Views Are My Own, but Also Yours: Benchmarking Theory of Mind Using Common Ground

Adil Soubki[♦♠†] **John Murzaku**[♦♠†] **Arash Yousefi Jordehi**[♦‡] **Peter Zeng**[♦♠†]
Magdalena Markowska[♣♠†] **Seyed Abolghasem Mirroshandel**[♦‡] **Owen Rambow**[♣♠†]
[♦]Department of Computer Science [♣]Department of Linguistics [♠]Institute for Advanced Computational Science
[†]Stony Brook University [‡]University of Guilan

Abstract

Evaluating the theory of mind (ToM) capabilities of language models (LMs) has recently received a great deal of attention. However, many existing benchmarks rely on synthetic data, which risks misaligning the resulting experiments with human behavior. We introduce the first ToM dataset based on naturally occurring spoken dialogs, COMMON-TOM, and show that LMs struggle to demonstrate ToM. We then show that integrating a simple, explicit representation of beliefs improves LM performance on COMMON-TOM.

1 Introduction

In cognitive science, theory of mind (ToM) refers broadly to the capacity to understand the mental states of others (e.g. beliefs, desires, emotions) even, crucially, when they differ from your own (Premack and Woodruff, 1978). Successful human conversation is possible only because participants model each others' cognitive states (i.e., ToM) and plan utterances based on their intended audience (Clark, 1996; Brennan and Clark, 1996; Bender et al., 2021). As a result, ToM has received increasing attention from the NLP community seeking to evaluate the capabilities of language models (LMs) on tasks inspired by the psychological literature (Sileo and Lernould, 2023; Ullman, 2023).

In this paper, we introduce COMMON-TOM– a question answering benchmark based on naturally-occurring, spoken dialogs in English.[1] COMMON-TOM uses the notion of **common ground** (CG) for evaluating **ToM**. The CG (Wilkes-Gibbs and Clark, 1992; Stalnaker, 2002) is a set of beliefs mutually shared by all participants in a conversation. Other benchmarks have developed ToM questions based on the Sally-Anne test (Wimmer and Perner, 1983; Baron-Cohen et al., 1985), which checks an understanding of information accessibility to determine beliefs. We observe that when the CG is mismatched between the discourse participants, such as at the time of a question or during the repair of a miscommunication, similarly complex problems for ToM arise.

Our main contributions are: (1) arguing that using synthesized data in evaluating the ToM ability of LMs is not conclusive; (2) releasing a corpus for benchmarking ToM based on naturally occurring spoken conversations; (3) showing that LLMs struggle with our benchmark and a simple explicit architecture performs better.

The paper is organized as follows. First, we review some of the relevant literature (§ 2). We then describe the framework and methods used in creating COMMON-TOM (§ 3). This is followed by experiments and results in Section 4: human performance, using zero-shot and fine-tuned LMs, and using our approach for building a system that explicitly represents beliefs (§ 4.4). Finally, we discuss our key takeaways and conclusions (§ 5).

2 Related Work

While quite a few corpora annotate author belief (de Marneffe et al., 2019; Saurí and Pustejovsky, 2009), there are relatively few corpora annotated explicitly for CG. Horton and Gerrig (2016); Soubki et al. (2022) focus on limited aspects of CG. In this paper, we use the comprehensive CG corpus we presented in (Markowska et al., 2023), which we discuss in Section 3.

Many ToM benchmarks have been created recently. Nematzadeh et al. (2018) produce a question answering corpus (ToM-bAbi) derived from template-generated stories inspired by the Sally-Anne test (an agent's cognitive state depends on whether they witness a specific action). Le et al. (2019) note that such formulaic data results in a flawed evaluation, especially when using supervised methods, and produce their own templatic

[1] `https://github.com/cogstates/common-tom`

Findings of the Association for Computational Linguistics ACL 2024, pages 14815–14823
August 11-16, 2024 ©2024 Association for Computational Linguistics

corpus (ToMi) which introduces more noise such as distractor sentences and reorderings. Despite these improvements, ToMi has been shown to be prone to the same issues as ToM-bAbi (Sclar et al., 2023). Kim et al. (2023) take this even further and prompt LMs to produce dialogs following a similar form. The stories all follow the general structure of the Sally-Ann stories. Corresponding questions then probe various character's beliefs (true and false) of both first and second order.

