

ToM in LLM is not ToM, but a Pragmatic Effect

Agnese Lombardi, Alessandro Lenci

University of Pisa
Via Santa Maria 36, Pisa
agnese.lombardi@phd.unipi.it, alessandro.lenci@unipi.it

Abstract

Instruction tuning has been shown to improve large language models' performance on pragmatic tasks, and recent work suggests that additional training can also enhance Theory of Mind (ToM)-like abilities. However, existing studies rarely examine how different alignment techniques and training data contribute to ToM-related behavior in language models. In this work, we investigate the respective roles of instruction tuning and preference learning in shaping pragmatic and ToM abilities. Using the LLaMA 3 8B architecture, we fine-tune models on either pragmatic or ToM-specific data and subsequently align them via Direct Preference Optimization. We evaluate the resulting models on benchmarks targeting both pragmatics and ToM. Our results show that pragmatic training can substantially improve ToM performance even without explicit belief-related supervision, and that instruction tuning plays a central role in model alignment. These findings clarify the relationship between pragmatics and ToM in large language models.

Keywords: theory of mind, pragmatics, instruction-tuning

1. Introduction

Instruction tuning, understood as the further refinement of language models to perform specific tasks, has been shown to be an effective technique for improving performance on pragmatic tasks (Wu et al., 2024; Ruis et al., 2023). From a theoretical perspective, this is particularly relevant given the nature of pragmatics itself. The domain of **pragmatics** encompasses a wide range of phenomena, including speech acts, implicature recognition, presupposition processing, the interpretation of figurative language, and the maintenance of discourse coherence and cohesion. These aspects are inherently context-sensitive, where context extends beyond the linguistic input to include extralinguistic factors such as communicative intentions, situational cues, and world knowledge (Bambini and Lecce, 2025).

Similarly, several studies report that additional training can lead to improvements in language models' abilities that extend beyond pragmatics (Cai et al., 2024; Ranaldi and Freitas, 2024; Puerto et al., 2025; Aw et al., 2023; Lu et al., 2025), including **Theory of Mind** (ToM).

ToM refers to the capacity to represent others' beliefs, intentions, and desires. It relies on mentalizing processes, that is, the ability to reflect on one's own mental states and to attribute distinct or shared mental states to others within a given context. More broadly, ToM involves recognizing that others may hold beliefs, perspectives, or knowledge that differ from one's own, and reasoning about those mental states in order to predict and interpret behavior.

Although ToM is often associated with pragmatics, the nature of their relationship remains controversial and actively debated (Bambini and Lecce, 2025). In this work, we support the view that the

two abilities only partially overlap, as suggested by evidence from psycholinguistics (Bosco et al., 2018; Kotila et al., 2025) and neurolinguistics (Schieche et al., 2025; Schieche and Uddén, 2025).

Given the nature of this partial overlap, it is important to clarify why further training of language models facilitates the emergence of ToM-like behaviors. In particular, studies reporting improvements in LLM performance following additional training (Lu et al., 2025) rarely examine the nature of the training data with respect to its pragmatic properties, nor do they explicitly consider how such properties relate to ToM.

Moreover, existing work does not directly address the distinction between different alignment techniques, specifically instruction tuning and preference optimization, and their respective contributions to ToM-related abilities.

In this work, we aim to clarify the roles of instruction tuning and preference learning in enhancing ToM capabilities, in comparison to their effects on pragmatic performance. We further investigate the quality and nature of the training data underlying these improvements by explicitly assessing whether pragmatic materials that do not contain belief-related information can nonetheless support improved performance on ToM tasks, without relying on explicit ToM supervision.

To this end, we adopt the Llama 3 8B base architecture and perform two supervised fine-tuning procedures: one using pragmatic data and one using ToM-specific data. Each model is subsequently aligned through preference learning using Direct Preference Optimization (DPO), while maintaining a clear distinction between pragmatic and ToM training materials across model variants.

We evaluate the resulting four models, together with the base model, on two benchmarks targeting

pragmatic and ToM abilities. Our results reveal a surprising contribution of pragmatic training to ToM performance and highlight the effectiveness of instruction tuning as an alignment strategy. These results suggest that the ToM-like behaviors observed in LLMs (Street et al., 2024; Kosinski, 2024a) may, at least in part, be driven by exposure to pragmatic material rather than reflecting advanced mentalizing capacities in the human sense. Rather than implementing a fully-fledged capacity for representing others’ mental states, LLMs may be leveraging advanced pragmatic reasoning skills, generalizing from patterns learned during training and benefiting from an enhanced ability to follow task-specific instructions. In light of the partial functional segregation between pragmatics and ToM, our findings are consistent with psycholinguistic evidence indicating that pragmatic competence may act as a driving force in the development of ToM abilities, rather than the reverse (Del Sette et al., 2025).¹

2. Related Works

Previous work has suggested that alignment with human beliefs, particularly through preference learning training, is a key mechanism underlying the emergence of ToM-like abilities in large language models (Street, 2024). Other studies have shown that larger, fine-tuned models encode mental state information more accurately than smaller ones, pointing either to a scaling effect at the level of internal representations (Bortoletto et al., 2024; Nguyen, 2025; Cui et al., 2024), or to the influence of instruction tuning (Aw et al., 2024), in-context learning or prompting techniques (Moghaddam and Honey, 2023).

Despite these findings, there is no consensus regarding which training or alignment technique is primarily responsible for increased ToM alignment. In this work, we aim to assess the contribution of instruction tuning by directly comparing it with preference-learning approaches.

Most prior studies in this area focus on reinforcement learning from human feedback (RLHF) (Christiano et al., 2017). Here, we instead adopt **Direct Preference Optimization** (DPO) (Rafailov et al., 2023), as it is simpler and computationally more efficient, and avoids the instability associated with reward-model training. DPO is an alignment algorithm that provides an alternative to RLHF by optimizing the policy directly via gradient ascent on the implicit RLHF objective, without requiring a separate reward model or explicit reinforcement learning.

To our knowledge, this is the first work to explicitly investigate whether pragmatic alignment and

ToM-like behavior in language models arise primarily from instruction tuning or from human feedback through preference-optimization techniques. Additionally, although the partial overlap between pragmatics and ToM is acknowledged in LLM evaluation (Hu et al., 2025), and several studies assess LLM performance on pragmatic abilities (Hu et al., 2023; Sravanthi et al., 2025; Ruis et al., 2023) and ToM tasks (Kosinski, 2024b; Ullman, 2023; Strachan et al., 2024), no previous work has systematically investigated the interaction between these two abilities during training.

