The Pragmatic Mind of Machines: Tracing the Emergence of Pragmatic Competence in Large Language Models

Anonymous ACL submission

Abstract

Current large language models (LLMs) have demonstrated emerging capabilities in social intelligence tasks, including implicature resolution (Sravanthi et al., 2024) and theory-of-mind reasoning (Shapira et al., 2024), both of which require substantial pragmatic understanding. However, how LLMs acquire this competence throughout the training process remains poorly understood. In this work, we introduce ALTPRAG, a dataset grounded in the pragmatic concept of alternatives, designed to evaluate whether LLMs at different training stages can accurately infer nuanced speaker intentions. Each instance pairs two contextually appropriate but pragmatically distinct continuations, enabling fine-grained assessment of both pragmatic interpretation and contrastive reasoning. We systematically evaluate 22 LLMs across key training stages: pre-training, supervised fine-tuning (SFT), and preference optimization, to examine the development of pragmatic competence. Our results show that even base models exhibit notable sensitivity to pragmatic cues, which improves consistently with increases in model and data scale. Additionally, SFT and RLHF contribute further gains, particularly in cognitive-pragmatic reasoning. These findings highlight pragmatic competence as an emergent and compositional property of LLM training and offer new insights for aligning models with human communicative norms.

1 Introduction

Human communication typically extends beyond the literal interpretation of utterances. Pragmatics, the branch of linguistics concerned with how context shapes meaning, is central to natural language understanding. It encompasses a range of phenomena such as implicature (Sadock, 1978), presupposition (Karttunen, 1974), and indirect speech acts (Searle, 1975). With the advent of LLMs, a growing body of research has begun to explore whether these models exhibit sensitivity to pragmatic cues.

Figure 1: Illustration of pragmatic alternatives. Two contrasting replies to the same question convey different pragmatic forces, one direct and explanatory, the other playful and implicitly affirmative. We prompt LLMs to infer the pragmatic intention behind each continuation and identify the contextual conditions under which one would be preferred over the other.

Recent studies have investigated LLMs' abilities to infer speaker intentions (Hu et al., 2023; Ruis et al., 2023; Sravanthi et al., 2024), perform theory-of-mind reasoning (Kosinski, 2024; Chen et al., 2024; Shapira et al., 2024), and even pass Turing tests in controlled settings (Jones and Bergen, 2025). These developments suggest that LLMs may exhibit emergent pragmatic behaviors, prompting deeper inquiry into their underlying capabilities.

However, it remains an open question at which stage of training LLMs acquire pragmatic understanding. Ruis et al. (2023) conducted an empirical study showing that only example-level instruction-tuned (IT) models will significantly outperform

random baselines on pragmatic tasks. Nonetheless, their evaluation faces two key limitations. First, their analysis is based on a binary classification task (George and Mamidi, 2020), in which models respond "yes" or "no" to specific utterances - an approach that may oversimplify the complexity of pragmatic reasoning and overlook important subtleties. Second, the category of example-level IT models they examine primarily includes proprietary models such as GPT-3.5 and GPT-4, for which the specific training procedures are not publicly known. In particular, it is unclear when techniques like reinforcement learning from human feedback (RLHF) are applied. As a result, it is difficult to draw reliable conclusions about how pragmatic competence correlates with specific phases of training.

In this paper, we introduce ALTPRAG, a human-in-the-loop annotated dataset grounded in the notion of alternatives in pragmatics. As illustrated in Figure 1, each dialogue instance is paired with two contextually relevant but pragmatically distinct continuations, highlighting contrasts in speaker intention. Using this dataset we test models' capacities to explain the intentions underlying these alternatives to probe the pragmatic capabilities of LLMs at different training stages, specifically, after pre-training, SFT, and preference optimization. To evaluate model performance, we adopt an LLM-as-a-judge framework, comparing model-generated interpretations of pragmatic intent with human-verified references. Our results and contributions can be summarized as follows:

- We present the first systematic analysis of how pragmatic competence evolves across different training stages of LLMs, using a free-form evaluation framework to capture nuanced linguistic judgments.
- In contrast to findings by Ruis et al. (2023), our results indicate that base LLMs already exhibit a degree of rational pragmatic competence, which scales with both model size and the volume of training data.
- We further show that both SFT and RLHF can enhance pragmatic understanding, particularly in capturing nuanced cognitive-pragmatic competence.

2 Related Work

Pragmatics in LLMs. The extent to which large language models (LLMs) understand and process pragmatic phenomena has been the focus of increasing scholarly attention. Hu et al. (2023) evaluated a range of LLMs on their ability to interpret pragmatic constructs such as deception, indirect speech, and irony, finding that the largest models approached human-level accuracy and exhibited similar error patterns. Building on this line of inquiry, Sravanthi et al. (2024) introduced a comprehensive benchmark designed to capture more nuanced aspects of pragmatic reasoning beyond conventional multiple-choice formats. Extending this evaluation paradigm further, Wu et al. (2024) proposed free-form pragmatic tasks and demonstrated that preference optimization can enhance pragmatic performance, suggesting it may serve as a "free lunch" for pragmatic competence. Other studies have investigated LLMs as pragmatic speakers using reference games (Shaikh et al., 2023; Jian and Siddharth, 2024), while Cong (2024) focused specifically on manner implicature to probe the fine-grained nature of pragmatic understanding in these models.

Training Phases of LLMs. The typical pipeline for developing deployment-ready LLMs involves several sequential training phases. First, models are pre-trained on large-scale text corpora to acquire general-purpose language representations. This is followed by instruction tuning, where models are trained on curated input-output pairs to better follow human instructions (Mishra et al., 2022; Longpre et al., 2023). We adopt the term "SFT" throughout this paper to align with current usage and emphasize its role as the first stage of alignment after pretraining. The final stage typically involves *preference optimization*, commonly implemented via Proximal Policy Optimization (Schulman et al., 2017) to align LLMs with human values. A recent and widely adopted PPO alternative: **Direct Preference Optimization (DPO)** simplifies PPO by avoiding reward modeling and policy optimization, instead directly optimizing model outputs to align with pairwise human preferences. Many open-source checkpoints (e.g., OLMo-2 (OLMo et al., 2025)) are released in DPO variants, which we adopt in our experimental comparisons.

A number of studies have investigated how these training stages affect downstream model behavior. For instance, Song et al. (2025) found that capabilities emerge at different rates during instruction tuning. Kirk et al. (2024) conducted a systematic analysis of SFT and RLHF, reporting that RLHF improves out-of-distribution generalization

but also reduces output diversity. Building on this line of work, we investigate these training phases in greater depth, with a particular focus on how each stage contributes to the emergence of pragmatic competence.

3 ALTPRAG

In both producing and interpreting utterances, speakers routinely consider alternative expressions, which are linguistic forms that could have been used but were not. These alternatives often play a crucial role in shaping the grammaticality and contextual appropriateness of an utterance (Degen, 2013). For instance, as illustrated in Figure 1, both candidate responses plausibly continue the question "Are you going to the gym?" However, they convey distinct pragmatic stances: the first reply, "Yes, I've been slacking off lately and need to catch up," provides a candid, self-reflective explanation, while the second, "Do I ever do anything else on a Friday?", takes a more playful and rhetorically indirect approach. Although semantically aligned in affirming the same activity, these alternatives differ in tone, social positioning, and expressive style.

In this work, we leverage such pragmatic contrasts to construct a dataset called ALTPRAG designed to probe LLMs' sensitivity to speaker intent, politeness, and implicature. Specifically, we use GPT4o (OpenAI et al., 2024) to generate a reference set of conversations with alternatives accompanied by human-verified explanations of pragmatic intent (Figure 2A).

3.1 First-round Data Generation

In the initial round of data generation, we build on the scenario-based dataset introduced by Hu et al. (2023) and the pragmatic benchmark proposed by Sravanthi et al. (2024). For each data point, we extract the scenario description as contextual background and treat the target sentence as the root of a dialogue. Using this setup, we prompt GPT-4o (OpenAI et al., 2024) to generate two contextually coherent but pragmatically distinct alternative continuations. The model is additionally instructed to provide natural language explanations detailing the pragmatic functions conveyed by each alternative, as well as why a speaker might prefer one over the other in a given context. This method allows us to elicit fine-grained pragmatic contrasts grounded in realistic and context-sensitive language use. Details on the prompt template and data postprocessing procedure are provided in Appendix A. In total, the first round of data generation yields 1298 datapoints.

3.2 Human-in-the-loop Refinement

Each datapoint was labeled as a *pass* only if it met the evaluation criteria described below and was independently approved by all three annotators. Otherwise, it was marked as a *fail*. Three authors with undergraduate training in pragmatics independently annotated each datapoint using the following criteria:

(1) Both continuations must be coherent and contextually appropriate responses to the initial utterance.

