Rethinking Pragmatics in Large Language Models: Towards Open-Ended Evaluation and Preference Tuning

Shengguang Wu[1], Shusheng Yang[2], Zhenglun Chen[1*], Qi Su[1†]
[1]Peking University [2]Huazhong University of Science and Technology
wushengguang@stu.pku.edu.cn, shushengyang@hust.edu.cn, danielchenbj@gmail.com, sukia@pku.edu.cn

Abstract

This study addresses the challenges of assessing and enhancing social-pragmatic inference in large language models (LLMs). We first highlight the inadequacy of current accuracy-based multiple choice question answering (MCQA) formats in assessing social-pragmatic reasoning, and propose the direct evaluation of models' free-form responses as measure, which correlates better with human judgment. Furthermore, we explore methods to improve pragmatic abilities in LLMs, advocating for preference optimization (PO) over supervised finetuning (SFT), given the absence of a definitive "gold" answer in social contexts. Our results show that preferential tuning consistently outperforms SFT across pragmatic phenomena and offers a near-free launch in pragmatic abilities without compromising general capabilities. Lastly, we examine the internal structure of LLMs, revealing that the significant boost in pragmatic reasoning is tied to deeper layer representations, analogous to human high-level thinking. Our experiments span a variety of pragmatic and social reasoning datasets, as well as an image referential game requiring a multimodal theory of mind (ToM). With our refined paradigms for evaluating and enhancing pragmatic inference, this paper offers key insights into building more socially aware language models.

1 Introduction

Social-pragmatic inference is a key aspect of human communication, involving the ability to understand and respond to implied meanings, intentions, and emotional states behind literal utterances (Horn, 1972; Grice, 1975; Green, 1998; Carston, 2004), as well as shared social conventions (Goffman, 1959). Pragmatics spans a broad range of phenomena, including implicatures, irony, humor, and metaphor, along with higher-level cognitive skills like theory of mind (ToM) (Premack and Woodruff, 1978), which are essential for interpreting non-literal language and context-sensitive messages. For example, a friend's statement, *It's chilly in here*, might be a polite request to close a window, rather than a simple observation about the temperature.

The importance of social-pragmatic intelligence in human communication underscores the need for large language models (LLMs) to develop similar capabilities in order to interact more naturally with users. However, current approaches to enhancing pragmatic abilities in LLMs face two lines of limitations: **1)** On the evaluation front, typical methods rely on measuring classification accuracy on benchmarks formatted as multiple (if not binary) choice question answering (MCQA) (Le et al., 2019; Ruis et al., 2023; Hu et al., 2023; Zhou et al., 2023; Gandhi et al., 2023; Sravanthi et al., 2024). While a model might correctly select the option label, it may still fail to respond pragmatically by itself. For example (see Fig. 1), a model might pick the right answer in an MCQA task without fully understanding the social complexity of *changing the subject*. Moreover, real-world social interactions seldom have a single "gold" answer. Relying on accuracy in selecting predefined responses undermines the assessment of a model's true pragmatic capability in flexible generations. **2)** On the side of improving pragmatic abilities, while inference-time methods such as few-shot prompt engineering (Moghaddam and Honey, 2023; Ruis et al., 2023) and external graph modules (Sclar et al., 2023) have been proposed to improve LLM performance in pragmatic tasks, little effort has been made to directly invoke the model's internal social-pragmatic intelligence, enabling it to autonomously generate pragmatically appropriate responses.

In this paper, we propose paradigm shifts on both fronts: **1)** For evaluation, we argue for an open-

[*]Work done during internship at Peking University.
[†]Corresponding author.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 22583–22599
November 12-16, 2024

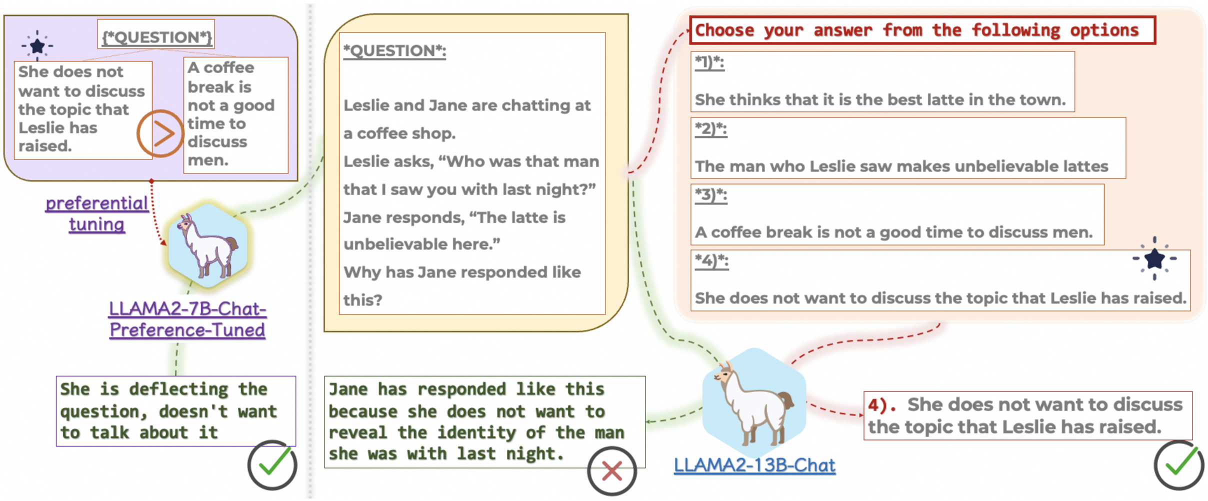

Figure 1: An example of LLM outputs when queried about a social-pragmatic scenario, taken from Hu et al. (2023). On the right, a LLAMA2-**13B**-Chat (Touvron et al., 2023) model correctly identifies the gold response ID in the MCQA format but fails to fully grasp the underlying pragmatic meaning when generating its own response. On the left, a smaller LLAMA2-**7B**-Chat model preference-tuned to contrast the gold answer with less pragmatic alternatives, produces an open-ended response that is equally good and as pragmatically sound as the provided "gold" answer.