3. Training

First, we performed two separate supervised fine-tuning (SFT) procedures on Llama 3 8B: one using pragmatic data and the other using ToM data. We then further trained each of these models using direct preference optimization (DPO) with pragmatic and ToM data. In total, this resulted in the evaluation of four model variants: SFT-Pragmatics, SFT-ToM, DPO-Pragmatics, and DPO-ToM.

3.1. SFT

We performed SFT using parameter-efficient fine-tuning with Low-Rank Adaptation (LoRA). All experiments were conducted using the Hugging Face `Transformers` and `TRL` libraries.

Two instruction-tuning conditions were considered:

- **ToM.** We used examples from the EXPLORE-ToM dataset, generated within the same framework (Sclar et al., 2024). ExploreToM is specifically designed to assess ToM abilities, as it comprises narratives in which different characters hold potentially divergent beliefs about the current state of the world and about one another’s beliefs. Each story is paired with questions intended to probe the model’s capacity to track and reason about these mental states. For instance, a story might include questions testing nested beliefs, such as whether “Anne does not know that Charles knows that the apple has been salted”, requiring the model to reason across multiple perspectives. From the training split, we used 6,000 instances for training and 5,000 for validation. The inputs consisted of structured or infilled story prompts, while the targets corresponded to the expected answers.
- **Pragmatics.** We combined two synthetic reasoning datasets: `argilla/sft`² and

¹Code for training and analysis: https://github.com/agneselombardi/ToM_is_not_ToM

²<https://huggingface.co/datasets/argilla/synthetic-concise-reasoning-sft>

lighteval/natural³. Both datasets include examples requiring concise reasoning that do not involve ToM, as illustrated by extended examples in Appendix A.1. The first dataset contributed 5,000 training examples, while from the second we selected 1,000 examples for training and 5,000 for validation. In all cases, questions were used as prompts and targets as completions.

All datasets were normalized to a unified format consisting of a *prompt* and a *completion*. During training, prompts and completions were concatenated into a single input sequence (prompt examples are reported in Appendix A.1).

We initialized the pretrained Llama3-8B model in half precision (FP16). Tokenization was performed using the corresponding Llama tokenizer, with the end-of-sequence token used as the padding token and right padding applied. Input sequences were truncated or padded to a maximum length of 2048 tokens.

LoRA was applied to both attention and feed-forward projection layers (`q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`) with rank 16, scaling factor 16, and no dropout. During training, only LoRA parameters were updated.

Instruction tuning was carried out using the SFT-Trainer. Training was conducted for three epochs with a learning rate of 2×10^{-4} , a per-device batch size of 2, and gradient accumulation over four steps. Evaluation was performed every 200 steps, and model checkpoints were saved every 600 steps, retaining the three most recent checkpoints. Early stopping with a patience of two evaluations was applied based on validation loss, and the best-performing checkpoint was selected at the end of training.

To stabilize training, we applied NEFTune noise with $\alpha = 5$. All experiments were logged using Weights & Biases.

3.2. DPO

Following instruction tuning, we further aligned the models using DPO. For each condition, DPO was applied to the corresponding instruction-tuned model, using preference datasets aligned with the same target domain (ToM or pragmatics).

In all experiments, the policy model consisted of the instruction-tuned Llama3-8B model with LoRA adapters, which remained trainable during alignment. Critically, we used the frozen SFT checkpoint from the previous stage as the reference model. Specifically, for ToM alignment, the ToM-SFT model

³https://huggingface.co/datasets/lighteval/synthetic_reasoning_natural

served as the reference; for pragmatic alignment, the Pragmatics-SFT model served as the reference. This design choice ensures that the KL penalty regularizes the policy toward the task-specific distribution learned during SFT, rather than toward a generic instruction-following distribution. All reference model parameters were kept fixed throughout training. Tokenization was performed using the Llama3 8B tokenizer, with the end-of-sequence token used as the padding token and right padding applied.

For ToM preference learning, we constructed a DPO training set by combining three publicly available preference datasets targeting social and mental-state reasoning⁴, as illustrated by examples in Appendix A.2. All datasets were normalized to a common format consisting of a *prompt*, a preferred (*chosen*) response, and a dispreferred (*rejected*) response. For dialogue-based datasets, user turns were used as prompts and assistant turns as candidate responses.

The combined dataset contained 14,369 examples total. We shuffled the data (seed 42) and split it into 90% training (12,932 examples) and 10% validation (1,437 examples). Training was conducted for up to three epochs using a learning rate of 5×10^{-6} , a per-device batch size of 1, and gradient accumulation over 32 steps (effective batch size of 32). We set the DPO temperature parameter to $\beta = 0.3$. The maximum prompt length was set to 1,024 tokens and the maximum sequence length to 2,048 tokens (matching SFT training). Training was conducted in bfloat16 precision with gradient checkpointing enabled using the AdamW optimizer with default hyperparameters.

To prevent overfitting, we evaluated on the validation set every 100 training steps and saved checkpoints at the same frequency, retaining the three most recent checkpoints. At the end of training, the checkpoint with the lowest validation loss was selected as the final model.

For pragmatic preference learning, we followed an analogous procedure using preference datasets focused on conversational and pragmatic quality. Specifically, we combined three different datasets⁵.

⁴<https://huggingface.co/datasets/Doctor-Shotgun/theory-of-mind-dpo> (539 examples), <https://huggingface.co/datasets/onyrotssih/social-i-qa-orpo-dpo-10k> (10200 examples) and https://huggingface.co/datasets/shayanfirouzan/SocialReasoning_DPO (3630 examples)

⁵<https://huggingface.co/datasets/flammenai/casual-conversation-DPO> (3730 examples); <https://huggingface.co/datasets/minchyeom/Thinker-DPO> (2670 examples); and https://huggingface.co/datasets/clembench-playpen/DPO_dialogue_1neg_old (6700 examples).

Each dataset was normalized to the standard prompt-chosen-rejected format. For multi-turn dialogue data (clembench dataset), conversations were flattened into single sequences by concatenating all turns in chronological order, with each turn’s content separated by newline characters.

The combined dataset contained 13,100 examples total, which we shuffled (seed 42) and split into 90% training (11,790 examples) and 10% validation (1,310 examples).

For both DPO conditions, we verified several implementation details to ensure correct training dynamics. First, we confirmed that gradients were applied exclusively to the policy model’s LoRA parameters while the reference model remained frozen. Second, we verified that the policy and reference models were properly initialized from the same SFT checkpoint by checking that their initial outputs were identical (mean absolute logit difference $< 10^{-6}$). Finally, we monitored that the models diverged during training, with the mean absolute logit difference increasing substantially by the end of training.