(2) Each natural language explanation must accurately capture the pragmatic function of its corresponding continuation and reflect nuanced speaker preferences.

Out of the initial 1,298 raw examples generated by the model, 650 passed this filtering stage.

To augment the dataset for evaluation purposes, we apply a symmetric transformation: for each validated datapoint, we generate a mirrored version by swapping the order of the two responses and their corresponding explanations. This enables us to probe model judgments about each sentence independently. After augmentation, our final dataset contains 1,300 examples. A representative example datapoint is shown in Table 1.

4 Experimental Setup

In evaluations, we provide models with the conversations from ALTPRAG and prompt models to generate analogous explanations of pragmatic intent to our gold references, using these to evaluate models' pragmatic reasoning via ten-point scoring and pairwise comparison metrics (Figure 2B).

4.1 Evaluated LLM Variants

To investigate how pragmatic competence develops across training stages, we evaluate a diverse set of open-source LLMs, covering different parameter scales and fine-tuning strategies:

- *OLMo-2 Series* (OLMo et al., 2025): We evaluate OLMo-2 models at 7B, 13B, and 32B parameter scales, each available in Base, SFT, and DPO variants. These models are trained

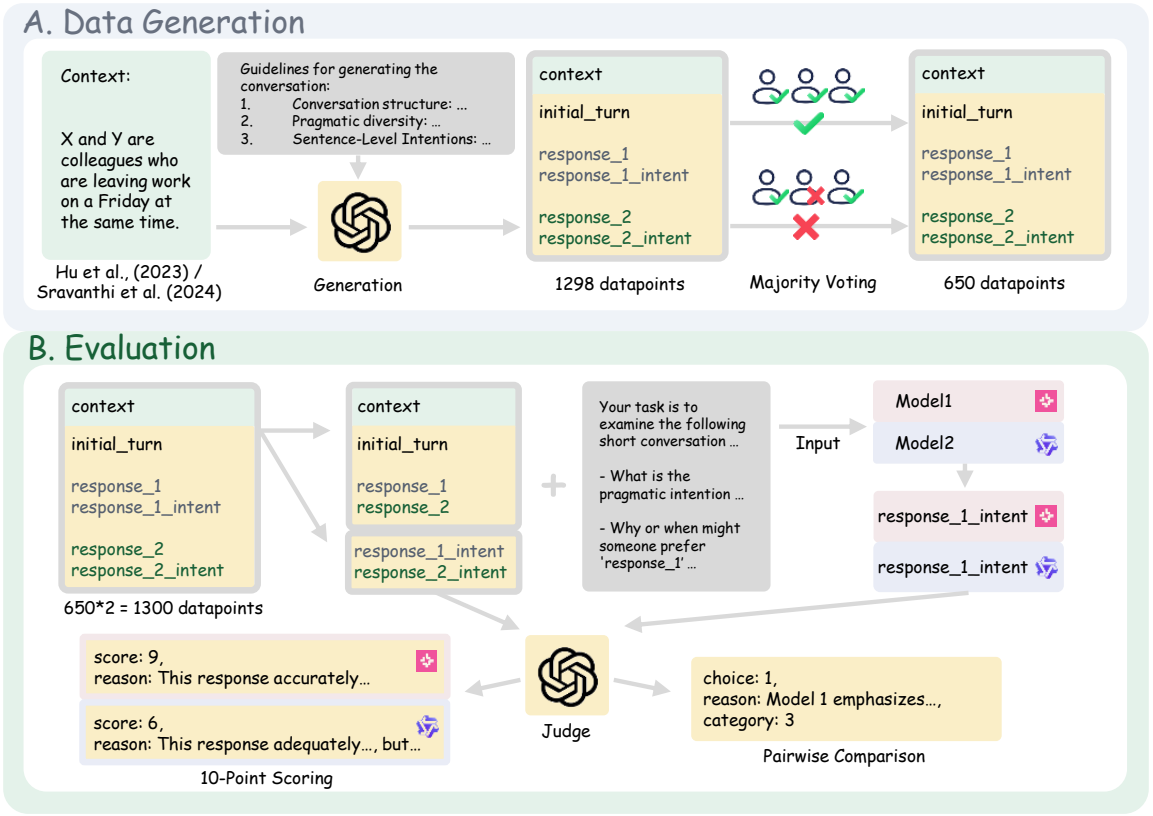


Figure 2: An illustration of the data generation process and evaluation workflow. After the majority voting phase, we construct a mirrored version by swapping the order of the two responses and their associated reference labels, resulting in a total of 1,300 data points.

on up to 6 trillion tokens and further refined using the Tülu 3 instruction-following and preference datasets.

- *OLMoE-1B-7B* (Muennighoff et al., 2025): This Mixture-of-Experts (MoE) model consists of 7 billion total parameters, with 1 billion active during inference. We evaluate the Base, SFT, and DPO checkpoints to assess their pragmatic behaviors.

- *LLaMA-3.1-Tülu-3 Series* (Grattafiori et al., 2024; Lambert et al., 2025): Based on Meta's LLaMA-3.1 foundation models, we evaluate 8B and 70B parameter variants, each trained with the Tülu 3 post-training pipeline. As with other families, we include the Base, SFT, and DPO versions to observe training-stage-specific effects.

To further probe the emergence of pragmatic competence in base models, we additionally evaluate *Qwen-3* base models at 0.6B, 1.7B, 4B, and 8B parameter sizes (Yang et al., 2025), as this project does not release instruction-tuned or preference-optimized checkpoints. This setup includes a broader comparison of baseline pragmatic abilities across model families and enables a fine-grained analysis of how pragmatic understanding emerges and evolves in LLMs, as well as the role of instruction tuning and preference optimization in shaping communicative competence.

4.2 Prompting Strategy and Setup

Evaluated models are prompted to generate an explanation of the pragmatic intention underlying each alternative in a conversational datapoint, as well as the pragmatic reasons why a speaker might prefer a given alternative to the other.

To mitigate the instability and underperformance commonly observed in interactions with base models, we adopt the URIAL prompt template introduced by Lin et al. (2023). This template is specifically designed to elicit more helpful and coherent outputs from base-stage LLMs without additional instruction tuning. For consistency and fairness across model stages, we apply the same template when evaluating SFT and DPO variants. The complete prompt template can be found in Appendix B.

To prevent evaluation inflation via format imita-

Field	Content
`context`	X and Y are colleagues who are leaving work on a Friday at the same time.
`initial_turn`	*Are you going to the gym?*
`response_1`	*Yes, I've been slacking off lately and need to catch up.*
`response_1_intent`	This response does not flout any maxim. It aims to directly acknowledge the intention to go to the gym while adding a self-reflective comment about recent habits. A speaker might prefer this reply over `response_2` because it is straightforward, honest, and avoids sarcasm, making the interaction clearer and more sincere. This approach fits well if the speaker wishes to be open and relatable.
`response_2`	*Do I ever do anything else on a Friday?*
`response_2_intent`	This response flouts the maxim of Quality. It aims to use sarcasm to humorously imply that going to the gym on Fridays is routine for the speaker. A speaker might prefer this reply over `response_1` because it injects playfulness and familiarity into the conversation, which can help maintain a lighthearted tone among colleagues. This style can strengthen rapport if both parties appreciate joking banter.

Table 1: An example datapoint showing a complete conversation with two pragmatically distinct continuations and annotated intentions.

tion, we adopt **zero-shot prompting** throughout, avoiding any in-prompt examples or structural cues. This ensures that models rely solely on their internal representations of pragmatic intent.

To control for variability, we fix decoding parameters across all runs: `max_new_tokens` = 256, `top_k` = 50, `top_p` = 1.0, and `temperature` = 0.5. Full configuration details appear in Appendix C.

4.3 Evaluation Metrics

We adopt two complementary LLM-as-a-Judge evaluation protocols (Lin and Chen, 2023; Fu et al., 2023), both employing GPT-4.1 (OpenAI et al., 2024) as the evaluator to assess the quality of model-generated explanations of pragmatic intent.

10-Point Scoring. In this setting, the evaluator is provided with the conversation, reference intent explanation, and model-generated hypothesis intent explanation, and asked to assign each explanation a score on a 10-point scale, accompanied by a brief justification. This method allows for direct, fine-grained comparison of explanation quality across different model variants. The full prompt template is provided in Appendix D.