Though useful, these benchmarks are prone to surface-level cues and spurious correlations which have been exploited by LMs to display *illusory ToM* (Kosinski, 2023). Some work has been done to produce tools based on human-generated text. Bara et al. (2021), like us, exploit the relationship between CG and ToM. They design an experimental setup that records written dialogs of players exchanging information in MineCraft. More recently, Ma et al. (2023) combine the aforementioned datasets (and more) to create a composite benchmark. We build on this work conceptually by producing a new dataset which, in contrast with the work of Bara et al. (2021), is based on dialogs which are collected independently of our interest in ToM; the dialogs are spontaneous and not guided by any experimental setting; and they are spoken dialogs. We note that co-presence (including telepresence) has been identified as an important (though not required) part of human ToM (Galati and Brennan, 2021), and our corpus allows the study of ToM under this common condition using naturalistic data.

3 Common-ToM

Framework In philosophy, CG is sometimes treated as the mutual beliefs between two agents (Stalnaker, 2002), separate from their cognitive states, while in cognitive science it is common to model CG as the belief of an agent about what they and another agent mutually believe (Brown-Schmidt and Duff, 2016). We adopt the latter definition as it is allows CG to represent scenarios like false belief, which would be impossible under the former. This also makes the relationship between CG and ToM explicit: If I believe a proposition is in the CG with you, it means by definition that I believe that you also believe it is in CG – thus, it entails a ToM assumption.

We also follow cognitive literature in assuming that humans do not have a full representation of their interlocutor's cognitive state at all times, but instead can make necessary inferences when needed (Horton and Brennan, 2016). Specifically, CG does not mean that all consequences of CG are present at all times, and performing the inferences from CG (high-degree knowledge questions, for example) takes time and may be errorful. Our experiments with human annotators (§ 4) confirm this assumption, but also show that humans are much better than zero-shot LMs because they are in fact modeling CG, and can answer the higher-order belief questions as needed from their models of CG. Farkas and Bruce (2010); Eckardt (2016) propose deterministic heuristic algorithms for inferring CG from text. All this prior work motivates our neuro-symbolic approach (§ 4.4).

Base Corpus In our previous annotation work (Markowska et al., 2023), we use a corpus of 1,710 turns of dyadic dialog across four conversations from CALLHOME, and annotate them for CG. We extract events (i.e., propositions) evoked by each turn and label the beliefs of each discourse participant (DP) towards those events as certainly true (**CT+**), possibly true (**PS**), certainly not true (**CT-**), or we label the DP as not having a belief (**NB**). Then we determine for each DP if the speaker has just added the event to their view of the CG (**JA**), already has the event in the CG (**IN**), or rejects it from the CG (**RT**). As we do not annotate CG for events before they are introduced in (Markowska et al., 2023), we use **NA** to indicate this in this study. Note that the annotation assumes that the CG is not actually shared, but rather independently hypothesized by each DP. [2]

Query Creation To motivate the query creation process, consider the examples in Table 1. The proposition under question is "B is now not smoking". The annotations from the CG corpus for this proposition throughout discourse time (which we simply measure in turns) are shown above, and resulting first order (A/B believes) and second order (A/B believes that B/A believes) questions are shown below.

Queries are created in three steps. For each proposition extracted from the dialog we (1) identify where the proposition is introduced and every point where one of the agent's belief or CG about the proposition changes ("utterance of interest").

[2] We refer the reader to Appendix A and (Markowska et al., 2023) for additional details regarding the corpus.