All experiments were conducted using a single NVIDIA A100 GPU with 40GB memory. The choice of $\beta = 0.3$ follows standard practices in the DPO literature, balancing the preference learning objective with KL regularization to prevent excessive deviation from the reference distribution. The sequence lengths (1,024/2,048 tokens) ensure that longer reasoning chains and multi-turn dialogues are not truncated during training.

4. Evaluation

We evaluated all model variants on two benchmarks targeting distinct but related abilities: **PUB** (Srivanthi et al., 2024), which explicitly assesses pragmatic reasoning, and **FANTOM** (Kim et al., 2023), which is specifically designed to evaluate ToM reasoning.

The PUB benchmark comprises 14 tasks, primarily targeting implicatures. Ten tasks provide a detailed assessment of models’ implicature-related abilities, while presupposition and reference are each evaluated through two tasks. Task 1 tests models’ ability to distinguish between direct and indirect responses using multiple-choice questions. Tasks 2 and 3 involve response classification: Task 2 categorizes indirect answers into five labels, whereas Task 3 extends this by incorporating implied meanings as additional cues. Task 4 focuses on implicature recovery in non-polar question-and-answer contexts. Task 5 (“Agreement Detection”) and Task 6 (“Understanding Sarcasm”) evaluate the model’s capacity to comprehend figurative language within dialogues. Tasks 7, 8, and 9 are based on the FLUTE dataset, presenting sentences

or premises expressed figuratively alongside corresponding hypotheses in simpler language. Task 10 assesses whether models can draw correct inferences when implicatures are involved. Finally, Tasks 11 and 12 evaluate presupposition understanding, either framed as natural language inference or by asking the model to judge whether a given presupposition is valid or invalid within the conversational context.

In contrast, the FANTOM benchmark is based on asymmetric conversational settings.⁶ In this setting, the full conversation is provided to the model, which thus has access to all exchanged information and effectively acts as an omniscient observer. However, a designated character (PersonX) is absent during part of the interaction, leading to an information asymmetry between the model and PersonX. Responses that attribute to PersonX information that was not available during their absence are taken as evidence of a failure to correctly model that character’s mental state.

For PUB evaluation we report F1 metric for each task, while for FANTOM evaluation we report F1 scores on the following metrics:

- **Belief (MCQA):** The model must choose between an omniscient belief and a PersonX-centric belief. The correct answer corresponds to the PersonX-centric belief.
- **Answerability List and Accessibility Info:** A response is considered correct only if it includes all characters who have access to the relevant answer or information, while excluding all characters who do not. Partial credit is not assigned.
- **Answerability Binary and Accessibility Binary:** The model must respond with *yes/true* for characters who have access to the answer or information, and with *no/false* for those who do not.
- **First Order:** Evaluation of first-order ToM reasoning.
- **Second Order (Overall):** Evaluation of second-order ToM reasoning. To further

⁶Small-talk conversations involving multiple characters, each centered around a topic (e.g., pets, risk-taking, personal growth). Each topic comprises several subtopics (e.g., the topic “pets” may include subtopics such as “breed” and “special moves”). Conversations begin with two or three characters, and as they progress, participants may join or leave while the subtopic evolves. During the absence of a character, information is exchanged among the remaining participants, creating a natural information asymmetry. When the absent character later rejoins the conversation, they lack access to the information shared in their absence.

analyze performance on second-order belief tasks, FANTOM distinguishes between **Second-Order Cyclic** and **Second-Order Acyclic** patterns. Cyclic second-order questions probe a character’s belief about another character’s belief concerning the first character (e.g., *What does Linda think about Kailey’s belief regarding the breed of Linda’s dog?*), whereas acyclic second-order questions focus on a character’s belief about another character’s belief concerning a third character (e.g., *What does David think about Kailey’s belief regarding the breed of Linda’s dog?*) (Kim et al., 2023).

5. Results

Figures 1 and 2 report a comparison of the three model variants on the PUB and FANTOM benchmarks, respectively. The PUB evaluation reveals substantial variability across tasks.

In Task 1, which focuses on **direct and indirect response classification**, both DPO and SFT pragmatic-based training achieve the best performance. A similar trend is observed in Task 2 and 3 which address response classification without and with implied meaning.

Task 4, targeting **implicature recovery**, shows a clear advantage for pragmatics-trained models over their ToM-trained counterparts, despite the fact that implicature resolution may involve mentalizing abilities. Task 5 (**agreement detection**) does not exhibit meaningful differences across model variants, including relative to the base model. In Task 6, which evaluates **sarcasm understanding**, pragmatics-trained models again display substantial improvements, even though sarcasm interpretation is often assumed to rely on ToM.

A similar pattern emerges in Tasks 7–9, which assess **figurative language understanding**. In all three cases, pragmatics-trained models outperform the ToM-trained versions, despite the fact that these tasks plausibly require ToM reasoning.

Tasks 10 and 11, addressing **implicature NLI** and **presupposition NLI** respectively, do not explicitly require mentalizing abilities. Accordingly, the best performance is achieved by the pragmatics-trained SFT model. Interestingly, Task 12, which focuses on **presupposition in QA**, does not show substantial improvements from either SFT or DPO. In Task 13 (**deictic QA**), the pragmatics-trained SFT model again performs best. Finally, Task 14, concerning **reference via metonymy**, shows a decrease in accuracy for both SFT and DPO models relative to the base model. This task appears to rely more on extralinguistic knowledge than on inferential reasoning.

Notably, the **SFT ToM-trained model** exhibits

Model	Data	Wins	F1 (M / Md)
SFT	ToM	4	0.363 / 0.334
SFT	Pragmatics	11	0.493 / 0.455
DPO	ToM	4	0.398 / 0.357
DPO	Pragmatics	12	0.491 / 0.435

Table 1: Performance of SFT and DPO aligned models on PUB compared to the base pretrained model. The *Wins* column indicates the number of tasks in which the model outperforms the reference model (in this case, the base model). The *Score* metric reports both the mean F1 and the median F1 achieved by the SFT or DPO model.

performance degradation on several tasks, particularly those involving implied meaning (Task 3), implicature recovery (Task 4), sarcasm (Task 6), and figurative language understanding (Tasks 7 and 8, with Task 9 showing performance comparable to the base model). Similar degradation is also observed for presupposition and metonymy tasks ⁷.

Table 1 further summarizes these findings by reporting the number of tasks in which each model outperforms the base model, together with mean and median F1 scores. Overall, **SFT and DPO with pragmatics-oriented training data yields the strongest improvements over the base model**.