ended assessment protocol that directly evaluates a model's ability to respond to social scenarios. We introduce the **Length-Normalized Relative Score** ($LNRS$) that rates a model's free-form response relative to the provided "gold" answer, with GPT-4 (OpenAI, 2023) as the judge. This scoring system is further de-biased to reduce length gameability (Dubois et al., 2024; Galambosi, 2024). Backed by human evaluation, our open-ended metric $LNRS$ is better correlated with human preferences than MCQA accuracy. **2)** For improving LLMs' pragmatic inference, we treat the non-selected answer options in MCQA-formatted datasets not as *incorrect*, but as *less pragmatically grounded* compared to the "gold" answer. We use preference optimization (PO) objectives, such as DPO (Rafailov et al., 2024), to finetune LLMs, allowing them to capture subtle nuances of pragmatic preferences. Our experiments show that preferential tuning yields significantly better results than conventional supervised finetuning (SFT) across pragmatic phenomena, with minimal impact on the model's other abilities inherited from the base LLM. Additionally, in the multimodal setting of the image referential game (Corona et al., 2019; Zhu et al., 2021; Liu et al., 2023) that explicitly requires theory of mind (ToM) (Premack and Woodruff, 1978), PO also results in a more capable, ToM-aware vision-language speaker model, which further demonstrates its superiority over SFT for enhancing pragmatic abilities.

To better understand how the internal components of a transformer-based LLM (Vaswani et al., 2017) are responsible for invoking social-pragmatic abilities, we explored finetuning specific transformer layers. Our results indicate that pragmatic understanding is closely tied to deeper-down layers in the model, which hints at a potential parallel with how human pragmatic inference relies on higher-level cognitive processes.

Overall, the main contributions of this paper are:

• Proposing open-ended evaluation of models' free-form responses instead of MCQA classification for assessing social-pragmatic understanding, which better aligns with human judgment;

• Proposing preference optimization (PO) over supervised finetuning (SFT) for improving LLMs' pragmatic abilities without degrading other core capabilities, as demonstrated through experiments across various pragmatic datasets and the multimodal theory of mind (ToM) task;

• Providing empirical insights into how only training deeper layers of LLMs can invoke significant gains in pragmatic performance, which potentially mirrors human high-level cognitive thinking.

2 Evaluating Pragmatic Abilities

2.1 Existing Evaluation

Existing works primarily assess a language model's pragmatic intelligence through multiple (or binary) choice question answering (MCQA) tasks. In such settings, for a given social scenario, the model must select an answer from a set of options (Le et al., 2019; Ruis et al., 2023; Hu et al., 2023; Zhou et al., 2023; Gandhi et al., 2023; Sravanthi et al., 2024), and the **accuracy** of choosing the annotated "gold" answer is used to gauge the model's pragmatic abilities ($MCQA$-Acc). In recent studies, the way to elicit a model's choice from the provided options can be generally divided into two categories:

• Metalinguistic[1] Probing: The model is explicitly prompted to choose from a set of answers linked to symbolic indicators, such as alphabetic letters (`A|B|C|D`) (Le et al., 2019; Sravanthi et al., 2024; Robinson and Wingate, 2023) or numerical indices (`1|2|3|4`) (Hu et al., 2023). The model then generates the corresponding symbol for the selected option.

• Probability Probing: The model is given the scenario and question text (context, $\mathbf{x}$), and we compute the likelihood of the model generating each answer option $\mathbf{y}_i$ conditioned on the context. The option with the highest probability is considered the model's choice. There are several normalization techniques for probability calculation (Brown et al., 2020; Robinson and Wingate, 2023; Holtzman et al., 2021), leading to different formulations: without normalization: $P\left(\mathbf{y}_i \mid \mathbf{x}\right)$; with length normalization over j tokens in $\mathbf{y}_i$: $\frac{\sum_{j=1}^{\ell_i} P\left(y_i^j | \mathbf{x}, \mathbf{y}^{1 \cdots j-1}\right)}{\ell_i}$; and with normalization by unconditional answer probability[2]: $\frac{P(\mathbf{y}_i|\mathbf{x})}{P(\mathbf{y}_i|\mathbf{x}_{\text{uncond}})}$.

These accuracy-based MCQA evaluations have several key limitations: **1)** This format diverges significantly from real-world social interactions, where no fixed answer exists. Even the "gold" answer provided in these benchmarks may not be the best response for a given scenario. For example, the preference-tuned model's response in Fig. 1 (left side) is equally valid from a social and pragmatic perspective. **2)** As noted by Robinson and Wingate (2023), different models show varying levels of proficiency in binding an option to its symbol (*multiple choice symbol binding, MCSB*), which can be confused with true pragmatic intelligence, particularly in the metalinguistic probing approach. **3)** Identifying the correct answer option does not necessarily mean the model understands the social scenario or can respond in a socially and pragmatically appropriate manner on its own (see the right side of Fig. 1), which is the actual ability desired for real-world human-LLM interactions.