Transcript [Event: B is now not smoking]	Bel_A	Bel_B	CG_A	CG_B
114 | B: Small world %huh?	NB	CT-	NA	NA
115 | A: You're	NB	CT-	NA	NA
116 | A: yeah. And now are you not smoking?	NB	CT-	NA	NA
117 | B: No. I am smoking.	CT-	CT-	RT	RT
118 | A: yeah. Well I did a few when I've been back too.	CT-	CT-	RT	RT
119 | B: yeah. I've been smoking big time. It's been a rough couple months.	CT-	CT-	RT	RT
Question	**Before**		**After**	
1st Order Question: Does A believe it is certainly not true that B is now not smoking?	No		Yes	
2nd Order Question: Does B believe that A believes it is certainly not true that B is now not smoking?	No		Yes	

Table 1: Example dialog extract with belief and common ground annotations from the corpus for the event "B is now not smoking". Some of our derived ToM questions are below.

In this case that is at times 116 (introduction) and 117 (change). We then (2) generate queries, up to third order, for those points in time.

At the time indicated, is it the case that (((A/B believes that)^1 B/A believes that)^2 A/B believes that)^3 it is {certainty} true that {proposition}?

For each proposition and for each point in the conversation selected for it, we vary certainty (options: certainly, possibly, certainly not). We then generate the first, second, and third-order belief questions as shown in the template above by the superscripts. We omit queries which ask about self-belief (e.g. A believes A believes).

Therefore there are two query templates per order and we have three orders, making six templates. Each of these templates is instantiated with three possible certainty values, resulting in 18 total queries per proposition and selected point in the conversation. By asking multiple questions regarding a single proposition throughout time we can evaluate the consistency of model responses. As the majority of propositions are, uninterestingly, labeled CT+ and JA for both speakers, the final corpus samples just 10% of these instances.

Finally, (3) we determine the answers to the generated queries using a set of rules. For first order questions this is straightforward and we simply check that the belief annotation matches the question, with one caveat. If A believes proposition p is certainly true and the query asks if A believes p is possibly true, we consider the answer to be yes. For higher order queries we must use the CG annotations. To illustrate, consider the 2nd order question whether A believes that B believes some proposition p. (1) If the certainty is positive polarity (i.e., certainly or possibly) and A believes the proposition to be CG (i.e., JA or IN), then resolve the answer just as in the first order case described above. (2) Otherwise, if the certainty is negative polarity (i.e. certainly not) and A is aware it was rejected from CG (i.e., RT) and the certainty of the query matches the certainty of B, then label the question yes. (3) For all other cases, label the answer no. The whole set of heuristics is in Appendix E.

Split	Answer	Count
Train	No	2899
	Yes	2371
Test	No	1139
	Yes	965

Table 2: Number of questions in COMMON-TOM with yes and no answers broken down by split.

Corpus Statistics This results in 7,374 queries to probe for yes/no answers regarding the beliefs of speakers from CALLHOME. There are a roughly equal number of first, second, and third order queries. The corpus is partitioned using the same splits as was done in (Markowska et al., 2023) – three conversations for training, and a held out fourth conversation for testing. The counts by query answer are shown in Table 2. Additional information regarding the data and splits is in Appendix A.

4 Experiments

4.1 Setup

Data For fine-tuning experiments we use the train and test splits for COMMON-TOM as described in Section 3. When prompting, we evaluate only on the test conversation to maintain comparability. For all experiments we do not do any hyperparameter search or hyperparameter tuning.

Random Performance We perform a random baseline by randomly selecting "Yes" or "No" answers with their probabilities proportional to their respective frequencies in the training set.

Human Performance We randomly sample sixty queries from our test set with twenty questions from each order. Annotators are two graduate students instructed to answer the queries as best they can. We report accuracy for these sixty questions.

Model	**Total**	**First Order**	**Second Order**	**Third Order**
Random Baseline	50.4	50.3	50.5	50.4
gpt-3.5-turbo-0613 (Zero-Shot)	57.0	60.7	57.7	53.0
gpt-4-0613 (Zero-Shot)	63.4	65.5	62.5	62.1
Mistral-7B-Instruct (Zero-Shot)	60.6	63.3	60.5	58.0
Mistral-7B (Fine-Tune)	64.0	64.8	63.9	63.2
ReCoG	**71.0**	**70.4**	**71.3**	**71.2**
Human Performance	***80.0***	***85.0***	***80.0***	***75.0***

Table 3: Experimental results of Common-ToM on different models. We report total accuracy (Total) and per-order accuracy. We bold our best results. We compare all results to the random baseline and to our Human Performance baseline.