Table 2 confirms that pragmatics-trained models consistently outperform ToM-trained models across tasks, with SFT (pragmatic) achieving the largest gains, while ToM-specific training shows limited or inconsistent effects.

Figure 2 reports the performance of all model variants on the FANTOM benchmark. The first task, which evaluates **belief interpretation**, shows the SFT model trained on ToM data substantially outperforming all other variants, whereas the DPO-trained models, particularly the DPO–ToM version, exhibit the weakest performance.

For the **Answerability List** and **Accessibility List** metrics, DPO pragmatic-based yields the best results, followed by SFT training on pragmatic material. Notably, for the Answerability List task, only the pragmatic-based variants yield a slight improvement over the base model, while the corresponding **Answerability Binary** metric benefits more clearly from ToM-based training, likely due to the binary response format. In contrast, for **Accessibility**, ToM-based models do not show meaningful gains in either formulation. In the binary version, pragmatic training proves to be the most effective, a result that contrasts with the pattern observed for

⁷In some cases, such as figurative language (Task 7), implicature NLI, presupposition, and metonymy, the DPO-trained model also underperforms relative to the base model.

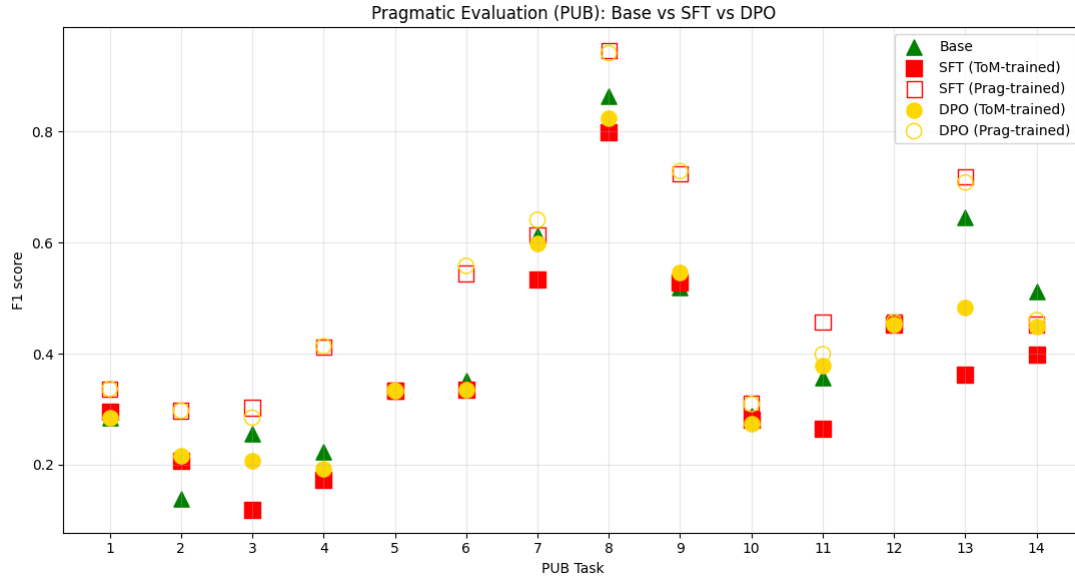


Figure 1: Results on the PUB benchmark for the base model, SFT, and DPO models.

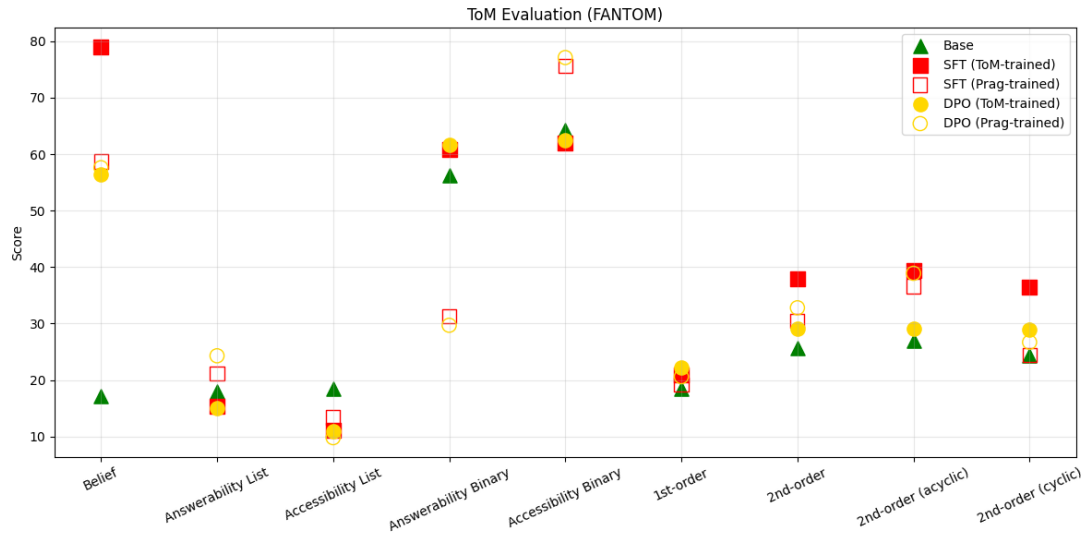


Figure 2: Results on the FANTOM benchmark for the base model, SFT, and DPO models.

Model / Task	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14
SFT (ToM) vs Base	3.5	48.9	-53.5	-22.9	0.0	-4.5	-13.0	-7.5	1.5	-2.5	-26.0	-0.9	-43.9	-22.2
SFT (Prag) vs Base	18.1	113.7	18.2	84.1	0.0	55.3	0.0	9.5	39.1	8.0	28.0	0.0	11.2	-11.7
DPO (ToM) vs Base	30.2	92.4	27.6	2.7	0.0	-1.3	-7.4	1.1	7.3	-12.2	-4.7	-0.2	-1.7	-11.0
DPO (Prag) vs Base	18.0	67.3	12.0	74.1	0.0	64.3	7.4	9.3	42.5	1.5	1.7	0.0	-2.3	-7.9

Table 2: Per-task percentage change in F1 score relative to the base model on PUB.

Answerability under the same task format.

For both **first-order** and **second-order** ToM reasoning, the ToM-trained SFT model achieves the strongest performance, followed by both DPO models. Surprisingly, in all ToM-related metrics, the DPO–ToM model underperforms the SFT–ToM.