Pairwise Win Rate. To mitigate potential scoring biases and highlight relative differences across training stages, we also conduct pairwise comparisons between model variants (e.g., Base vs. SFT, SFT vs. DPO). For each pair, the evaluator is asked to determine which explanation better captures the speaker's pragmatic intent. Drawing on the framework of pragmatic competence from Mao and He (2021), we further instruct the evaluator to categorize each winning explanation into one of three dimensions:

1. *Cognitive-pragmatic competence*: The explanation goes beyond literal meaning and identifies the speaker's underlying communicative goal or intention.

2. *Pragmalinguistic competence*: The explanation highlights rhetorical strategies such as humor, irony, or self-deprecation and explains how these are used to manage interpersonal meaning.

3. *Sociopragmatic competence*: The explanation demonstrates awareness of social norms, roles, relationships, or context-sensitive appropriateness in the speaker's choice.

Together, these two evaluation protocols enable both absolute assessment of explanation quality and nuanced, comparative analysis of pragmatic competence across training stages.

5 Results

5.1 General Results

We present our overall findings from both the 10-point scoring and pairwise win rate comparisons, focusing on how pragmatic competence develops across model training stages.

10-Point Scoring. As shown in Figure 3, models generally achieve higher scores as they progress from base to SFT to DPO stages. We conduct a Wilcoxon test (Wilcoxon, 1992) between the model and its immediately following training stage, and the results suggest that all score rises are statistically significant. First, we found that base models

Model	Stages	Win Rate
OLMo-2-7B	Base vs SFT	26.9% vs 73.1%
OLMo-2-7B	SFT vs DPO	19.8% vs 80.2%
OLMo-2-7B	Base vs DPO	10.8% vs 89.2%
OLMo-2-13B	Base vs SFT	27.6% vs 72.4%
OLMo-2-13B	SFT vs DPO	23.1% vs 76.9%
OLMo-2-13B	Base vs DPO	10.9% vs 90.1%
OLMo-2-32B	Base vs SFT	56.7% vs 43.2%
OLMo-2-32B	SFT vs DPO	7.7% vs 92.3%
OLMo-2-32B	Base vs DPO	14.2% vs 85.8%
OLMoE-1B-7B	Base vs SFT	32.4% vs 67.6%
OLMoE-1B-7B	SFT vs DPO	44.7% vs 55.3%
OLMoE-1B-7B	Base vs DPO	19.7% vs 80.3%
Llama3-8B	Base vs SFT	26.3% vs 73.7%
Llama3-8B	SFT vs DPO	32.0% vs 68.0%
Llama3-8B	Base vs DPO	16.4% vs 83.6%
Llama3-70B	Base vs SFT	38.7% vs 61.3%
Llama3-70B	SFT vs DPO	9.6% vs 90.4%
Llama3-70B	Base vs DPO	11.0% vs 89.0%

Table 2: Pairwise win rate comparisons across model stages. Win rates are reported as *A% vs B%*, where A and B correspond to the stages listed.

already demonstrate surprising competence, with average scores around 6 out of 10 for models with 7-8B parameters, indicating that early-stage models are already capable of non-trivial pragmatic inference without instruction tuning or preference optimization, likely benefiting from implicit exposure to pragmatic phenomena during large-scale pretraining. At the DPO stage, responses generally receive scores of 8 or higher, reflecting a marked alignment between model output and the intent conveyed in reference annotations. Example responses can be found in Appendix F.

Pairwise Win Rate. Consistent with the scoring results (Table 2), DPO models achieve the highest win rates in all head-to-head comparisons, followed by SFT and then base models. This pattern holds across different model families and parameter scales, reinforcing the observation that both SFT and DPO contribute to improved pragmatic sensitivity. These results provide empirical support for the view that pragmatic competence emerges gradually, with measurable gains at each fine-tuning stage.

5.2 Does Pragmatic Competence Scale?

We further analyze how pragmatic competence scales with two key factors: model size and pretraining data volume.

Scaling with Model Size. We observe that larger models tend to achieve better pragmatic competence across families. This trend holds across evaluated model families, including OLMo-2 (OLMo et al., 2025) and LLaMA-3.1-Tülu-3 (Grattafiori et al., 2024; Lambert et al., 2025). However, since OLMo-2 models are trained on different amounts of pretraining data at each scale, we cannot strictly attribute these gains to model size alone.

To more precisely isolate the effect of parameter scaling, we conduct a controlled comparison between LLaMA-3.1 7B and 70B models, which share the same pretraining corpus. In our head-to-head comparison, the 70B model achieves a substantially higher win rate (66%) than the 7B model (34%), providing evidence that increased model capacity also plays an important role in improving pragmatic competence. Similarly, our evaluation on Qwen-3 models (Yang et al., 2025), which vary in parameter size but share the same large-scale pretraining data, shows clear scaling trends: larger Qwen-3 models generally achieve higher scores, reinforcing the positive effect of model size.

Scaling with Pretraining Data. We also find that pretraining data volume contributes significantly to base model performance. The Qwen-3 series, trained on 36T tokens (Yang et al., 2025), shows relatively strong pragmatic competence across parameter scales, performing better than other models of similar size trained on smaller corpora—such as the OLMo-2 series, which is trained on 4T/5T/6T tokens for the 7B/13B/32B models, respectively (OLMo et al., 2025). Notably, the Qwen-3 1.7B model achieves a higher average score (6.48) than the OLMo-2 7B model (6.13), illustrating how pretraining scale alone can improve models' ability to infer pragmatic intent, suggesting that larger pretraining corpora can also contribute to enhancing pragmatic abilities.

Taken together, our results show that both parameter scaling and pretraining corpus size shape pragmatic ability. While larger models tend to perform better, our findings highlight the often underappreciated role of pretraining data quality and scale—particularly in the emergence of early-stage pragmatic competence. These results underscore the need to consider pretraining data as an important factor in shaping pragmatic abilities.

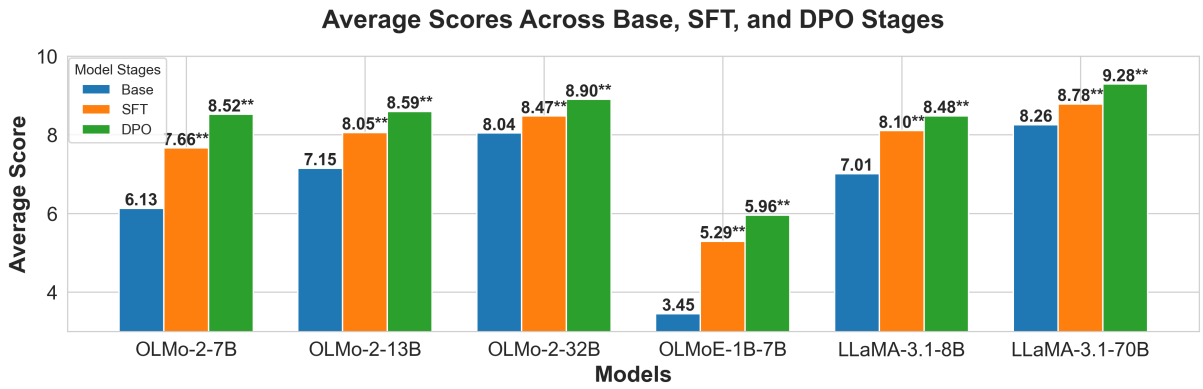


Figure 3: Average ten-point quality scores across Base, SFT, and DPO stages for different model families. Significance codes are based on Wilcoxon signed-rank tests comparing each stage with the previous one (e.g., SFT vs. Base, DPO vs. SFT). Asterisks denote statistical significance: * $p < 0.05$, ** $p < 0.01$. Base-stage results are not assigned significance codes as they are used as reference baselines.

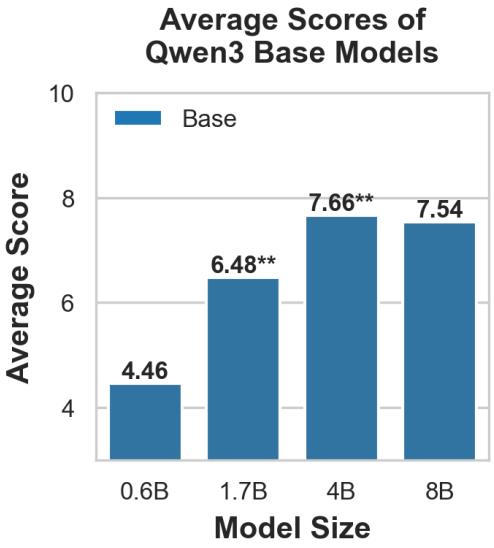


Figure 4: The Qwen-3 series achieves comparatively higher scores with fewer parameters, illustrating that scaling pretraining data size can enhance a model's capacity for pragmatic reasoning.

