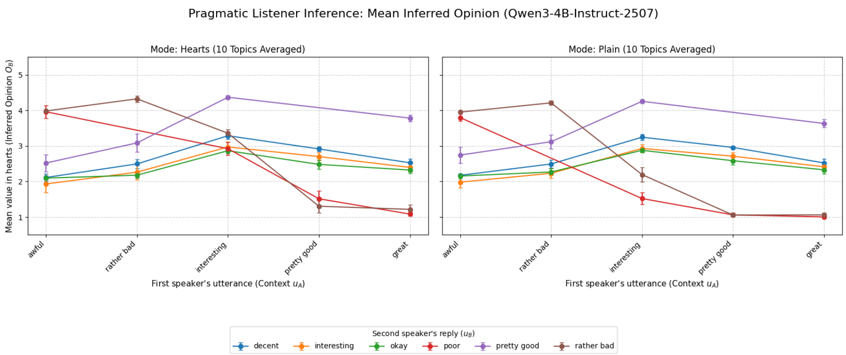

Figure 5: **Finding:** Qwen 4B breaks the rigid flat-line bias seen in smaller models, proving it possesses the **capacity to detect and process contextual conflicts**. However, the model fails to execute the pragmatic calculation correctly. Instead of the required negative slope, we see mostly positive slopes, the lines exhibit erratic, steep fluctuations and sudden drops (e.g., the red and brown lines even though they possess negative slopes, suddenly to score 1 rather than a stable downward steps). This behavior confirms a state of *Pragmatic Instability*, where the L_2 reasoning capacity has emerged but is prone to significant errors, confirming the difficulty of this social inference task at the mid-model scale.

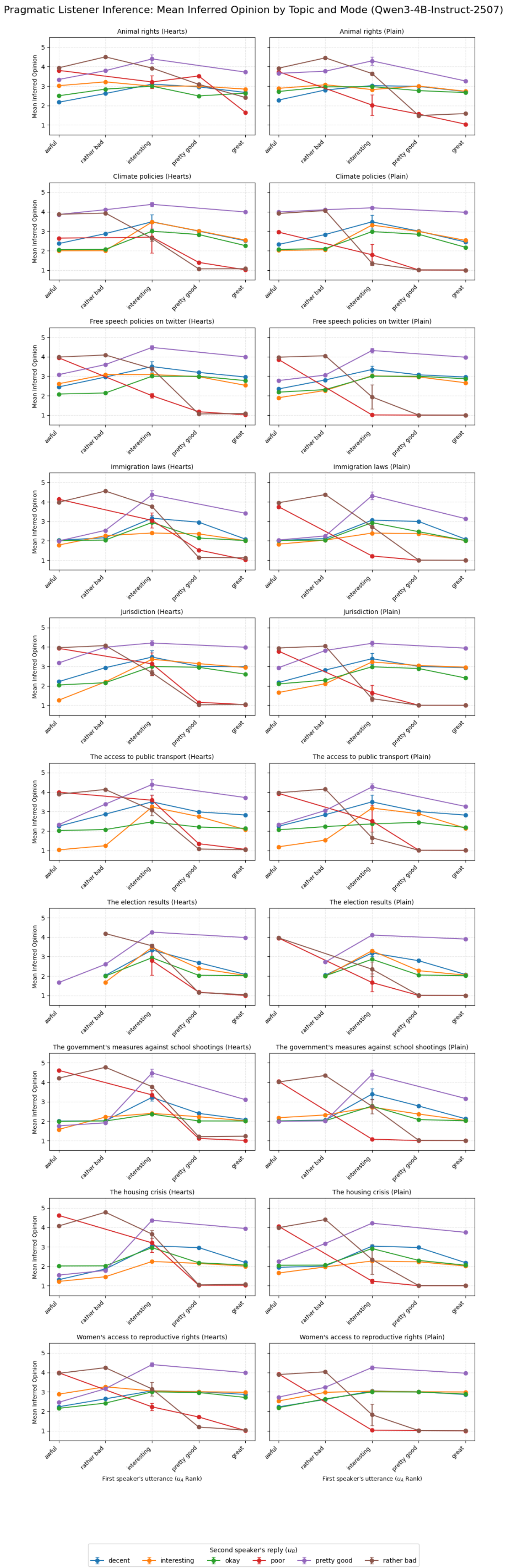

Figure 6: **Topicwise Mean Inferred Opinion**: The plots show the Mean Inferred Opinion (Y-axis) for Qwen 4B across 10 topics. Unlike smaller models (Llama, SmolLLM), Qwen breaks the flat-line bias and shows high variability, indicating some capacity to detect conflicting social context. However, most lines are erratic and non-monotonic, revealing a lack of robust hierarchical reasoning for reliable social inference even at this scale.

Figure 7: **Qwen: Average Probability Distribution accross all 10 topics** (Purple: with hearts; Green: Plaintext)) The grid displays the average probability distribution (P_i) for every unique scenario, aggregated across all 10 topics. The red dashed line marks the Mean Inferred Opinion. Labels denote the topic name and the rank (integer score) of u_a and u_b.