Table 3 corroborates these observations by summarizing the number of metrics in which each model

improves over the base version, together with their mean and median scores. Overall, **DPO with pragmatics-oriented training data emerges as the most effective alignment strategy on FANTOM**. Both SFT variants also lead to improvements, with a slight advantage for ToM-based SFT training.

Table 4 confirms these trends: SFT (ToM) produces the largest improvements on first-order be-

Model	Data	Wins	Score (M / Md)
SFT	ToM	6	40.31 / 37.90
SFT	Pragmatics	6	34.54 / 30.50
DPO	ToM	6	35.10 / 29.10
DPO	Pragmatics	7	35.28 / 29.70

Table 3: Performance of SFT- and DPO-aligned models on the FANTOM benchmark compared to the base pretrained model. The *Wins* column indicates the number of tasks in which the model outperforms the reference model (in this case, the base model). The *Score* metric reports both the mean F1 and the median F1 achieved by the SFT or DPO model.

lief reasoning (M1), DPO (pragmatics) yields stable gains on later metrics (M7–M9) involving second-order cyclic and acyclic reasoning, and SFT (pragmatics) shows mixed results. Overall, SFT excels on core ToM tasks, whereas DPO provides more reliable performance on higher-order, multi-agent reasoning.

5.1. SFT vs. DPO

Tables 5 and 6 compare the effect of DPO alignment if compared to instruction tuning. In this case we want to see for ToM tasks which training is the most effective and if model’s correct answers on ToM data were determined from the ability of the model to follow instruction and to better understand the task or if they derive from preference learning alignment.

Results on the PUB dataset indicate that DPO is more effective than SFT only when training is performed with ToM data. In contrast, for pragmatic training, DPO does not produce a significant advantage over SFT.

Results on the FANTOM dataset instead show that DPO trained on ToM data is never effective for ToM tasks, whereas DPO applied to pragmatic training yields better results. Moreover, the comparison within the ToM setting shows that the SFT–ToM model consistently outperforms the DPO–ToM variant, indicating that preference learning is effective only when applied to pragmatic content.

5.2. Pragmatics vs. ToM

Tables 7 and 8 compare pragmatics-based and ToM-based training across the two evaluation benchmarks.

On the PUB benchmark, results clearly indicate that pragmatic training is more effective than ToM training across both alignment conditions. On FANTOM, a different pattern emerges. ToM-oriented data prove more effective only in the SFT condition, whereas pragmatic training yields superior

performance for DPO-aligned models. It is clear that **ToM-specific training does not appear to provide a substantial advantage over pragmatic training** in enhancing ToM abilities.

6. Discussion

We can finally provide answers to the two research questions: i.) Is preference learning more effective when compared to instruction tuning? ii.) Is pragmatic training content sufficient to explain patterns observed in previous experiments?

Our results support the idea that what we observe in models might be the effect of instruction tuning rather than human alignment. Indeed, our first finding is that DPO yields better results when compared to SFT only in PUB evaluation (both DPO models) and in FANTOM if trained with pragmatic data. This suggests that **preference learning helps the model to solve pragmatic tasks, but not ToM tasks.**

Consequently, it is possible that what we observe in these models is not a genuine general capacity of reasoning or reproducing mentalizing processes, but instead an ability to learn how to perform specific tasks and to generalize this competence to closely related tasks. This phenomenon is also observed on the FANTOM benchmark, which is explicitly designed to isolate ToM abilities. Therefore, a more general methodological issue emerges, raising the question of how closely what we observe and evaluate in language models truly corresponds to the cognitive abilities we intend to test (Hu et al., 2025).

Our second finding suggests that training on pragmatic data is more effective at improving model performance than training on ToM data. This observation supports two points widely discussed in the psycholinguistic literature: i.) the necessity and difficulty of isolating ToM reasoning from pragmatic reasoning, and ii.) the dissociation between pragmatics and ToM in language models. We observed that models tend to dissociate these abilities at the output level, exhibiting stronger performance on pragmatic tasks and weaker performance on ToM tasks. We hypothesize that this is because models do not possess genuine ToM abilities; rather, what we observe in ToM evaluations is the manifestation of certain pragmatic inferences that emerge in some ToM tests.

This interpretation is further supported by the finding that pragmatic alignment benefits both ToM and pragmatic tasks. Therefore, what appears to be ToM reasoning in the models may instead reflect the generalization of specific pragmatic competencies. In this sense, our results align with Grice’s view that ToM reasoning can be considered an enrichment that follows pragmatic interpretation (Grice,

Model / M	M1	M2	M3	M4	M5	M6	M7	M8	M9
SFT (ToM) vs Base	61.8	-2.5	-7.3	4.6	-2.3	2.5	12.3	12.5	12.0
SFT (Prag) vs Base	41.6	3.3	-5.0	-24.9	11.3	0.8	4.9	9.7	0.0
DPO (ToM) vs Base	33.5	-3.1	-9.4	3.8	-3.9	-5.2	-4.2	-1.2	-7.4
DPO (Prag) vs Base	37.6	3.6	-5.1	-34.3	13.5	10.9	22.8	29.7	15.9

Table 4: Per-metric score differences on FANTOM across training conditions.

Comp.	W	F1 M	F1 Md
DPO(ToM) vs SFT(ToM)	10	0.398	0.357
DPO(Prag.) vs SFT(Prag.)	7	0.491	0.435

Table 5: DPO vs SFT on PUB. The W column indicates the number of tasks in which the DPO model outperforms the SFT model. The mean F1 ($F1 M$) and the median F1 ($F1 Md$) refers to the DPO model.

Comp.	W	M	Md
DPO(ToM) vs SFT(ToM)	3	35.10	29.10
DPO(Prag.) vs SFT(Prag.)	6	35.28	29.70

Table 6: DPO vs SFT on FANTOM. The W column indicates the number of tasks in which the DPO model outperforms the SFT model. The mean F1 ($F1 M$) and the median F1 ($F1 Md$) refers to the DPO model.

1967). The models may then only perform the initial pragmatic step, and it is precisely this pragmatic information that produces apparent success on some ToM tasks.

7. Conclusion

In this work, we demonstrated that pragmatics can play a crucial role in eliciting ToM-like behaviour in LLMs. We also observed that preference learning appears to contribute differently to performance on ToM tasks compared to pragmatic tasks, and that SFT may be sufficient to improve performance on certain ToM evaluations. Although we did not construct the training and evaluation datasets our-

Comp.	W	F1 M	F1 Md
SFT (Prag.) vs SFT (ToM)	13	0.493	0.455
DPO (Prag.) vs DPO (ToM)	13	0.491	0.435

Table 7: Cross-training comparisons between pragmatic and ToM supervision on PUB. The W column indicates the number of tasks in which models trained on pragmatic material outperforms models trained on ToM material. The mean F1 ($F1 M$) and the median F1 ($F1 Md$) refers to SFT and DPO models with pragmatic training.