5.3 Where Do SFT and DPO Help?

To better understand the role of fine-tuning and preference optimization in shaping pragmatic competence, in Figure 5 we visualize the distribution of winning explanations across the three aforementioned categories of pragmatic competence from Mao and He (2021): cognitive-pragmatic, pragmalinguistic, and sociopragmatic.

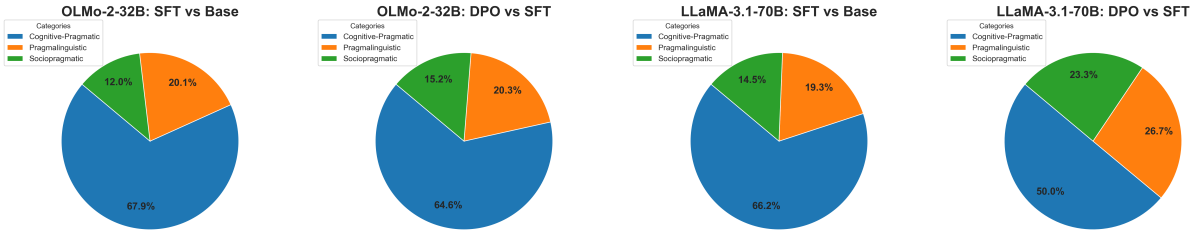
We find that *cognitive-pragmatic competence*, the ability to go beyond literal meaning and infer the speaker's communicative goal, is the primary justification for wins across all model stages. This trend is especially pronounced in SFT stage, where cognitive-pragmatic explanations account for the majority of wins over base models. In the OLMo-2-32B SFT variant, 66.7% of winning explanations fall into this category, suggesting that supervised fine-tuning primarily strengthens the model's ability to capture intended meaning.

While cognitive-pragmatic competence remains the dominant strategy in DPO, we observe a continued strengthening of this ability compared to SFT stage, indicating that preference optimization further refines models' understanding of speaker intent. In parallel, we also observe a shift toward more *sociopragmatic competence*—the ability to recognize social roles, politeness strategies, and contextual appropriateness—suggesting that the DPO stage broadens the scope of pragmatic strategies beyond purely cognitive interpretations. The full comparison results can be found in Appendix H.

6 Discussion

In this paper, we revisit the findings presented in Ruis et al. (2023) and utilize the concept of 'alternative' to construct a dataset for evaluating LLMs at various training stages. Our results provide a complement to the previous findings: although instruction tuning and DPO surely help, base models already show non-trivial pragmatic competence.

The Role of Pretraining. Our findings underscore the foundational role of pretraining in shaping LLMs' pragmatic competence. Even base models—those without instruction tuning or preference optimization—exhibit non-trivial sensitivity to speaker intent and context. This effect is



(a) OLMo-2-32B SFT vs Base

(b) OLMo-2-32B DPO vs SFT

(c) LLaMA-3.1-70B SFT vs Base

(d) LLaMA-3.1-70B DPO vs SFT

Figure 5: Distribution of winning explanation categories across selected model comparisons. While both SFT and DPO stages are dominated by cognitive-pragmatic explanations, the DPO stage shows a notable increase in sociopragmatic responses, indicating enhanced sensitivity to social context and appropriateness.

particularly evident in the Qwen-3 series, which, despite having fewer parameters than many of its counterparts, performs competitively across pragmatic tasks. Notably, Qwen-3 models are trained on 36 trillion tokens, one of the largest publicly reported pretraining corpora, suggesting that the scale and quality of pretraining data can significantly enhance pragmatic reasoning, independent of model size.

These observations align with findings from Yue et al. (2025) and Essential AI et al. (2025), who report that base models already demonstrate strong reasoning capabilities and comparable pass@K performance to post-trained reasoning models. Our results extend this evidence from the perspective of pragmatics, indicating that much of an LLM's ability to handle nuanced, context-sensitive communication is already rooted in the pretraining phase. This highlights the importance of pretraining not only for general language modeling and reasoning but also for the development of socially and pragmatically competent behavior.

Revisiting Goldilocks: Improved Base Model Performance. To further contextualize our findings and make a fair comparison, we replicate the experiments proposed in Ruis et al. (2023) using the zero-shot setting. When applied to modern base models such as the OLMo-2 and Qwen-3 series, we observe substantially improved performance compared to the earlier-generation models used in the original study (Table 3). Through well-conducted pretraining, current LLMs even surpass the original 175B GPT-3 model. In particular, OLMo-2 base models with larger parameter counts reach accuracy levels above 70%, highlighting the increased pragmatic competence brought by better quality of pretraining without instruction tuning. These results also exhibit clear scaling patterns, with accuracy improving as model size increases. These results further underscore the importance of pretraining as a mechanism for establishing strong foundations for pragmatic competence.

Model	Accuracy (%)
GPT-2-xl (Goldilocks)	51.3
OPT-13B (Goldilocks)	61.0
GPT-3-175B (Goldilocks)	57.7
OLMo-2-7B	71.7
OLMo-2-13B	70.5
OLMo-2-32B	75.5
Qwen3-0.6B-Base	63.8
Qwen3-1.7B-Base	67.3
Qwen3-4B-Base	69.5

Table 3: Accuracy on the Goldilocks implicature reasoning task. Top section shows original results reported by Ruis et al. (2023); bottom section reports our own evaluation of modern base models using the same experimental setup.

Beyond Pretraining: Refining Pragmatic Competence through SFT and DPO. While pretraining establishes a baseline for pragmatic reasoning, our results show that SFT significantly enhances cognitive-pragmatic competence, enabling models to more effectively infer speaker intent beyond literal content. Preference optimization through DPO provides additional gains, particularly in sociopragmatic competence, by improving the model's sensitivity to social context, roles, and politeness norms. Wu et al. (2024) argue that preference optimization may offer a "near-free lunch," improving pragmatic ability without degrading general performance. Our findings reinforce this view, highlighting the critical role of preference optimization in advancing pragmatic competence, especially in socially grounded interpretations.

Limitations

This work faces two primary limitations. First, the current dataset does not explicitly distinguish among different types of pragmatic phenomena, such as humor, indirect speech, or irony, which limits our ability to analyze how models at various training stages handle specific subcategories of pragmatics. Second, while we include multiple model families and architectures, all evaluated models across training stages are developed by the same organization (AI2). This shared provenance may introduce systematic biases, potentially limiting the generalizability of our findings.

References

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. 2024. ToMBench: Benchmarking theory of mind in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15959–15983, Bangkok, Thailand. Association for Computational Linguistics.

Yan Cong. 2024. Manner implicatures in large language models. *Scientific Reports*, 14(1):29113.

Judith Degen. 2013. *Alternatives in pragmatic reasoning*. University of Rochester.

Essential AI, :, Darsh J Shah, Peter Rushton, Somanshu Singla, Mohit Parmar, Kurt Smith, Yash Vanjani, Ashish Vaswani, Adarsh Chaluvaraju, Andrew Hojel, Andrew Ma, Anil Thomas, Anthony Polloreno, Ashish Tanwer, Burhan Drak Sibai, Divya S Mansingka, Divya Shivaprasad, Ishaan Shah, and 10 others. 2025. Rethinking reflection in pre-training. *Preprint*, arXiv:2504.04022.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *Preprint*, arXiv:2302.04166.

Elizabeth Jasmi George and Radhika Mamidi. 2020. Conversational implicatures in english dialogue: Annotated dataset. *Procedia Computer Science*, 171:2316–2323. Third International Conference on Computing and Network Communications (CoCoNet'19).

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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*.

Mingyue Jian and N. Siddharth. 2024. Are llms good pragmatic speakers? *Preprint*, arXiv:2411.01562.

Cameron R. Jones and Benjamin K. Bergen. 2025. Large language models pass the turing test. *Preprint*, arXiv:2503.23674.

Lauri Karttunen. 1974. Presupposition and linguistic context.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2024. Understanding the effects of rlhf on llm generalisation and diversity. *Preprint*, arXiv:2310.06452.

Michal Kosinski. 2024. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, and 4 others. 2025. Tulu 3: Pushing frontiers in open language model post-training. *Preprint*, arXiv:2411.15124.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. 2023. The unlocking spell on base llms: Rethinking alignment via in-context learning. *Preprint*, arXiv:2312.01552.

Yen-Ting Lin and Yun-Nung Chen. 2023. Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. *Preprint*, arXiv:2305.13711.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 22631–22648. PMLR.