For these reasons, we argue for a shift in the evaluation of machine pragmatics towards an **open-ended** assessment of the model's autonomous response, while keeping the annotated "gold" answer as a reference.

2.2 Open-Ended Evaluation

We introduce **Length-Normalized Relative Score** ($LNRS$) to quantitatively assess how well a model's own response compares to the provided "gold" answer. Rather than giving the model a set of options, we directly obtain its free-form response to the pragmatic question describing a social scenario. Then, we query GPT-4 (OpenAI, 2023) to score the model's response relative to the annotated "gold" answer.

GPT-4 Judge. We employ GPT-4 as the judge, because it is the most reliable model available for robust and human-matching performance across various social-pragmatic tasks (Gandhi et al., 2023; Sap et al., 2023; Zhou et al., 2023; Ruis et al., 2023; Kosinski, 2023). Additionally, GPT-4 has been widely used in numerous automatic settings, such as instruction-following evaluations (Chiang et al., 2023; Li et al., 2023; Dubois et al., 2024, 2023; Wang et al., 2023a), and even as a "teacher" for guiding other LLMs in reasoning tasks (Shridhar et al., 2023; Hsieh et al., 2023). To reduce potential position bias, we query GPT-4 twice, reversing the order of the model's answer and the "gold" answer. The prompt template for querying GPT-4 (`gpt-4-1106-preview`) is provided in Appx.A.

After parsing GPT-4's responses into pairs of scores, we compare the average score of the model's response to that of the "gold" answer. For all test questions T, we compute the **Relative Score** (RS) of the model's response a_{model} with respect to the "gold" answer a_{gold} as

$$RS = \frac{\sum_{q \in T} \mathrm{JS}(a_{model})}{\sum_{q \in T} \mathrm{JS}(a_{gold})}$$

where JS is the GPT-4 judge's score. This measures how closely the model's responses align with or even surpass the quality of the "gold responses, reflecting the model's understanding of social norms and pragmatic rules.

Length Normalization. Inspired by recent advancements in LLM evaluation, such as AlpacaEval-2.0 (Dubois et al., 2024; Galambosi,

[1]Term adopted from Hu and Levy (2023), also known as *multiple choice prompting* in Robinson and Wingate (2023).

[2]Also referred to as *domain conditional point-wise mutual information* by Holtzman et al. (2021).

2024), we carefully control for the influence of response length on GPT-4's judgment (referred to as *length gameability* in Dubois et al. (2024)). We adopt a *logistic length normalization* technique (Galambosi, 2024; Dubois, 2024)[3] for our open-ended evaluation. Specifically, the Length-Normalized Relative Score ($LNRS$) adjusts the RS by applying a temperature-weighted sigmoid function to the length difference between the model's and the "gold" response:

$$LNRS = \frac{\sum_{q \in T} \mathrm{JS}(a_{model})}{\sum_{q \in T} \mathrm{JS}(a_{gold})} \cdot \sigma\left(\frac{1}{\tau \cdot T} \sum_{q \in T} (\mathrm{Len}(a_{gold}) - \mathrm{Len}(a_{model}))\right) \tag{1}$$

where τ is a temperature hyperparameter, and JS and Len represent the judge's score and the token length, respectively.

In §4.1, we empirically demonstrate that $LNRS$ outperforms $MCQA\text{-}Acc$, showing a stronger correlation with real user preferences, as confirmed by our human evaluation.

3 Improving Pragmatic Abilities

On top of the open-ended evaluation paradigm that more closely reflects real-world scenarios, we also aim to explore how to intrinsically enhance the social-pragmatic capabilities of LLMs. Different from previous works (§5) that primarily focus on adding external modules for better cognitive abilities (Sclar et al., 2023; Takmaz et al., 2023) or rely on few-shot prompt engineering (Moghaddam and Honey, 2023; Ruis et al., 2023), our approach is centered on aligning the model's intrinsic representation toward a more socially and pragmatically grounded distribution.

Let $\mathbf{p}_\theta$ represent an LLM parameterized by θ. In our context, $\mathbf{p}_\theta$ takes a question q as input, which describes a pragmatics-involved social scenario, and a_{gold} is the annotated correct response.

Supervised Finetuning (SFT). The straightforward approach is to apply SFT using the question q and the gold answer a_{gold} from each MCQA-formatted data source $\mathcal{D}$. The objective here is to minimize the negative log-likelihood loss for predicting each token in the gold answer a_{gold} conditioned on the question q:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(q, a_{gold}) \sim \mathcal{D}} \left[\log \mathbf{p}_\theta(a_{gold}|q)\right] \tag{2}$$

While SFT is a simple and widely used method, it does not allow the model to discern between nuanced, socially acceptable responses, but instead forces the selection of the predefined "gold" answer. This may prevent the model from developing the pragmatic flexibility needed to handle complex social scenarios.

Preference Optimization (PO). In social contexts, there is rarely a single definitive *right* answer. For instance, in MCQA-formatted datasets such as the one in Fig. 1, we might not consider `option 3)` a *wrong* answer, but rather a response that is less socially and pragmatically appropriate than `option 4)`. This nuanced understanding – weighing possible responses based on their pragmatic soundness and social appropriateness – is the kind of reasoning we aim to instill in the model.

To address this, we turn to the preference optimization (PO) paradigm, specifically using the simplified *direct preference optimization (DPO)* objective (Rafailov et al., 2024). Unlike SFT, DPO does not rely solely on maximizing the likelihood of the annotated answer. Instead, it focuses on optimizing the model parameters θ to favor more desirable responses over less desirable ones.