Comp.	W	M	Md
SFT (Prag.) vs SFT (ToM)	3	34.54	30.50
DPO (Prag.) vs DPO (ToM)	5	35.28	29.70

Table 8: Cross-training comparisons between pragmatic and ToM supervision on FANTOM. The W column indicates the number of tasks in which models trained on pragmatic material outperforms models trained on ToM material. The mean F1 ($F1 M$) and the median F1 ($F1 Md$) refers to SFT and DPO models with pragmatic training.

elves, and although one potential limitation of this study is that pragmatic datasets may already embed ToM components, our findings nonetheless open the way to more fine-grained investigations into the role of specific linguistic features in training data. In addition, we did not systematically explore the effect of different dataset sizes for SFT and direct preference optimization DPO, which represents another avenue for future work. Future research could extend this analysis to pretraining data, enabling a more comprehensive understanding of how different stages of training contribute to ToM-related behaviour. More broadly, our results underscore the importance of disentangling ToM abilities from other concurrent capacities, such as pragmatic reasoning, when designing and interpreting ToM evaluations. This distinction is essential if we aim to make robust and theoretically grounded claims about the presence or extent of ToM abilities in large language models.

8. Limitation

While the EXPLORE-ToM dataset used for SFT is theory-driven and explicitly designed to target ToM, the composition of the DPO materials presents a more complex landscape.

Regarding the DPO datasets used for ToM optimization, we selected three datasets and manually inspected them to verify that their responses were grounded either in the recognition of others' beliefs or in the recognition of feelings and emotions. On this basis, we classified them as explicitly requiring ToM reasoning, ensuring that the ToM DPO condition was indeed targeting mentalizing processes.

Moreover, since the DPO datasets were not

custom-built for this study, we cannot definitively rule out the presence of latent ToM requirements in every entry. However, we do not view this as a primary pitfall. The central contribution of this work is demonstrating that non-ToM material enhances ToM elaboration more effectively than ToM-specific material itself. Consequently, any minor ToM overlap within the DPO set is acknowledged as a limitation rather than a structural failure.

Conversely, the inclusion of explicit ToM reasoning in our pragmatic datasets is highly improbable due to the following factors:

SFT Constraints: These datasets focus on concise, functional reasoning that lacks the recursive mentalizing characteristic of ToM.

DPO Domain Specificity: The datasets were selected based on three criteria that prioritize pragmatic coherence over mental state attribution: casual conversation (targets textual coherence rather than perspective-taking); mathematical reasoning (focuses on contextual interpretation and adherence to communicative norms) and game dialogue (while interactive, these entries require adherence to historical dialogue flow rather than an analysis of the player's internal perspective).

Isolating ToM is inherently difficult, as evidenced by the specialized framework required for EXPLORE-ToM. We define our pragmatic datasets as "pragmatic-only" because they lack explicit mentalizing prompts or reasoning elicitation. While some entries may involve phenomena that sit at the boundary of mentalizing (such as figurative language), this ambiguity is an intrinsic characteristic of the partial overlapping nature of pragmatics and ToM.

9. Bibliographical References

- Khai Loong Aw, Syrielle Montariol, Badr AlKhamissi, Martin Schrimpf, and Antoine Bosselut. 2023. [Instruction-tuning aligns llms to the human brain](#). *ArXiv*, abs/2312.00575.
- Khai Loong Aw, Syrielle Montariol, Badr AlKhamissi, Martin Schrimpf, and Antoine Bosselut. 2024. [Instruction-tuning aligns llms to the human brain](#).
- Valentina Bambini and Serena Lecce. 2025. [At the heart of human communication: new views on the complex relationship between pragmatics and theory of mind](#). *Philosophical Transactions of the Royal Society B: Biological Sciences*, 380(1932):20230486.
- Matteo Bortoletto, Constantin Ruhdorfer, Adnen Abdessaied, Lei Shi, and Andreas Bulling. 2024. [Limits of theory of mind modelling in dialogue-based collaborative plan acquisition](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4856–4871, Bangkok, Thailand. Association for Computational Linguistics.
- Francesca M. Bosco, Maurizio Tirassa, and Ilaria Gabbatore. 2018. [Why pragmatics and theory of mind do not \(completely\) overlap](#). *Frontiers in Psychology*, 9.
- Huanqia Cai, Yijun Yang, and Zhifeng Li. 2024. [System-2 mathematical reasoning via enriched instruction tuning](#).
- Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17*, page 4302–4310, Red Hook, NY, USA. Curran Associates Inc.
- Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024. [Ultrafeedback: boosting language models with scaled ai feedback](#). In *Proceedings of the 41st International Conference on Machine Learning, ICML'24*. JMLR.org.
- Paola Del Sette, Valentina Bambini, Elena Tonini, and Serena Lecce. 2025. [Are theory of mind and metaphor comprehension causally related? a training study in middle childhood](#). *Language Acquisition*, 32(4):432–452.
- Paul H. Grice. 1967. The logic of conversation. *William James Lectures*.
- Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. [A fine-grained comparison of pragmatic language understanding in humans and language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213, Toronto, Canada. Association for Computational Linguistics.
- Jennifer Hu, Felix Sosa, and Tomer Ullman. 2025. [Re-evaluating theory of mind evaluation in large language models](#). *Phil. Trans. R. Soc.*
- 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. 2024a. [Evaluating large language models in theory of mind tasks](#). *Proceedings of the National Academy of Sciences of the United States of America*, 121(45):e2405460121.
- Michal Kosinski. 2024b. [Evaluating large language models in theory of mind tasks](#). *Proceedings of the National Academy of Sciences of the United States of America*, 121(45):e2405460121.
- A. Kotila, L. Mäkinen, E. Leinonen, and S. Loukusa. 2025. [Interrelatedness of false-belief reasoning, structural language, and pragmatic communication in typically developing 4- to 6-year-old children](#). *First Language*, 45(1):47–65. Online first published 2024.
- Yi-Long Lu, Chunhui Zhang, Jiajun Song, Lifeng Fan, and Wei Wang. 2025. [Do theory of mind benchmarks need explicit human-like reasoning in language models?](#)
- Shima Rahimi Moghaddam and Christopher J. Honey. 2023. [Boosting theory-of-mind performance in large language models via prompting](#).
- Hieu Minh "Jord" Nguyen. 2025. [A survey of theory of mind in large language models: Evaluations, representations, and safety risks](#).
- Haritz Puerto, Tilek Chubakov, Xiaodan Zhu, Harish Tayyar Madabushi, and Iryna Gurevych. 2025. [Fine-tuning on diverse reasoning chains drives within-inference CoT refinement in LLMs](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3789–3808, Vienna, Austria. Association for Computational Linguistics.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. [Direct preference optimization: your language model is secretly a reward model](#). In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23*, Red Hook, NY, USA. Curran Associates Inc.
- Leonardo Ranaldi and Andre Freitas. 2024. [Self-refine instruction-tuning for aligning reasoning in language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, page 2325–2347. Association for Computational Linguistics.
- Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2023. [The goldilocks of pragmatic understanding: fine-tuning strategy matters for implicature resolution by llms](#). In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23*, Red Hook, NY, USA. Curran Associates Inc.
- Christoffer Forbes Schieche, Manu Mahal, William Hedley Thompson, and Julia Uddén. 2025. [Pragmatics partially segregated from theory of mind: evidence from resting-state functional connectivity](#). *Philos Trans R Soc Lond B Biol Sci*.
- Christoffer Forbes Schieche and Julia Uddén. 2025. [Neural adolescent pragmatic development mirrors pragmatic differences in adulthood: an fmri study](#). *bioRxiv*. Preprint.
- Melanie Sclar, Jane Yu, Maryam Fazel-Zarandi, Yulia Tsvetkov, Yonatan Bisk, Yejin Choi, and Asli Celikyilmaz. 2024. [Explore theory of mind: program-guided adversarial data generation for theory of mind reasoning](#). In *The Thirteenth International Conference on Learning Representations*.
- Settaluri Lakshmi Sravanthi, Meet Doshi, Tankala Pavan Kalyan, Rudra Murthy, Pushpak Bhattacharyya, and Raj Dabre. 2024. [Pub: A pragmatics understanding benchmark for assessing llms' pragmatics capabilities](#).
- Settaluri Lakshmi Sravanthi, Kishan Maharaj, Sravani Gunnu, Abhijit Mishra, and Pushpak Bhattacharyya. 2025. [Understand the implication: Learning to think for pragmatic understanding](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23778–23790, Vienna, Austria. Association for Computational Linguistics.
- James W. A. Strachan, Davide Albergo, Gabriele Borghini, et al. 2024. [Testing theory of mind in large language models and humans](#). *Nature Human Behaviour*, 8:1285–1295.
- Winnie Street. 2024. [Llm theory of mind and alignment: Opportunities and risks](#).
- Winnie Street, John Oliver Siy, Geoff Keeling, Adrien Baranes, Benjamin Barnett, Michael McKibben, Tatenda Kanyere, Alison Lentz, Blaise Aguera y Arcas, and Robin I. M. Dunbar. 2024. [Llms achieve adult human performance on higher-order theory of mind tasks](#).
- Tomer Ullman. 2023. [Large language models fail on trivial alterations to theory-of-mind tasks](#).
- Shengguang Wu, Shusheng Yang, Zhenglun Chen, and Qi Su. 2024. [Rethinking pragmatics in large language models: Towards open-ended evaluation and preference tuning](#). In *Proceedings of the*