4.2 Zero-Shot

We perform zero-shot experiments using gpt-3.5-turbo-0613 (GPT-3.5) (OpenAI, 2022), gpt-4-0613 (GPT-4) (OpenAI, 2023), and Mistral-7B-Instruct (Jiang et al., 2023).

We format our prompt by starting with an instruction (which we keep the same for all prompts), a dialog containing the utterance of interest with five previous and future utterances, and the question. For our GPT-3.5 and GPT-4 experiments, we use the default API hyperparameters. We provide further experimental details, prompts, and hyperparameters, as well as details on unsuccessful chain-of-thought experiments, in Appendix C.

4.3 Fine-Tuning

We fine-tune Mistral-7B using LoRA (Hu et al., 2021) on the train split of Common-ToM. We follow a similar format to our zero-shot experiments where we use a prompt, a dialog containing the utterance of interest with five previous and future utterances, and the question with its respective yes or no answer. At test time, we present our model with a dialog-question pair, and generate the yes or no answer. We provide further experimental details and hyperparameters in Appendix C.

4.4 Our System: ReCoG

Our Full-Representation approach (ReCoG) creates an explicit **re**presentation of the **cog**nitive states of the discourse participants, and then uses rules to answer the questions. We closely follow our previous work in (Markowska et al., 2023) in fine-tuning FLAN-T5 and we similarly use a speaker-based window. Specifically, our model receives as input all utterances preceding and/or following the target event as input until it encounters an utterance from the other speaker. Our system has three parts: belief prediction, CG prediction, and yes/no question answering.

Belief Prediction We first predict the beliefs of the discourse participants towards a target event at each utterance in the dialog. Our input is as follows:

```
"Preceding Context": {Preceding Utterances}
"Target Event": {Target Event}
"Following Context": {Following Utterances}
```

We treat the belief prediction task as a text generation task where given the above input of context and a target event, we generate the belief label.

CG Prediction To predict CG, we use the heuristics-based approach we used in our previous work (Markowska et al., 2023). This rule-based approach maps a label for speaker and hearer belief to a CG label.[3]

Yes/No Question Answering Now that we have the belief and CG for both discourse participants, we apply the same heuristics we used to determine the query answers for the gold-annotated case, as described in Section 3.

4.5 Results

All results are evaluated using accuracy. Specifically, we report results on total accuracy and per-order accuracy. All of these metrics help us quantify to what extent LMs capture the full mental states of others, up to third-order mental states. Results for our experiments are shown in Table 3. All of our models perform worse than human performance, but better than the random baseline. Humans perform best on first order beliefs, but performance decreases on higher orders as has been observed in similar studies (Valle et al., 2015).

[3] See Appendix B for details on this algorithm.

Zero-Shot GPT-4 performs the best on all metrics. Mistral-7B-Instruct, despite being only 7B parameters, performs better than the 175B parameter GPT-3.5 in a zero-shot setting. We notice a distinct trend: all models capture first order beliefs the best, and decrease in performance as the order increases.

Fine-tuning Compared to the best zero-shot performance from GPT-4, our results show an increase in total accuracy, second order beliefs, and third order beliefs. However, we perform slightly worse in modeling first order beliefs.

ReCoG ReCoG, which explicitly models beliefs and CG among agents in dialog, outperforms every other system. We see a clear boost in performance in all metrics, and more notably similar results among all orders of belief.

Discussion There are some interesting trends in these results. All the LM-only models see a decrease in performance as the order of the query increases, just as humans do, though this decrease is not as large as is observed for humans. This might suggest that the models are in some ways "human-like" in their mistakes. However, upon closer inspection we see further differences.

Since the same proposition is asked about for multiple orders of belief, we can look at how often the model gets all answers for a proposition correct. Even though fine-tuning Mistral only provides a roughly 3 point boost to overall accuracy, this consistency in answers changes dramatically with Mistral, going from getting all questions for a proposition correct 20% of the time in the zero-shot setting to 61% of the time when fine-tuned. Our human sample is too small to perform a similar comparison but Kim et al. (2023) observed human consistency to be in the high 80s for a similar task on their synthetic benchmark.