For each question q, we create pairwise triples (q, a_{gold}, a_{other}), where a_{gold} is the provided "gold" and thus preferred response over any other answer option a_{other}. Given a data source $\mathcal{D}$, the PO objective can be formulated as:

$$\mathcal{L}_{\text{DPO}}(\mathbf{p}_\theta; \mathbf{p}_{\text{ref}}) = -\mathbb{E}_{(q, a_{gold}, a_{other}) \sim \mathcal{D}} \left[\log \sigma\left(\beta \log \frac{\mathbf{p}_\theta(a_{gold}|q)}{\mathbf{p}_{\text{ref}}(a_{gold}|q)} - \beta \log \frac{\mathbf{p}_\theta(a_{other}|q)}{\mathbf{p}_{\text{ref}}(a_{other}|q)}\right)\right] \tag{3}$$

where σ is the sigmoid function, and β controls the impact of preference differences.

Compared to SFT, the DPO objective encourages the model to learn to distinguish between responses based on their pragmatic preferences, allowing for more socially grounded reasoning.

[3] The *length control* method used in AlpacaEval-2.0 (Dubois et al., 2024) can not be directly applied to our evaluation without prior win-rate data. So we used *length normalization* that achieves similar performance.

4 Experiments

4.1 Pragmatic Question Answering

Setup. We conducted experiments using four popular social and pragmatic inference data sources – *SOCIAL-IQA* (Sap et al., 2019), *PRAGMEGA* (Floyd, 2022; Hu et al., 2023), *LUDWIG* (Ruis et al., 2023), *PUB* (Sravanthi et al., 2024). These datasets cover a wide range of pragmatic phenomena, including implicature, metaphor, irony, and various social norms. Tab. 4 summarizes the dataset details. We experimented with three base LLMs of varying pretraining data and model sizes: `PYTHIA-6.9B-Tulu` (Wang et al., 2023b), `LLAMA2-7B-Chat`, and `LLAMA2-13B-Chat` (Touvron et al., 2023).[4] Details of the training configurations are listed in Tab. 5.

Human Evaluation. To further support our argument for open-ended assessment of pragmatic abilities, we recruited 12 voluntary human participants from top educational institutions to evaluate the quality of different responses. Given a social-pragmatic context and related question, human evaluators were presented with four types of responses (the dataset-annotated "gold" answer, the base LLM's response, and responses from DPO-tuned and SFT-tuned models) in random order. Evaluators were asked to rank the responses based on their pragmatic understanding and fitness to the context scenario. Detailed instructions used for this study are provided in Appx.B. The ranking of the four responses was converted into scores, with the highest-ranked response receiving 4 points, and the lowest-ranked response receiving 1 point. In total, we randomly sampled 192 data points with the corresponding four responses. Each evaluator was randomly assigned 16 data points for assessment.

Results. Fig. 2, Fig. 5, and Tab. 1 present the performance of LLMs finetuned with different paradigms (PO *vs.* SFT) – evaluated using the open-ended framework (§2.2), the MCQA format[5] (§2.1), and user study (described in the paragraph above). The results reveal the following patterns:

PO-tuned LLMs consistently outperform their SFT-trained counterparts, achieving substantial gains in pragmatic inference over the base models across nearly all configurations of base models, training data, test sets, and evaluation paradigms (MCQA/open-ended/human-eval). There are very few exceptions, such as the marginally lower $LNRS$ score on the *LUDWIG_Test* set for the `PYTHIA-6.9B-Tulu` model DPO-tuned on *PUB* compared to SFT. Additionally, in the MCQA setup, the DPO-tuned `LLAMA2-13B-Chat` underperforms relative to SFT on *PRAGMEGA_Test*, which however contrasts strongly with human evaluations (Tab. 1), where the PO version of `LLAMA2-13B-Chat` is ranked highest in response quality.

The open-ended evaluation paradigm shows better alignment with human judgment than the MCQA results. Tab. 1 clearly demonstrates that humans prefer responses generated by PO-tuned models, which are ranked the best (even surpassing the annotated "gold" answer) for both `LLAMA2` models, and second only to the "gold" answer for `PYTHIA`. In contrast, SFT-tuned models receive lower ratings than their base LLMs, indicating that SFT can even degrade pragmatic performance. These human evaluation findings resonate with the $LNRS$ comparisons in Fig. 2, where similar trends of PO's superiority and SFT's negative impact on pragmatics are observed.

The PO objective facilitates stronger generalization to "out-of-domain" pragmatic phenomena. Our test sets were intentionally designed to include both "in-domain" data (*i.e.*, similar data source and phenomena as the training sets, such as *SOCIAL-IQA_Train/_Test*) and "out-of-domain" data (*i.e.*, different data sources and phenomena from the training sets). We occasionally observe even greater performance gains for PO on data from different sources. For example, on the *SOCIAL-IQA_Test* set, `LLAMA2-13B-Chat` DPO-finetuned on *PUB* (which focuses on implicature, presupposition, etc.) even outperforms the version finetuned on the same social norm dataset.

The PO objective has minimal impact on other abilities inherited from the base LLMs. As shown in Tab. 3, across almost all benchmarks – including professional exams (Hendrycks et al., 2020; Zhong et al., 2023; Clark et al., 2018), math (Cobbe et al., 2021), and reading comprehension (Mihaylov et al., 2018) – models trained with DPO on pragmatic data consistently outperform their SFT counterparts, often by significant margins. This suggests that, despite being finetuned on prag-

[4] We used instruction-tuned chat models as baselines to ensure they started with reasonable instruction-following abilities, especially considering the limited availability of social-pragmatic data, which may not be sufficient for general-purpose alignment tuning.