2024 Conference on Empirical Methods in Natural Language Processing, pages 22583–22599, Miami, Florida, USA. Association for Computational Linguistics.

A. Appendix

A.1. SFT: Dataset and Prompts

Example of a prompt from the dataset used in SFT training that involves ToM reasoning:

EXPLORE Prompt: The sun shone through the large glass doors of the hotel lobby, illuminating the marble floor and casting a warm glow over the comfortable seating areas. Soft music filled the air, mingling with the gentle hum of conversation and the occasional chime of the elevators in the bustling hotel. As Kaylee entered the hotel lobby her eyes quickly scanned the space, taking in every detail to ensure everything was in order for the upcoming event. Kaylee’s task of tidying the lobby extended to the small, silver item, which she carefully stowed away in the desk drawer, and Liam, observing from across the room, felt his interest in the object grow, his mind racing with questions about its significance and purpose. With her tasks in the lobby complete, Kaylee stepped out into the fresh air, the sounds of the bustling hotel lobby fading into the background as the glass doors slid shut behind her. Liam moved across the lobby floor, his footsteps silent on the marble as he walked towards the empty reception desk where the mystery item was now hidden. Moments later, Kaylee stepped back through the glass doors of the hotel, joining Liam in the lobby once again. Liam’s interest in the silver item led him to reposition it, now resting snugly within the hotel lobby’s nearby leather briefcase. Meanwhile, Kaylee began to run a final check on the lobby, seemingly unconcerned by the briefcase’s new contents. In which container was the silver letter opener at the beginning? **Completion:** wooden desk drawer

Examples of pragmatic reasoning prompts from datasets used in SFT training:

Argilla/SFT Prompt: What is the primary function of mitochondria within a cell? The primary function of mitochondria is to produce energy for the cell through a process called cellular respiration. **Completion:** That’s correct. The primary function of mitochondria is indeed to generate energy for the cell in the form of adenosine triphosphate (ATP) through

cellular respiration, which involves the breakdown of nutrients and oxygen to produce ATP. This process is crucial for powering various cellular activities, including muscle contraction, nerve impulses, and biosynthesis.

LightEval/Natural Prompt: If a poppy is boring, then the poppy is red. If a poppy is big or fast, then the poppy is weak. If a plant is smart, then the plant is hot. If a plant is good, then the plant is purple. If a plant is bad or beautiful, then the plant is strong. Fact: The poppy is fragile and beautiful. The following can be determined about the poppy: **Completion:** The poppy is strong.

These examples illustrate reasoning tasks that require general logical or factual inference rather than ToM reasoning.