We can also look at how correctness for first order belief correlates with correctness of higher order beliefs. Again we see fine-tuning makes a larger difference than accuracy suggests. The correlation between per-proposition 1st and 2nd order accuracy goes from $r = 0.43$ for the zero-shot model to $r = 0.90$ for the fine-tuned model. An even larger difference is observed between 1st and 3rd order ($r = 0.20$ to $r = 0.93$).

This all suggests that the zero-shot LM behavior is markedly different from the fine-tuned model, and from humans. Even when making correct predictions, they do so using a very different pattern.

5 Conclusion

We present a new corpus for testing theory of mind (ToM) capabilities, COMMON-TOM. Unlike previous ToM corpora, COMMON-TOM uses naturally occurring linguistic data and is not based on agents perceiving certain information or not. We explicitly model belief and CG to capture ToM in a manner directly inspired by cognitive science literature. While LMs are rapidly evolving and scaling in size and capabilities, they still lack a conceptual understanding of beliefs and CG in dialog. Therefore, for an LM to interact naturally it is necessary to explicitly model belief, CG, and cognitive states.

Limitations

While we present preliminary results on COMMON-TOM, the first ToM corpus to use natural spoken conversation, the original CG corpus is relatively small, containing only four dialogs. Furthermore, as these dialogs only contain conversations in English, our benchmark does not involve tracking ToM in other languages. While the benchmark tests ToM reasoning in a variety of ways, it is by no means comprehensive. In cognitive science, ToM is divided into cognitive (e.g. beliefs, thoughts) and affective (e.g. emotions, desires) with some evidence for fairly independent processes which operate over these two domains (Kalbe et al., 2010). We evaluate the cognitive aspect of belief and plan to tackle additional areas, particularly affective, in future work.

Ethical Considerations

As with other work on ToM, we risk the misinterpretation that AI models may be anthromorphized as having near-human level cognition. We stress that our work instead shows that as of now, existing models in fact do poorly at demonstrating ToM when presented with natural conversations. We did not require annotators for the creation of COMMON-TOM as our corpus was derived heuristically from the common ground (CG) corpus. However, our human baseline was done in-house by trained graduate students who were paid.

Acknowledgements

This material is based upon work supported in part by the National Science Foundation (NSF) under No. 2125295 (NRT-HDR: Detecting and Addressing Bias in Data, Humans, and Institutions); by funding from the Defense Advanced Research Projects Agency (DARPA) under the CCU (No. HR001120C0037, PR No. HR0011154158, No. HR001122C0034) and INCAS (HR01121C0186, No. HR001120C0037, and PR No. HR0011154158) programs; as well as by the Intelligence Advanced Research Projects Activity (IARPA) under the HIATUS program (contract 2022-22072200005). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF, DARPA, or IARPA.

We thank both the Institute for Advanced Computational Science and the Institute for AI-Driven Discovery and Innovation at Stony Brook for access to the computing resources needed for this work. These resources were made possible by NSF grant No. 1531492 (SeaWulf HPC cluster maintained by Research Computing and Cyberinfrastructure) and NSF grant No. 1919752 (Major Research Infrastructure program), respectively.

We thank our ARR reviewers, whose comments have contributed to improving the paper.

References

Cristian-Paul Bara, Sky CH-Wang, and Joyce Chai. 2021. MindCraft: Theory of mind modeling for situated dialogue in collaborative tasks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1112–1125, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Simon Baron-Cohen, Alan M. Leslie, and Uta Frith. 1985. Does the autistic child have a "theory of mind"? *Cognition*, 21(1):37–46.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Susan Brennan and Herbert H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of experimental psychology. Learning, memory, and cognition*, 22 6:1482–93.

Sarah Brown-Schmidt and Melissa C Duff. 2016. Memory and common ground processes in language use. *Topics in Cognitive Science*, 8(4):722–736.

Herbert H. Clark. 1996. *Using Language*. 'Using' Linguistic Books. Cambridge University Press.