[5] We used the *length-normalized probability probing* variant in our implementation.

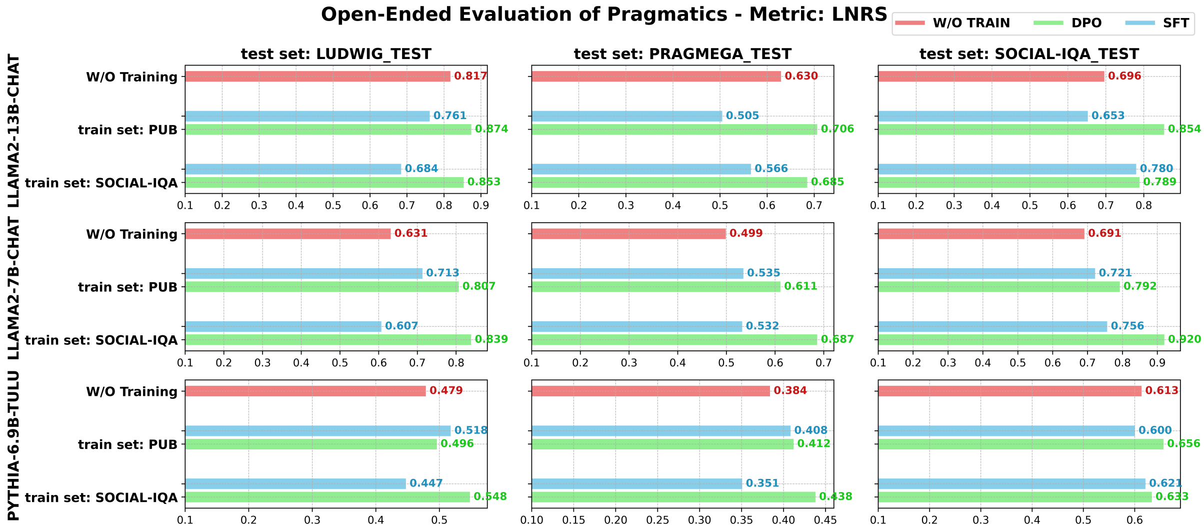

Figure 2: $LNRS$ comparisons across models, data sources, and training paradigms (PO *v.s.* SFT).

matic datasets, the preference-optimized version provides a **near-free launch** of pragmatic abilities, while even improving the various other skills learned by the base models. On the contrary, the SFT-tuned models perform far worse in retaining these inherited abilities.

In addition to the quantitative metric results, we provide **qualitative analyses** in Appx.D. In particular, Tab. 7 presents examples where the model's responses are even better than the reference "gold" answer, as rated by our GPT-4 judge. These examples support our motivational insight that the human-annotated "gold" response might not always be the optimal answer in social-pragmatic scenarios (§1).

Base Models	Base	+SFT	+PO	"Gold"
LLAMA2-7B-Chat	<u>2.75</u>	2.11	**2.81**	2.34
LLAMA2-13B-Chat	2.44	2.05	**2.81**	<u>2.72</u>
PYTHIA-6.9B-Tulu	2.33	2.19	<u>2.66</u>	**2.83**

Table 1: Average human evaluation scores elicited from our user study ranking different responses (§4.1). The **best** and <u>second best</u> results are highlighted.

4.2 Image Referential Game with ToM

In this section, we extend our method for improving models' pragmatic inference from the pure text world (§4.1) to multimodal environments using large vision-language models (LVLMs). We focused on the well-established *image referential game* task (Zhu et al., 2021; Liu et al., 2023; Takmaz et al., 2023), which explicitly requires a theory of mind (ToM) (Premack and Woodruff, 1978) – a key aspect of social-pragmatic capabilities.

Task Formulation. The image referential game involves two interlocutors: a speaker and a listener. Given an image i_{target}, the speaker generates a descriptive caption $c_{speaker}$, which the listener uses to identify the target image i_{target} from a set of images containing both the target and several distractor images $i_{distractor} \in I_{distractor}$. ToM is vividly present in this task, as the speaker must anticipate the listener's understanding and frame the caption in such a way that the listener correctly identifies the target image. Following the methodology from §4.1, we improve the speaker VLM's intrinsic ToM using the same SFT and PO objectives described in §3 and §4.1, with the addition of visual conditions represented by image encodings.

Setup. The base VLM-speaker is implemented as `LLaVA-1.5-7B` (Liu et al., 2024a), while the listener is modeled using the discriminative `OpenCLIP-ViT-B/32` (Ilharco et al., 2021), which matches the target image i_{target} with the speaker's caption $c_{speaker}$ based on image-text similarity. More finetuning configurations are detailed in Tab. 6. Our data source for the image referential game is *COCO-CAPTION* (Lin et al., 2014) which includes 5 captions for each image. We used the Karpathy-split[6] – training on *COCO-Karpathy-Train* and testing on *COCO-Karpathy-Val*. To build preferential caption pairs {`preferred caption`, `dispreferred caption`} for PO, we used a pre-trained CLIP (Ilharco et al., 2021) to compute sim-