Tiaoyuan Mao and Shanhua He. 2021. An integrated approach to pragmatic competence: Its framework and properties. *SAGE Open*, 11(2):21582440211011472.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume*

1: Long Papers), pages 3470–3487, Dublin, Ireland. Association for Computational Linguistics.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Pete Walsh, Oyvind Tafjord, Nathan Lambert, Yuling Gu, Shane Arora, Akshita Bhagia, Dustin Schwenk, David Wadden, Alexander Wettig, Binyuan Hui, Tim Dettmers, Douwe Kiela, and 5 others. 2025. Olmoe: Open mixture-of-experts language models. *Preprint*, arXiv:2409.02060.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, and 21 others. 2025. 2 olmo 2 furious. *Preprint*, arXiv:2501.00656.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

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.

Jerrold M Sadock. 1978. On testing for conversational implicature. In *Pragmatics*, pages 281–297. Brill.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *Preprint*, arXiv:1707.06347.

John R Searle. 1975. Indirect speech acts. In *Speech acts*, pages 59–82. Brill.

Omar Shaikh, Caleb Ziems, William Held, Aryan Pariani, Fred Morstatter, and Diyi Yang. 2023. Modeling cross-cultural pragmatic inference with codenames duet. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6550–6569, Toronto, Canada. Association for Computational Linguistics.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2024. Clever hans or neural theory of mind? stress testing social reasoning in large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2257–2273, St. Julian's, Malta. Association for Computational Linguistics.

Chiyu Song, Zhanchao Zhou, Jianhao Yan, Yuejiao Fei, Zhenzhong Lan, and Yue Zhang. 2025. Dynamics of instruction fine-tuning for Chinese large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10345–10366, Abu Dhabi, UAE. Association for Computational Linguistics.

Settaluri Sravanthi, Meet Doshi, Pavan Tankala, Rudra Murthy, Raj Dabre, and Pushpak Bhattacharyya. 2024. PUB: A pragmatics understanding benchmark for assessing LLMs' pragmatics capabilities. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12075–12097, Bangkok, Thailand. Association for Computational Linguistics.

Frank Wilcoxon. 1992. Individual comparisons by ranking methods. In *Breakthroughs in statistics: Methodology and distribution*, pages 196–202. Springer.

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.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. 2025. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *Preprint*, arXiv:2504.13837.

A Prompt Template for Data Generation

We use a detailed prompt to instruct GPT-4o during the initial data generation phase. This prompt guides the model to construct a tree-structured dialogue rooted in a given scenario, with three semantically coherent but pragmatically distinct alternative replies. These three replies will then be further paired to construct three datapoints that have two alternative replies. It also requests natural language justifications that explain each reply's pragmatic function and its potential conversational effect. For all evaluation tasks, we use four H100 GPU. We obtain all open-source models from HuggingFace. The full prompt used in generation is shown below.

Prompt Template for Data Generation

Task Overview: In this task, we will explore **pragmatic alternatives** in conversation by constructing a multi-round, **tree-structured** dialogue. The goal is to analyze why a speaker might choose one alternative over another based on **pragmatic effects** rather than purely semantic differences.

Guidelines for Generating the Conversation:
1. Conversation Structure (Tree)
- The conversation will expand in a **1-3** structure (4 sentences total).
- The second turn introduces **three alternative responses**, each reflecting a **distinct pragmatic intention**.
- The conversation alternates speakers:
- Root: A initiates the conversation.
- Second layer: B responds with three alternatives (B1, B2, B3).

2. Pragmatic Diversity (Not Constrained to Specific Effects)
- Each pair of responses must be **semantically similar but pragmatically different**.
- Responses should vary across a **broad range of pragmatic effects**, such as:
- **Politeness & Indirectness** (*e.g.,* "Could you open the window?" vs. "Open the window.")
- **Emphasis & Focus** (*e.g.,* "I think you did great." vs. "You did great, no doubt about it.")
- **Formality & Register** (*e.g.,* "He passed away." vs. "He died.")
- **Tone & Emotion** (*e.g.,* "I'm not sure." vs. "I don't think so.")
- **Hedging & Certainty** (*e.g.,* "It might work." vs. "This will definitely work.")
- **Conversational Strategies** (*e.g.,* "That's a good idea!" vs. "I see where you're coming from.")
- **Avoid overly simplistic or directly opposing responses** (e.g., "Yes." vs. "No.").
- **Make sure the conversation not being too neutral, be drastic if necessary.**
- **Strictly avoid repeating the context or simple agreement.** E.g., if context says A, avoid saying "yes, it's A". Be natural!

3. Violation of Gricean Maxims for Each Alternative
Sometimes, speakers intentionally violate these maxims to imply something else (implicature). For example:
- **Flouting Quantity:** Giving too much or too little information to imply something (e.g., "Some students passed the exam" implies not all did).
- **Flouting Quality:** Using sarcasm or irony (e.g., "Oh, great! Another surprise quiz!" when unhappy).
- **Flouting Relation:** Responding indirectly to suggest something (e.g., A: "Did you finish the report?" B: "I had a really busy weekend.").
- **Flouting Manner:** Being intentionally vague (e.g., "Let's just say things didn't go as planned.").
- **Ensure at least 2 of the 3 alternatives contain a violation of Gricean Maxims.**

4. Sentence-Level Intentions
Each response must explicitly state its **pragmatic intention** in the JSON output:
- **GM:** Which Gricean Maxim type does the response flout. If none, type "None".
- **Intention:** Explanation including:
- How the three alternatives differ pragmatically.
- When or why a speaker might choose each alternative.
- The effect each alternative has on the conversation (*e.g.,* softens tone, strengthens commitment, redirects focus).

5. Demonstration
Below is an example of the JSON output:
Context: X and Y are colleagues who are leaving work on a Friday at the same time.
Question: Do you want me to give you a lift home?
Root: Do you want me to give you a lift home?
B1: Oh, that would be great! I was dreading the crowded train.
B2: Well, I suppose if you're absolutely sure my presence won't ruin your evening...
B3: Nah, I could use the walk. Been sitting all day.
Intentions:
B1:
GM: None
Intention: Directly accepts and expresses gratitude. Enthusiastic tone. A speaker might use it to show appreciation and comfort.
B2:
GM: Quality
Intention: Uses sarcasm and self-deprecation. Adds humor while implying acceptance. Useful for playful or noncommittal tone.
B3:
GM: Relation
Intention: Deflects the offer indirectly. Politely declines without confrontation. Useful for maintaining autonomy while being polite.

6. JSON Output Requirements
- `"Context"`: The paraphrase of the given text, but do not shorten it.
- `"Root"`: The initial statement (A's opening statement).
- `"B1"`, `"B2"`, `"B3"`: The three replies to the root.
- `"Intentions"`: Each reply must include `"GM"` and `"Intention"`.
- Ensure responses are semantically similar but pragmatically distinct.
- Do not output anything outside of the specified JSON format.

7. Final Instruction
When ready, check the conversation again to ensure all responses convey distinct pragmatic functions. If not, revise them and their associated intentions. **ONLY OUTPUT THE JSON FILE.** Do not include any explanations or additional commentary.

B Prompt Templates for Response Generation

To obtain model's pragmatic analysis of each datapoint, we adopt the full assistant prompt format proposed by Lin et al. (2023), including the instruction preamble and example completions. The only component we add is the final task-specific question shown below:

Prompt Segment (Appended after URIAL Prompt)

Your task is to examine the following short conversation and assess:
- What is the pragmatic intention behind 'response_1'?
- Why or when might someone prefer 'response_1' over 'response_2' pragmatically?
Please answer in 1 paragraph.
Answer:

C Hyperparameter Settings for Response Generation

To ensure comparability across models and avoid extraneous variance, we apply a unified set of decoding parameters for all generations, regardless of model architecture or size. The configuration is summarized below:

Generation Settings (Applied to All Models)

`max_new_tokens`	256
`do_sample`	True
`top_k`	50
`top_p`	1.0
`temperature`	0.5
`repetition_penalty`	1.0

This configuration was consistently applied to all evaluated models, including OLMo-2, OLMoE, LLaMA-3.1-Tülu-3, and Qwen-3 families.

D Prompt Templates for 10-Point Scoring

In this setting, we evaluate the quality of intention explanations generated by models in different stages. To assess alignment with the intended pragmatic goal, we compare each model output to a human-annotated reference and ask GPT-4.1 to assign a score between 1 and 10, or "Invalid" if the response is incoherent. This helps filter out degenerate completions common in base models. Data points marked "Invalid" are excluded during score aggregation.

Prompt Template for 10-Point Scoring

You are an expert evaluator of language model responses. Your task is to score how well a model's response captures the idea of a response to a given sentence.