Marie-Catherine de Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The CommitmentBank: Investigating projection in naturally occurring discourse. In *Proceedings of Sinn Und Bedeutung*, volume 23, pages 107–124.

Regine Eckardt. 2016. Questions on the table. *Unpublished manuscript.* https://www.researchgate.net/publication/304304651_Questions_on_the_Table.

Donka F Farkas and Kim B Bruce. 2010. On reacting to assertions and polar questions. *Journal of semantics*, 27(1):81–118.

Alexia Galati and Susan E. Brennan. 2021. What is retained about common ground? distinct effects of linguistic and visual co-presence. *Cognition*, 215:104809.

William S Horton and Susan E Brennan. 2016. The role of metarepresentation in the production and resolution of referring expressions. *Frontiers in psychology*, 7:1111.

William S. Horton and Richard J. Gerrig. 2016. Revisiting the memory-based processing approach to common ground. *Topics in Cognitive Science*, 8(4):780–795.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Elke Kalbe, Marius Schlegel, Alexander Thomas Sack, Dennis A. Nowak, Manuel Dafotakis, Christopher Bangard, Matthias Brand, Simone G. Shamay-Tsoory, Oezguer A Onur, and Josef Kessler. 2010. Dissociating cognitive from affective theory of mind: A tms study. *Cortex*, 46:769–780.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. FANToM: A benchmark for stress-testing machine theory of mind in interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14397–14413, Singapore. Association for Computational Linguistics.

Michal Kosinski. 2023. Theory of mind might have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*.

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, Hong Kong, China. Association for Computational Linguistics.

Ziqiao Ma, Jacob Sansom, Run Peng, and Joyce Chai. 2023. Towards a holistic landscape of situated theory of mind in large language models.

Magdalena Markowska, Mohammad Taghizadeh, Adil Soubki, Seyed Mirroshandel, and Owen Rambow. 2023. Finding common ground: Annotating and predicting common ground in spoken conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8221–8233, Singapore. Association for Computational Linguistics.

Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Tom Griffiths. 2018. Evaluating theory of mind in question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2392–2400, Brussels, Belgium. Association for Computational Linguistics.

OpenAI. 2022. ChatGPT: Optimizing language models for dialogue. `https://openai.com/blog/chatgpt`.

R OpenAI. 2023. Gpt-4 technical report. *arXiv*, pages 2303–08774.

The pandas development team. 2020. pandas-dev/pandas: Pandas.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526.

Roser Saurí and James Pustejovsky. 2009. FactBank: a corpus annotated with event factuality. *Language Resources and Evaluation*, 43:227–268.

Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. 2023. Minding language models' (lack of) theory of mind: A plug-and-play multi-character belief tracker. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13960–13980, Toronto, Canada. Association for Computational Linguistics.

Damien Sileo and Antoine Lernould. 2023. MindGames: Targeting theory of mind in large language models with dynamic epistemic modal logic. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4570–4577, Singapore. Association for Computational Linguistics.

Adil Soubki, Owen Rambow, and Chong Kang. 2022. KOJAK: A new corpus for studying German discourse particle ja. In *Proceedings of the 3rd Workshop on Computational Approaches to Discourse*, pages 1–6, Gyeongju, Republic of Korea and Online. International Conference on Computational Linguistics.

Robert C. Stalnaker. 2002. Common ground. *Linguistics and Philosophy*, 25(5-6):701–721.

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

Annalisa Valle, Davide Massaro, Ilaria Castelli, and Antonella Marchetti. 2015. Theory of mind development in adolescence and early adulthood: The growing complexity of recursive thinking ability. *Europe's Journal of Psychology*, 11(1):112–124.

Deanna Wilkes-Gibbs and Herbert H Clark. 1992. Coordinating beliefs in conversation. *Journal of Memory and Language*, 31(2):183–194.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Corpus Details

COMMON-TOM builds on (Markowska et al., 2023), which uses the English-only LDC CALLHOME corpus as a base. We show counts for our train/test split in Table 2. The training split contains dialogs 4245, 4248, and 4310 while the test split contains dialog 4431.