[6] https://cs.stanford.edu/people/karpathy/deepimagesent/coco.zip

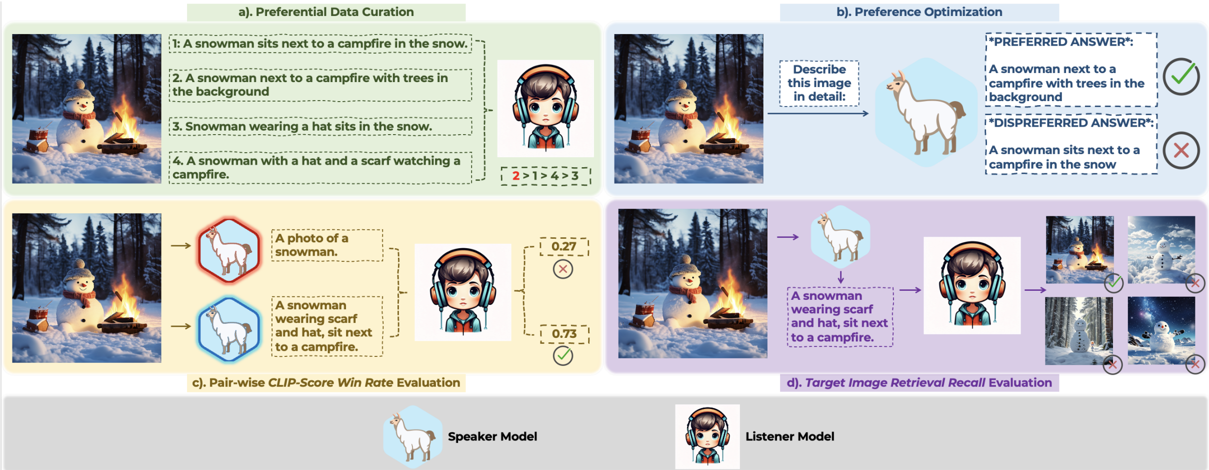

Figure 3: Illustrations of our image referential game experiment with the preferential tuning objective DPO (Rafailov et al., 2024): **a)** Data curation of paired preferential captions; **b)** DPO-finetuning a base speaker VLM; **c)** Evaluating different output captions in terms of *CLIP-Score Win Rate*; **d)** Evaluating caption's *Target Image Retrieval Recall*.

ilarity scores between each image and its 5 associated captions. The caption with the highest image-text similarity was selected as the preferred caption, while a random alternative was chosen as the dispreferred caption. We evaluated the speaker VLM's ToM using two metrics specific to the image referential game:

• *CLIP-Score Win Rate*: This metric compares the captions generated by different models based on their similarity to the target image, using CLIP-Score (Hessel et al., 2021) to determine the winner. The win rate reflects which model generates captions with higher fidelity to the target image.

• *Target Image Retrieval Recall*: This metric measures the recall of the target image from among the distractors, given the speaker's caption. It directly simulates the listener's task of selecting the correct image from a set of distractors.

Fig. 3 illustrates our data curation, preferential tuning process, and evaluation pipeline.

Results. Tab. 2 presents the evaluation results for the base `LLaVA-1.5-7B` speaker, alongside the SFT- and PO-finetuned version. The *CLIP-Score Win Rate* compares captions between each pair among the three models, while *Target Image Retrieval Recall* is calculated at different levels (R@k for $k \in \{1, 5, 10\}$), with k indicating the number of retrieved candidates. The results show:

The PO-finetuned speaker outperforms both the base VLM and the SFT-trained version across all metrics in this multimodal experiment – similar to the textual-domain results (§4.1). The +`PO` model generates captions that achieve the highest CLIP-score similarity with the target image and consistently leads to the highest retrieval success on the listener's part, which directly indicates the best image referential game success.

SFT leads to a slight decline in performance compared to the base pretrained VLM. The +`SFT` speaker wins fewer than 50% of the caption comparisons against the base `LLaVA-1.5-7B`, and its retrieval recall is consistently lower across all k values. This further proves that forcing a single correct answer, as done in SFT, can even impair a model's ToM, which requires flexibility in the face of dynamic social scenarios and the listener's knowledge space.

The consistent performance of PO across both text-based pragmatic QA (§4.1) and image referential game (§4.2) highlights its effectiveness in developing pragmatic abilities within the model's internal representations, regardless of the modality. This in turn supports our notion that learning pragmatics requires comparing more grounded options against less grounded ones, rather than force-memorizing of fixed answers.

4.3 Layer Depth

Human social reasoning and pragmatic prediction with ToM are integral to high-level cognitive processes (Sperber and Wilson, 1986; Bara, 2011). Inspired by this fact, we explore how the depth[7] of trainable network layers in a Transformer-based LLM (Vaswani et al., 2017) relates to its pragmatic

[7] In our terminology, layer 1 (closest to the input) is considered the "deepest" layer, while layer 32 (closest to the output) is considered the most "shallow" layer.

(a) CLIP-Score Win Rate

	LLaVA-1.5-7B	+SFT	+PO
LLaVA-1.5-7B	-	56.6	45.4
+SFT	43.4	-	41.2
+PO	**54.6**	**58.8**	-

(b) Target Image Retrieval Recall

R@1	R@5	R@10
31.0	56.9	68.4
$30.5_{\downarrow 0.5}$	$56.0_{\downarrow 0.9}$	$67.1_{\downarrow 1.3}$
$\mathbf{31.9}_{\uparrow 0.9}$	$\mathbf{58.0}_{\uparrow 1.1}$	$\mathbf{69.4}_{\uparrow 1.0}$

Table 2: Image referential game evaluation results on *COCO-Karpathy-Val* in terms of the *CLIP-Score Win Rate* and *Target Image Retrieval Recall*. We compare three versions of the speaker: the base VLM `LLaVA-1.5-7B` as well as the SFT-tuned (+SFT) and PO-tuned (+PO) `LLaVA` model. The **best** scores are boldfaced.