The tested model will see a conversation structure like this:
"context": ...
"initial turn": ...
"response_1": ...
"response_2": ...
And they are asked to evaluate:
What is the intention behind 'response_1'?
Why or when might someone prefer 'response_1' over 'response_2'?

We have the reference golden intention for you: ..., this is the core idea of the intention of response_1.
Model's response: ...
You should understand that the tested models are base models, thus not deliberately instructed to format their responses and may output random strings and sentences. You just need to check the content of their responses instead of the format.

Score the model's response on a scale of 1 to 10, or Invalid, where:
- 10: The model's response perfectly captures the intention described in the reference
- 1: The model's response poorly captures the intention described in the reference
- Invalid: The model's response is nonsense or invalid
Return ONLY a JSON object with the following format:
{{
"score": <number between 1 and 10 or Invalid>,
"reason": "<brief explanation of your score, no more than 25 words>"
}}
Do not include any other text, just the JSON object.

E Prompt Templates for Pairwise Comparison

We use the following prompt to evaluate two model explanations for a single datapoint. GPT-4.1 is asked to choose which model better captures the speaker's intention and to classify the difference into one of three pragmatic categories. If either

model's response is incoherent or if no clear winner can be determined, GPT-4 is instructed to return "Invalid".

Prompt Template for GPT-4 Comparison

You are an expert evaluator of language model analysis. Your task is to compare two model analysis of the intention of response_1,
which is a response to the "initial turn", and choose which one better aligns with the golden label's intention.

The conversation structure is:
"context": ...
"initial turn": ...
"response_1": ...
"response_2": ...
"golden_label": ...

The models are asked to evaluate:
What is the intention behind 'response_1'?
Why or when might someone prefer 'response_1' over 'response_2'?

Model 1's response: ...
Model 2's response: ...

Choose which response better captures the intention described in the golden label.

After comparing the two responses, you also need to categorize how the better response
is better than the other.

Here are the categories:
Category #1: Cognitive-Pragmatic Competence
description: The better response goes beyond the literal meaning of the sentence and
identifies the speaker's underlying social goal, such as softening a refusal or signaling indirect disagreement.

Category #2: Pragmalinguistic Competence
description: The better response identifies and explains rhetorical techniques—like humor,
irony, or self-deprecation—and clarifies how these strategies function to manage emotion
or social tension.

Category #3: Sociopragmatic Competence
description: The better response shows sensitivity to the social context, including roles,
relationships, or timing, and explains why the speaker's choice fits the situation appropriately.

If you feel the better response does not fall into any of the categories,
you can choose "Invalid", and explain how it better than the other in the "reason" field.

Return ONLY a JSON object with the following format:
{{
"choice": "1" or "2" or "Invalid",
"reason": "<brief explanation in 20 words or less>",
"category": "1" or "2" or "3" or "Invalid"
}}

You should understand that the tested models might be base models,
thus not deliberately instructed to format their responses and may output
random strings and sentences. You need to purely focus on the content and
compare them instead of having preference on the format.

Return "Invalid" if:
- Either response is empty, nonsense, unrelated to the question, or
as long as you feel it is not answering the question
- You cannot make a clear choice between the two responses

F Examples from 10-Point Scoring Evaluation

To better illustrate how we assess explanation quality in our 10-point scoring evaluation, we present representative examples from the evaluation process. Each example includes the conversation context, potential responses, the human-annotated reference explanation (`response_1_intent`), the model-generated explanation, and the resulting evaluation score and rationale.

Example of 10-Point Scoring Evaluation A: LLaMA-3.1-70B Base

`context:` John is a teacher at an elementary school. When talking with the principal about a new student, who did poorly on her entrance examination, John said, "This one is really sharp."

`initial_turn:` This one is really sharp.

`response_1:` I think she will fit right in with some guidance.

`response_2:` She's just like any other student who needs a bit of encouragement.

`response_1_intent:` This response does not flout any maxim. It aims to express optimism about the student's ability to succeed with proper support and highlights that she has potential. A speaker might prefer this reply over response_2 because it directly affirms the student's strengths and offers a solution, presenting the student in a positive and proactive light. This approach helps reinforce a sense of capability and possibility.

`model_response:` The pragmatic intention behind response_1 is to express optimism and confidence in the new student's ability to succeed with some guidance. This statement implies that the student has potential and can thrive in the school environment with the right support.
Someone might prefer response_1 over response_2 pragmatically in situations where they want to emphasize the student's potential and the importance of providing guidance to help them succeed. response_1 focuses on the student's strengths and the teacher's belief in their ability to fit in and do well, while response_2 takes a more neutral stance and simply acknowledges that the student needs encouragement like any other student.
In summary, choosing response_1 over response_2 pragmatically can help communicate a more positive and supportive attitude towards the new student, highlighting their potential and the teacher's confidence in their ability to succeed with the right guidance.

`scoring:`

Score: 10
Reason: Accurately explains the optimistic intention, focus on student potential, and why one might prefer this over a more neutral response.

Example of 10-Point Scoring Evaluation B: Olmo2-7B Base

`context:` Callie promised to bring a salad to Jason's dinner party. She arrived two hours late. Jason said, "I'm so glad that you were able to fit us into your busy schedule."

`initial_turn:` I'm so glad that you were able to fit us into your busy schedule.

`response_1:` Well, it's a good thing I'm fashionably late, right?

`response_2:` I know, right? Had to fight through traffic, but I'm here now.

`response_1_intent:` This response flouts the maxim of Quality by playfully embracing the notion of being "fashionably late" instead of seriously addressing the tardiness. It aims to use humor and irony to diffuse any awkwardness and avoid direct confrontation. A speaker might prefer this reply over response_2 because it turns the situation into a joke, attempting to lighten the mood rather than provide an explanation. This approach could help smooth over tension if the relationship is friendly enough for such banter.

`model_response:` In this context, the pragmatic intention behind the response_1 is to express gratitude and appreciation for the other person's effort in accommodating the delay. This is because the speaker, Jason, is acknowledging that Callie was able to attend the dinner party despite being late. The speaker may prefer response_1 over response_2 pragmatically because it is more lighthearted and humorous, which can help to ease the tension or awkwardness that often arises from being late. Additionally, response_1 does not explicitly acknowledge the reason for the delay, which may be considered more polite in certain contexts. On the other hand, response_2 explicitly states the reason for the delay, which can be seen

as more direct and assertive. This can be more appropriate in certain situations, such as when the speaker wants to emphasize the importance of punctuality or when they are trying to avoid any misunderstandings about the reason for the delay. Overall, the pragmatic intention behind response_1 is to express gratitude and appreciation for the other person's effort, and it may be preferred in certain situations due to its lighthearted and humorous tone.

`scoring:`
Score: 5
Reason: Captures the humor and lightheartedness, but incorrectly says the intention is gratitude/appreciation, missing the playful, ironic deflection central to the reference.

G Examples from Pairwise Comparison Evaluation

To further illustrate how different models perform in pragmatic reasoning, we present selected examples from our pairwise comparison evaluation. In this setting, two model-generated responses are compared against a human-annotated pragmatic interpretation (`response_1_intention`) to determine which aligns better with the intended meaning. For the winning explanation, we further categorize it into one of three pragmatic dimensions as discussed in Section 4.3.

We provide one illustrative example per category to showcase the types of reasoning improvements observed in our evaluations.

Pairwise Comparison Example A: LLaMA-3.1-70B Base vs SFT

`context:` A boy from New York was being led through the swamps of Florida and is curious about whether alligators will attack if he carries a flashlight.

`initial_turn:` Is it true that an alligator won't attack you if you carry a flashlight?

`response_1:` Well, let's just say alligators have their own mind and aren't that predictable.

`response_2:` That's actually just a myth, but keeping one handy is a good idea for other reasons.

`response_1_intention:` This response flouts the maxim of Relation. It aims to answer indirectly, emphasizing the unpredictability of alligators rather than directly addressing the myth. A speaker might prefer this reply over response_2 because it introduces a sense of caution and intrigue, perhaps leaving the listener more alert or curious. The indirectness could serve to prompt further questions or reinforce respect for wildlife unpredictability.

`model_1_response:` The pragmatic intention behind 'response_1' is to express uncertainty, skepticism, or a lack of definitive knowledge about the behavior of alligators. By saying "alligators have their own mind and aren't that predictable," the speaker is suggesting that it's difficult to make general statements about how alligators will react in specific situations. This may be a more appropriate response if the speaker is unsure of the veracity of the claim or wants to avoid making a definitive statement about something they don't know for certain.
Someone might prefer 'response_1' over 'response_2' pragmatically in situations where they want to express uncertainty or skepticism, or when they don't want to make a definitive statement about something they aren't sure of. This could be particularly relevant in the context of the conversation, where the speaker is being asked about a potentially dangerous situation and may want to avoid giving misleading information.