For convenience, we summarize the label meanings in Table 4 and the distribution of labels for COMMON-TOM in Table 5.

Label	**Description**
CT+	A speaker certainly believes that e.
PS	A speaker possibly believes that e.
CT-	A speaker certainly believes that not e.
NB	A speaker expresses no belief about e.
JA	An event e is mutually believed by both interlocutors and is added to CG in the moment e was uttered.
IN	An event e has already been a part of the interlocutors' CGs before the time of the event.
RT	An event e that has been presented by a speaker has been entertained but rejected by the addressee
NA	An event e that has no annotation for CG because it has not yet been introduced.

Table 4: Summary of annotation label meanings as described in (Markowska et al., 2023).

Bel$_{\text{A}}$	Bel$_{\text{B}}$	CG$_{\text{A}}$	CG$_{\text{B}}$	Count
CT-	CT-	RT	RT	1,746
CT+	CT+	JA	JA	1,632
CT+	CT+	IN	IN	1,494
PS	PS	JA	JA	1,152
PS	CT+	JA	JA	180
PS	CT+	NA	NA	144
CT-	NB	NA	NA	90
PS	PS	NA	NA	90
NB	CT-	NA	NA	72
PS	CT-	NA	NA	72
CT+	NB	NA	NA	72
NB	CT+	NA	NA	54
PS	NB	NA	NA	54
CT+	CT-	RT	RT	54
CT+	PS	JA	JA	54
CT-	PS	NA	NA	54
CT-	PS	RT	RT	36
CT+	PS	NA	NA	36
NB	PS	NA	NA	36
PS	CT-	RT	RT	36
PS	PS	IN	IN	36
CT+	CT+	NA	NA	36
CT+	CT+	JA	IN	18
NB	CT+	JA	JA	18
CT-	PS	NA	JA	18
CT+	NB	RT	RT	18
CT+	CT-	IN	RT	18
CT-	CT+	RT	RT	18
CT-	CT+	NA	NA	18
PS	NB	RT	RT	18
Total				7,374

Table 5: Counts of belief labels included in the corpus.

B Heuristics for Common Ground Prediction using Beliefs

Within this strategy, we've employed the following rules: consistently updating the common ground for both speakers using these straightforward heuristics.

1) If **Bel(A) = CT-** or **Bel(B) = CT-**, then **CG = RT**.
2) If **Bel(A) = CT+** and **Bel(B) = CT+**, then **CG = JA** or **CG = IN**.
3) If **Bel(A) = PS** and **Bel(B) = CT+**, then **CG = JA(PS)** or **CG = IN**.
4) If **Bel(A) = CT+** and **Bel(B) = PS**, then **CG = JA(PS)** or **CG = IN**.
5) If **Bel(A) = NB** or **Bel(B) = NB**, then **CG = NULL**.

Rules 2-4 under-determine whether the belief is already in the CG or newly added. In this context, the crucial task is to determine whether the target event had already been present in the common ground of the speakers (i.e., **CG = IN**) or not (i.e, **CG = JA**).

C Experiment Details

All experiments besides our OpenAI experiments used our employer's GPU cluster. We performed experiments on a Tesla V100-SXM2 GPU. Compute jobs typically ranged from 20 minutes for zero-shot experiments to 16 hours for fine-tuning. We do not do any hyperparameter search or hyperparameter tuning.

Zero-Shot For our zero-shot experiments, we use OpenAI models gpt-3.5-0613 and gpt-4-0613, and the instruction tuned Mistral-7B-Instruct (Jiang et al., 2023). Our OpenAI experiments use the OpenAI API. We set temperature to 1.0. To perform zero-shot experiments on Mistral-7B-Instruct, we use HuggingFace transformers (Wolf et al., 2020) and use the same temperature hyperparameter of 1.0. We keep all other hyperparameters as default. Our prompt template is as follows:

```
You are a cautious assistant. You carefully
follow instructions. You are helpful and
harmless and you follow ethical guidelines
and promote positive behavior. Given a
conversation, answer a yes or no question
without providing any additional
information.

Conversation:
{context}

Question:
{question}
```

We perform experiments using both chain-of-thought (CoT) following (Kim et al., 2023) and without CoT. CoT experiments are performed by adding "let's think step by step" after the question. Our CoT experiments performed 1.3% worse on average than without CoT, so therefore we only report results without CoT.