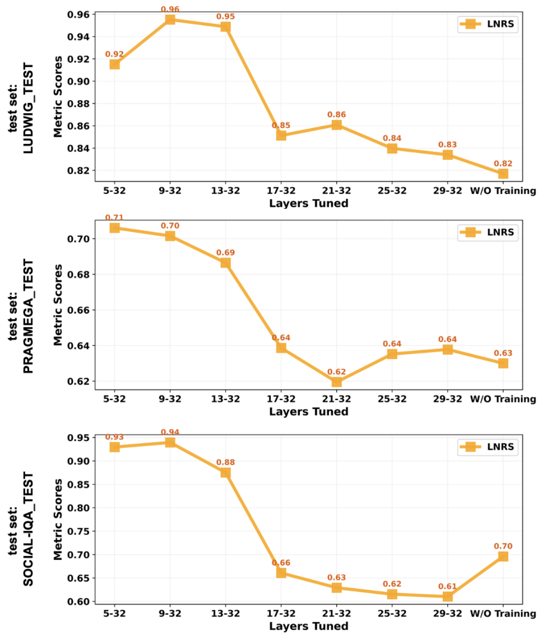

Figure 4: Impact of trainable `LLAMA2-7B` transformer layer depth on PO-tuned pragmatic performance.

reasoning abilities.

Setup. Following the framework in §4.1, we applied DPO to `LLAMA2-7B-Chat` (Touvron et al., 2023) with 32 transformer layers as a demonstrative model, and used *SOCIAL-IQA_Train* as an example training set. We controlled the trainable `layer_id` (starting from 1) combinations, using a 4-layer interval: `(5-32)`, `(9-32)`, `...`, `(29-32)`. Evaluation was performed across three test sets: *SOCIAL-IQA_Test*, *PRAGMEGA_Test*, and *LUDWIG_Test* (Tab. 4), using the open-ended assessment metric $LNRS$ (§2.2).

Results. Fig. 4 reveals a clear overall trend: **as we train progressively shallower layers, the model's performance in pragmatic inference declines**. While preference-tuning deeper layers significantly improves performance compared to the base `LLAMA2-Chat`, training only shallower layers yields limited benefits and can even degrade the model's performance. This underscores the necessity of engaging deeper layers for effective pragmatic learning. Additionally, the LLM's ability to learn pragmatic inference drops sharply starting from approximately the midpoint of the transformer stack, with minimal gains observed after finetuning beyond the `21`st layer. The best results are obtained by training the deep-down `5-` or `9-32` layers. Interestingly, skipping the `5-8`th layers produces a slightly higher $LNRS$ score, though the difference is not significant.

This contrast between the effectiveness of preferential tuning in deeper versus shallower transformer layers suggests a possible correspondence with the pattern of human cognition. Just as complex social-pragmatic reasoning in humans relies on **higher-level** cognitive processes, our results (Fig. 4) demonstrate that **deeper** layers in an LLM significantly invoke pragmatic performance, while training shallower layers offer little improvement.

5 Related Work

Machine Pragmatics. Rooted in linguistic theory (Grice, 1975; Austin, 1962; Searle, 1975; Sperber and Wilson, 1986), the study of pragmatics within machine learning has recently been explored in terms of how LLMs perform in scenarios involving various pragmatic phenomena (Hu et al., 2023; Lipkin et al., 2023; Ruis et al., 2023; Qi et al., 2023; Sravanthi et al., 2024) or subtle social norms (Sap et al., 2023; Shapira et al., 2023). Theory of mind (ToM) (Premack and Woodruff, 1978) has been tested in tasks such as false-belief reasoning (Kosinski, 2023; Ullman, 2023), story comprehension (Jones et al., 2023), and multi-turn interactive contexts (Kim et al., 2023). Additionally, Gandhi et al. (2023) proposed a framework for using LLMs themselves to generate ToM evaluation samples, revealing that `GPT-4` (OpenAI, 2023) is the only model matching human capabilities whereas all other LLMs struggle. To improve ToM inference in LLMs, Moghaddam and Honey (2023) employed

Base Model	Finetuning Dataset	Finetuning Method	MMLU 5-shot	ARC-E 5-shot	ARC-C 25-shot	AGIEval 0-shot	GSM8K 8-shot	OpenBookQA 0-shot
LLAMA2-7B-Chat	-	-	47.4	80.9	53.2	37.0	23.2	43.8
	SOCIQL-IQA	PO	47.5	**83.0**	**58.4**	**37.3**	**23.4**	**46.6**
	SOCIQL-IQA	SFT	**48.1**	81.1	52.6	36.7	20.2	44.6
	PUB	PO	**48.1**	**81.2**	**55.3**	**37.8**	**24.3**	**44.2**
	PUB	SFT	47.2	80.8	51.9	36.7	23.0	42.6
LLAMA2-13B-Chat	-	-	53.6	83.5	59.7	39.0	35.4	44.0
	SOCIQL-IQA	PO	**54.0**	**85.3**	**62.8**	**39.2**	**35.7**	**46.4**
	SOCIQL-IQA	SFT	53.4	84.2	58.8	38.7	33.2	45.4
	PUB	PO	**54.4**	**84.8**	**61.6**	**39.5**	**35.9**	**44.8**
	PUB	SFT	53.9	83.0	58.1	38.5	32.7	44.2
PYTHIA-6.9B-Tulu	-	-	34.0	67.9	39.7	31.9	11.7	38.4
	SOCIQL-IQA	PO	**34.6**	**70.3**	**43.0**	**33.0**	**11.5**	**40.6**
	SOCIQL-IQA	SFT	33.3	67.8	38.9	32.5	10.8	36.8
	PUB	PO	**35.2**	**68.9**	**40.2**	**32.7**	**11.4**	**41.0**
	PUB	SFT	33.9	67.5	39.2	32.2	9.9	36.0

Table 3: Various benchmark performances of the base LLMs along with their versions PO- and SFT-finetuned on pragmatic datasets. The **best** metric scores are boldfaced.

few-shot prompting with chain-of-thought (Wei et al., 2022) and step-by-step reasoning (Kojima et al., 2022), while Sclar et al. (2023) proposed a graph module for tracking each character's mental state. For the image referential game, approaches have been developed to explicitly build a simulated ToM-listener that externally models ToM and guides the speaker's output (Zhu et al., 2021; Liu et al., 2023; Takmaz et al., 2023).