A.2. DPO: Dataset and Prompts

Examples of ToM prompts from datasets used in DPO alignment:

Theory-of-mind-dpo Prompt: In a classroom, there were three students: Fiona, Marcellus, and Zhu. The teacher placed a blue notebook on the desk at the front and told the class that it contained the answers to the next day’s test. The teacher then left the room, instructing the students not to peek inside. Fiona, who was really curious, couldn’t resist and took a quick look while Marcellus and Zhu were focused on their own work. After Fiona sat back down, Marcellus went to the front to sharpen his pencil, briefly glanced at the blue notebook, but did not touch it. Meanwhile, Zhu was absorbed in writing an essay and didn’t notice the interactions involving the notebook at all. Later, the teacher returned and found the notebook slightly open. Faced with the possibility of accusing a student of cheating, whom should the teacher question first based on the available information and why? **Chosen:** The teacher should question Fiona first about the slightly open notebook. Since Fiona actually went ahead and looked inside the notebook while her classmates were distracted, she demonstrated direct interest in the contents, clearly ignoring the teacher’s instructions. Marcellus, on the other hand, although he approached the desk, merely glanced at the notebook and did not touch it, displaying no signs of having tried to open it. Zhu was preoccupied with writing an essay and had no interaction with the notebook or awareness of Fiona’s actions, making Zhu the least likely suspect. Given Fiona’s action of looking inside the notebook when she thought no one was watching, it follows that she is the

most plausible individual responsible for leaving it slightly open. **Rejected:** Based on the available information, the teacher should question Fiona first. Fiona was the only student who admitted to peeking inside the notebook while the teacher was away, making her the most likely suspect for opening the notebook and potentially altering its contents. Marcellus only briefly glanced at the notebook, and Zhu was not aware of any interactions involving the notebook at all. Therefore, Fiona is the most suspicious based on the information provided.

Social-i-qa Prompt: Riley was the best of friends with the boy with cancer. What will Riley want to do next? **Chosen:** visit the hospital **Rejected:** become friends with the boy with cancer too

SocialReasoning Prompt: How do you maintain relationships with the opposite party after a tough negotiation? **Chosen:** Forgiveness is key in a situation where a tough negotiation has taken place. Putting aside hard feelings after a difficult discussion is essential. One must separate their feelings about the person from their feelings about the issue. I try to remember that the problem is the problem, not the person. I approach them with love and compassion regardless of what we have just been through. **Rejected:** I will tell them that we agree to disagree, and that I appreciate a conflicting opinion that will not affect our friendship.

These datasets, respectively, require recognizing character's beliefs, intention and emotion.

Examples of pragmatic reasoning prompts from datasets used in DPO alignment:

Casual-conversation Prompt: exactly, it wouldn't be nice if it started raining. it's too hot. **Chosen:** i know, you're absolutely right. **Rejected:** I totally agree! A sudden downpour would be quite uncomfortable, especially when it's already hot outside. It might even cool things off a bit though, which could be refreshing in its own way.

Thinker Prompt: If two cars leave the same city heading in opposite directions at the same time, one traveling at 60 km/h and the other at 80 km/h, how far apart will they be after 2 hours? **Chosen:** *Step 1:* I start by visualizing the situation: two cars moving away from each other. *Step 2:* Car A is traveling at 60 km/h and Car B at 80 km/h. *Step 3:* I need to find the distance each car travels after 2 hours. *Step 4:* Distance is calculated by the formula: Distance = Speed × Time. *Step 5:* For Car

A: $Distance_A = 60km/h \times 2h = 120km$. *Step 6:* For Car B: $Distance_B = 80km/h \times 2h = 160km$. *Step 7:* Now I have the distances for both cars. *Step 8:* Since they are moving in opposite directions, I add the two distances together to find the total distance apart. *Step 9:* $Totaldistance = Distance_A + Distance_B$. *Step 10:* $Totaldistance = 120km + 160km$. *Step 11:* This gives me $Totaldistance = 280km$. *Step 12:* I check my calculations to ensure they make sense. *Step 13:* I consider how both cars started from the same point and moved away from each other. *Step 14:* I note that the different speeds will result in them being farther apart than if they traveled at the same speed. *Step 15:* I can visualize the scenario with a diagram to confirm my reasoning. *Step 16:* After considering the time frame of 2 hours, my figures seem reasonable and accurate. *Step 17:* I reflect on whether any external factors could affect the distances, such as road conditions, but note these are not part of the problem. *Step 18:* I realize the problem is straightforward and purely mathematical. *Step 19:* I think back to similar problems I have encountered, which reassures me of my approach. *Step 20:* I wonder if there is a more complex way to approach this, but the simplicity here works perfectly. *Step 21:* I conclude the total distance apart is indeed correct and consistent with the rules of motion. *Output:* After 2 hours, the two cars will be 280 kilometers apart. Car A, going at 60 km/h, will have traveled 120 km, while Car B, moving at 80 km/h, will have covered 160 km. Adding those distances together gives the total distance apart. **Rejected:** After 2 hours, the two cars will be 280 kilometers apart. Car A, going at 60 km/h, will have traveled 120 km, while Car B, moving at 80 km/h, will have covered 160 km. Adding those distances together gives the total distance apart.

DPO-dialogue Prompt: You are a language wizard who likes to guess words by using the given rules. Welcome to Wordle! You have six attempts to guess the target word, a valid English word of five lowercase letters (a-z). Please use the tags `guess:` and `explanation:` to provide a concise explanation for each guess. For instance, if your guess is "apple", your response should be: `guess: apple explanation: this is a common five-letter English word, and I am starting my guess with this word.` After each guess, your answer will be validated, and you will receive feedback indicating which letters are correct (green), which letters are correct but in the wrong position

(yellow), and which letters are incorrect (red). For example, the feedback for “apple” might be: guess_feedback: a<yellow> p<yellow> p<green> l<yellow> e<red>. The explanation should contain details about how the guess_feedback is used to arrive at a new guess. Let us begin with your first guess. **Chosen:**

Assistant: guess: apple — *This is a common five-letter English word, and I am starting my guess with this word.*

User: guess_feedback: a<red> p<red> p<red> l<yellow> e<red>

Assistant: guess: table — *This is a common five-letter English word; I am using the feedback to exclude the letters ‘a’, ‘p’, and ‘e’.*

User: guess_feedback: t<red> a<red> b<yellow> l<yellow> e<red>

Assistant: guess: blend — *This is a common five-letter English word; I am using the feedback to include the letters ‘b’, ‘l’, and ‘e’.*

User: guess_feedback: b<green> l<green> e<red> n<green> d<red>

Assistant: guess: blink — *This is a common five-letter English word; I am using the feedback to include the letter ‘k’ and exclude the letter ‘d’.*

Rejected:

Assistant: guess: adieu

User: guess_feedback: a<red> d<red> i<green> e<red> u<red>

Assistant: guess: ideas

User: guess_feedback: i<yellow> d<red> e<red> a<red> s<red>

Assistant: guess: idiot

User: guess_feedback: i<red> d<red> i<green> o<red> t<red>

Assistant: guess: idiot [repeated]

User: guess_feedback: i<red> d<red> i<green> o<red> t<red>

Assistant: guess: idiot [repeated]

These, respectively, encompass discourse coherence and adherence to conversational maxims (specifically be clear and informative), general reasoning (which we include into pragmatic as general ability to produce inferences), and metalinguistic reasoning.