`model_2_response:` The pragmatic intention behind 'response_1' is to express uncertainty and caution about the predictability of alligators' behavior. The speaker may prefer 'response_1' over 'response_2' when they want to emphasize the unpredictable nature of alligators and the importance of being cautious in general, rather than focusing on the specific myth about flashlights. This approach may be more appropriate when the speaker wants to encourage the listener to be vigilant and aware of potential dangers in the swamps, regardless of the presence of a flashlight.

`comparison_result:`

Choice: 2
Reason: Model 2 highlights indirectness and caution, matching the golden label's focus on unpredictability and vigilance.
Category: 1

Pairwise Comparison Example B: LLaMA-3.1-8B Base vs SFT

`context`: X wants to know about Y's food preferences by asking if Y enjoys Chinese cuisine.

`initial_turn`: Are you into Chinese cuisine?

`response_1`: Well, I guess I've survived a couple of meals.

`response_2`: It's okay, but I'm more of a Mediterranean fan.

`response_1_intention`: This response flouts the maxim of Quality by using sarcasm to imply only a minimal or reluctant enjoyment of Chinese food. It aims to add humor and ambiguity, downplaying direct expression of preference. A speaker might prefer this reply over response_2 to keep the conversation playful and avoid direct criticism, especially if they want to maintain a light tone.

`model_1_response`: The pragmatic intention behind 'response_1' is to provide a noncommittal answer to the question about liking Chinese cuisine. It is a way to avoid giving a direct answer, which might be interpreted as a personal preference. 'response_1' is preferred over 'response_2' pragmatically because it offers a more neutral and noncommittal response. By saying "I guess I've survived a couple of meals," the speaker is implying that they have had some experience with Chinese cuisine, but they are not necessarily a fan of it. This response allows the speaker to save face and avoid appearing too negative about Chinese cuisine, while still conveying that they may not be a big fan. Additionally, it leaves room for the conversation to continue without being closed off by a definitive answer.

`model_2_response`: The pragmatic intention behind 'response_1' is to express a lack of strong preference or enthusiasm for Chinese cuisine, possibly indicating a neutral or indifferent attitude towards it. Someone might prefer 'response_1' over 'response_2' pragmatically in a situation where they do not want to completely rule out the possibility of enjoying Chinese food, but also do not want to commit to a strong positive stance. 'response_1' could be considered a middle ground or a way to provide a non-committal response, which might be a preferred strategy in polite conversation or when the speaker does not want to appear overly critical or dismissive of the food. This response can be useful in maintaining a neutral tone and avoiding potential awkwardness or conflict in the conversation.

`comparison_result`:
Choice: 1
Reason: Model 1 recognizes the use of humor and indirectness, aligning with the golden label's focus on sarcasm and playfulness.
Category: 2

Pairwise Comparison Example C: LLaMA-3.1-7B SFT vs DPO

`context`: On his first flight while learning to skydive, a man nervously anticipated his first jump, so he said to his instructor, "I packed the parachute myself, but I'm sure it won't open."

`initial_turn`: I packed the parachute myself, but I'm sure it won't open.

`response_1`: Don't worry, we've double-checked it. You're in safe hands.

`response_2`: You'll be just fine. And if not, it's been nice knowing you!

`response_1_intention`: This response does not flout any maxim. It aims to provide direct reassurance to the nervous jumper by emphasizing that the equipment has been double-checked and he is safe. A speaker might prefer this reply over response_2 because it maintains a calming and professional tone, offering clear comfort and build-

ing trust in a potentially stressful situation.

`model_1_response`: The pragmatic intention behind "Don't worry, we've double-checked it. You're in safe hands." is to reassure the speaker's partner that he is in good care and that the parachute has been thoroughly checked, which would be preferred over response_2 pragmatically because it provides a sense of security and confidence in the safety measures that are in place, which is important in a potentially dangerous situation like skydiving. This sentence would be more comforting to the partner and would also serve to alleviate any concerns the partner may have about the safety of the parachute.

`model_2_response`: The pragmatic intention behind the sentence "Don't worry, we've double-checked it. You're in safe hands." ('response_1') is to provide reassurance and comfort to the nervous skydiver. This sentence is chosen to pragmatically offer support and confidence, alleviating the skydiver's anxiety about the parachute. It's likely preferred because it communicates a sense of professionalism and thoroughness in safety checks, thereby reassuring the skydiver that all precautions have been taken. This sentence is appropriate when someone needs to be comforted and needs to feel secure in a potentially risky situation. In contrast, "You'll be just fine. And if not, it's been nice knowing you!" ('response_2') is a more lighthearted and possibly flippant response that might not provide the necessary comfort or professionalism that the skydiver is seeking in this context. While 'response_2' might be used in a more joking or casual setting, it lacks the assurance and emphasis on safety that the skydiver needs to feel secure about jumping out of a plane.

`comparison_result`:
Choice: 2
Reason: Model 2 explains both reassurance and professionalism, matching the golden label's emphasis on tone and trust.
Category: 3

H Pairwise Comparison Category Distributions

To supplement our main findings, we present the full set of pairwise comparison results across all evaluated model pairs. Each figure below visualizes the distribution of winning explanations across three pragmatic competence categories.

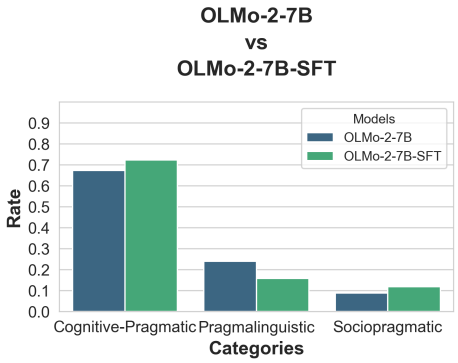


Figure 6: OLMo-2-7B Base vs SFT.

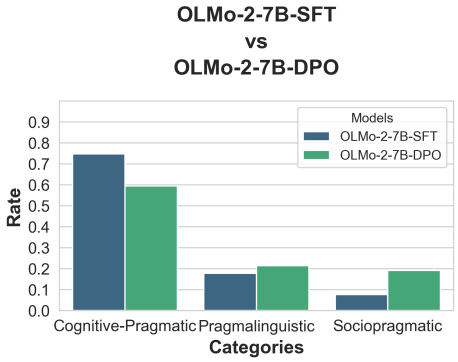


Figure 7: OLMo-2-7B SFT vs DPO.

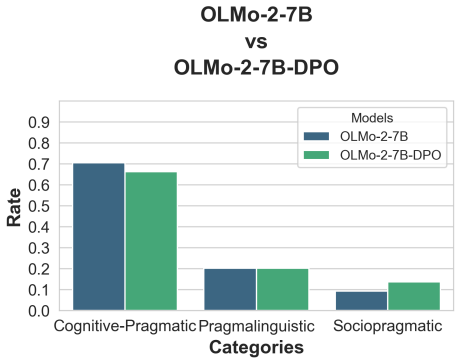


Figure 8: OLMo-2-7B Base vs DPO.

I Generative AI Statement

We use generative AI tools to assist with both the implementation and writing processes in this

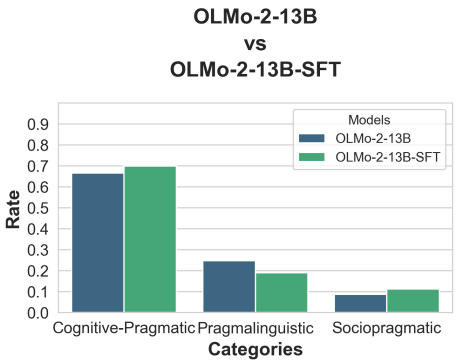


Figure 9: OLMo-2-13B Base vs SFT.

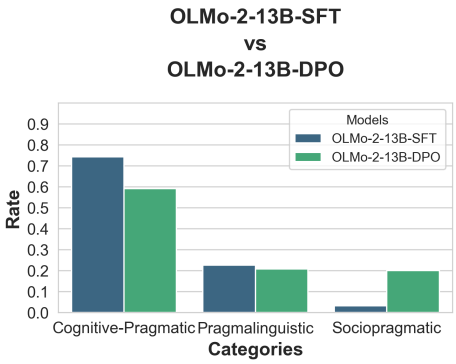


Figure 10: OLMo-2-13B SFT vs DPO.

project. Specifically, we employed **Cursor**, an AI-assisted development environment, and **ChatGPT** (GPT-4o) to support the coding of evaluation tasks. Additionally, ChatGPT was used to aid in formatting sections of the paper, as well as generating LaTeX tables and figure templates. All outputs were carefully reviewed, edited, and verified by the authors to ensure factual accuracy and scholarly integrity.