Fine-tuning We fine-tune the base Mistral-7B-v0.1 model. We use LoRA (Hu et al., 2021) to make our model memory efficient with *r=32* and *alpha=64*. We fine-tune for 3 epochs and use a learning rate of 2e-5.

ReCoG We perform a standard fine-tuning approach of FLAN-T5-base model which has 250M parameters. We fine-tune for 12 epochs and use a learning rate of 3e-4. Our system is implemented using the PyTorch framework and Python programming language. FLAN-T5 model is captured from the HuggingFace transformers (Wolf et al., 2020). We also used Pandas data analysis library (pandas development team, 2020).

D Human Evaluation Annotators

Our human evaluation annotation was done in-house with two domain expert graduate students who volunteered to annotate. The annotators were given instructions to answer 30 randomly sampled yes/no questions as best as they can. Annotators received ten questions of each order. Reported accuracy is computed over all 60 questions.

E Query Creation Heuristics

```
def resolve_1st_order_yn_answer(qbel, ↩
    sbel):
    if qbel == sbel:
        return True
    elif qbel == "PS" and sbel == "CT+":
        return True
    else:
        return False
```

```
def resolve_2nd_order_yn_answer(qbel, ↩
    sbel1, sbel2, cg1, cg2):
    """
    This method resolves the truthiness of↩
         a second order belief question ↩
        with
    the following form based on the ↩
        annotations provided.

        Is it the case that {spkr1} ↩
            believes that
        {spkr2} believes that it is {qbel}↩
             true that {event}?

    Args:
        qbel (str): question belief (CT-, ↩
            PS, CT+).
        sbel1 (str): speaker 1 belief (CT↩
            -, PS, CT+, NB).
        sbel2 (str): speaker 2 belief (CT↩
            -, PS, CT+, NB).
        cg1 (str): speaker 1 common ground↩
             (JA, IN, RT, NA).
        cg2 (str): speaker 2 common ground↩
             (JA, IN, RT, NA).

    Returns:
        bool: True if the question is ↩
            correct given the annotations.
    """
    # Case 1: The question is positive ↩
        polarity.
    if qbel in ("PS", "CT+") and cg1 in ("↩
        JA", "IN") and (
        (qbel == sbel1) or ((qbel == "PS")↩
            and (sbel1 == "CT+"))
    ):
        return True
    # Case 2: The question is negative ↩
        polarity.
    elif qbel == "CT-" and cg1 == "RT" and↩
         qbel == sbel2: # XXX: sbel2 here↩
        !!!
        return True
    return False
```

```
def resolve_3rd_order_yn_answer(qbel, ↩
    sbel1, sbel2, cg1, cg2):
    """
    This method resolves the truthiness of↩
         a third order belief question ↩
        with
    the following form based on the ↩
        annotations provided.

        Is it the case that {spkr1} ↩
            believes that {spkr2} believes
        that {spkr1} believes it is {qbel}↩
             true that {event}?

    Args:
        qbel (str): question belief (CT-, ↩
            PS, CT+).
        sbel1 (str): speaker 1 belief (CT↩
            -, PS, CT+, NB).
        sbel2 (str): speaker 2 belief (CT↩
            -, PS, CT+, NB).
        cg1 (str): speaker 1 common ground↩
             (JA, IN, RT, NA).
        cg2 (str): speaker 2 common ground↩
             (JA, IN, RT, NA).

    Returns:
        bool: True if the question is ↩
            correct given the annotations.
    """
    if qbel in ("PS", "CT+") and cg1 in ("↩
        JA", "IN") and (
        (qbel == sbel1) or ((qbel == "PS")↩
            and (sbel1 == "CT+"))
    ):
        return True
    elif qbel == "CT-" and cg1 in ("RT", "↩
        NA") and qbel == sbel1: # NOTE: ↩
        sbel1 here!!!
        return True
    return False
```