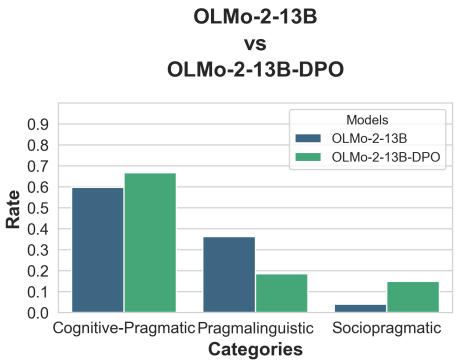


Figure 11: OLMo-2-13B Base vs DPO.

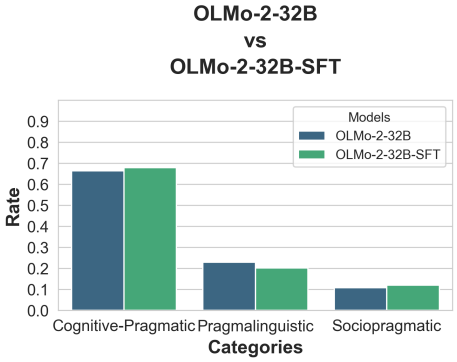


Figure 12: OLMo-2-32B Base vs SFT.

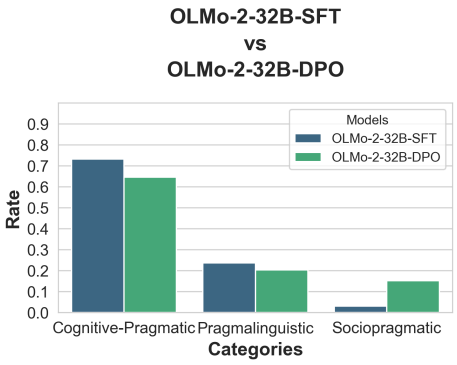


Figure 13: OLMo-2-32B SFT vs DPO.

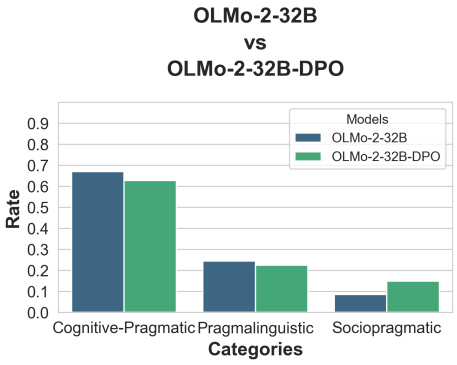


Figure 14: OLMo-2-32B Base vs DPO.

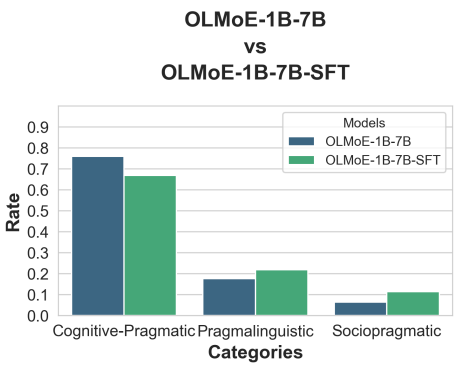


Figure 15: OLMoE-1B-7B Base vs SFT.

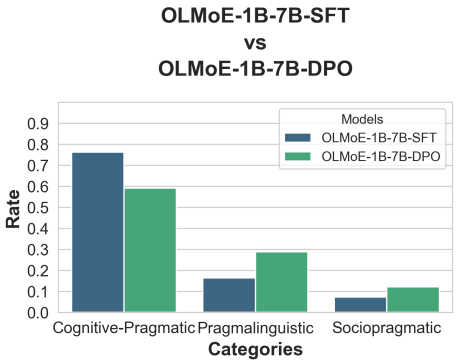


Figure 16: OLMoE-1B-7B SFT vs DPO.

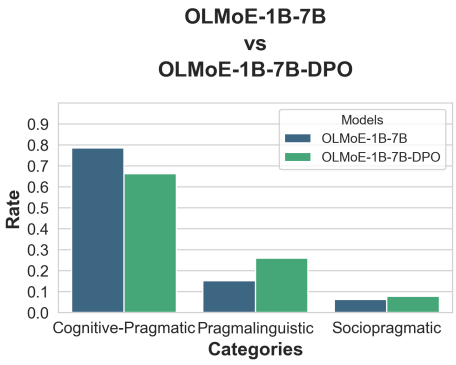


Figure 17: OLMoE-1B-7B Base vs DPO.

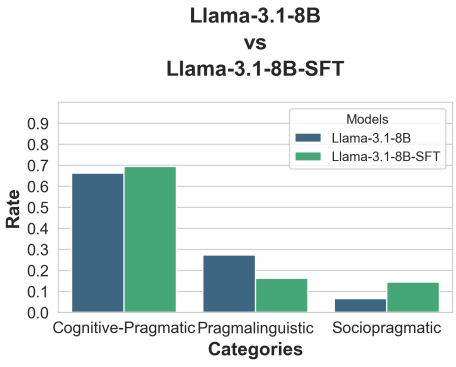


Figure 18: LLaMA-3.1-8B Base vs SFT.

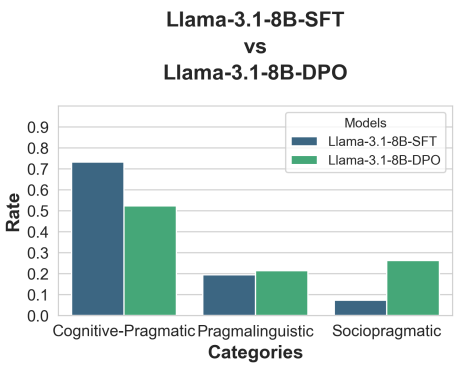


Figure 19: LLaMA-3.1-8B SFT vs DPO.

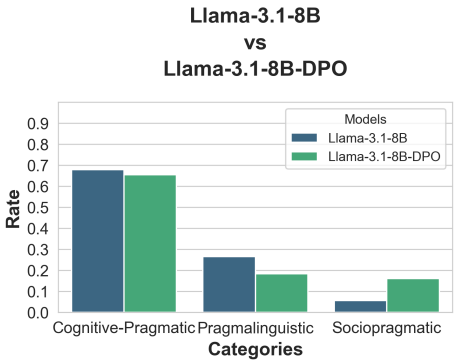


Figure 20: LLaMA-3.1-8B Base vs DPO.

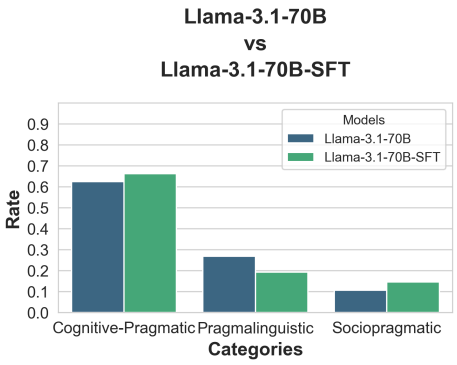


Figure 21: LLaMA-3.1-70B Base vs SFT.

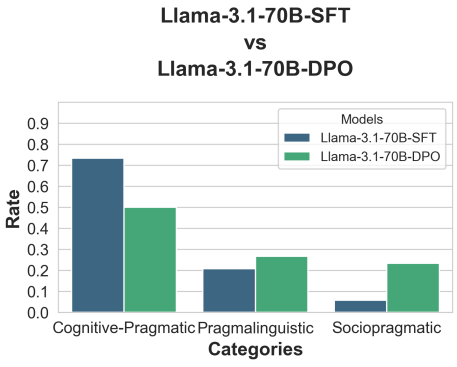


Figure 22: LLaMA-3.1-70B SFT vs DPO.

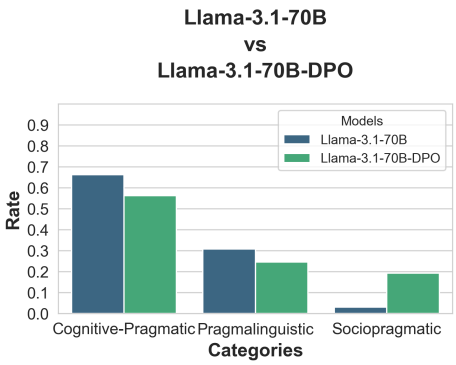


Figure 23: LLaMA-3.1-70B Base vs DPO.