Finetuning Methods of LLMs. Pretrained LLMs undergo finetuning that better aligns these models with human instructions and conversational behaviors. Supervised finetuning (SFT) – also referred to as instruction tuning, follows the language modeling loss on `{human instruction, response}` data that directly train the LLMs to follow human instructions and respond like the given "gold" `response`. Instruction-tuned LLMs, such as `InstructGPT` (Ouyang et al., 2022), outperform pretrained base models like `GPT-3` (Brown et al., 2020) in generating more natural, human-like conversations. Preference optimization (PO) steers LLMs towards outputs that align with human preferences. Reinforcement learning from human feedback (RLHF) (Christiano et al., 2017; Ziegler et al., 2019) uses human feedback in the form of paired data `{preferred response, dispreferred response}` to train a reward model for interpreting human feedback, which then guides the LLM's outputs to align with the human preferences under a reinforcement learning framework. However, RLHF can be complex to implement and prone to unstable training. Recent works such as DPO (Rafailov et al., 2024) and SimPO (Meng et al., 2024) simplify and improve the training process by eliminating the need for a separate reward model or reference model, thereby making preference optimization more efficient.

6 Conclusion

This paper addresses two lines of challenges related to social-pragmatic abilities in LLMs. First, we argue for a shift from the traditional MCQA format to open-ended evaluation that directly measures the soundness of the model's generated responses in social scenarios. Second, we propose to enhance LLMs' intrinsic pragmatic abilities via preference optimization (PO) over supervised finetuning (SFT). Through PO, models learn to capture the subtle nuances between preferred and dispreferred social interactions. Our experiments across multiple pragmatic datasets, coupled with human evaluation, and further examined within a multimodal theory of mind setting through the image referential game, all effectively demonstrate both the advantages of our free-form evaluation protocol and the superiority of PO over SFT in pragmatic scenarios. Additionally, we also reveal the impact of trainable layer depth on the model's pragmatic performance gains, suggesting a potential mirroring with the higher-level cognitive processes involved in human social reasoning.

Limitations

In our open-ended evaluation paradigm, we used GPT-4 (OpenAI, 2023) as the judge to score the models' generated responses. While this approach was effective, it relies on an API that offers limited control over how the judge's evaluations are conducted. Future work should explore more transparent and controllable methods for quantifying the quality of free-form outputs.

The benefits of preference optimization (PO) for improving machine pragmatics are both intuitively motivated by the absence of a single "gold" answer in social interactions and empirically validated by our experiments across modalities. But our models also inherit certain issues associated with PO, such as verbosity (Appx.D, Tab. 8). Addressing how to refine these inherent limitations in PO algorithms remains an open question for future research.

Finally, as shown in our layer-depth studies (§4.3), the social-pragmatic abilities of LLMs are closely tied to deeper representation, which may reflect a similarity to the role of high-level cognitive processes in human pragmatic reasoning. This potential connection between machine learning and human cognition should inspire future research on possibly bridging human cognitive science with language modeling.

Ethics Statement

In this project, all data and pretrained models are publicly available. They are collected and processed in adherence to the respective data, checkpoints, and API usage policy. We acknowledge that our finetuned models may generate unsafe content, and we advise all users of careful verification before deploying this work in real-world applications.

References

John Langshaw Austin. 1962. *How to do things with words*. William James Lectures. Oxford University Press.

Bruno G Bara. 2011. Cognitive pragmatics: The mental processes of communication.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Robyn Carston. 2004. Stephen c. levinson, presumptive meanings: the theory of generalized conversational implicature. cambridge, ma: Mit press, 2000. pp. xxiii+ 480. *Journal of linguistics*, 40(1):181–186.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Rodolfo Corona, Stephan Alaniz, and Zeynep Akata. 2019. Modeling conceptual understanding in image reference games. *ArXiv*, abs/1910.04872.

Yann Dubois. 2024. Length controlled alpacaeval.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *Preprint*, arXiv:2404.04475.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback. *Preprint*, arXiv:2305.14387.

Sammy Floyd. 2022. Pragmega materials.

Balazs Galambosi. 2024. Advanced length-normalized alpacaeval 2.0. `https://github.com/tatsu-lab/alpaca_eval/issues/225`.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. 2023. Understanding social reasoning in language models with language models. *Preprint*, arXiv:2306.15448.

Erving Goffman. 1959. The moral career of the mental patient. *Psychiatry*, 22(2):123–142.

Mitchell S Green. 1998. Direct reference and implicature. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 91(1):61–90.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Geoffrey Hinton. 2014. Coursera lecture slides - neural networks for machine learning lecture 6.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7038–7051.

Laurence R. Horn. 1972. On the semantic properties of logical operators in english' reproduced by the indiana university lin.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *arXiv preprint arXiv:2305.02301*.

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

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. *Preprint*, arXiv:2212.06801.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. Openclip. If you use this software, please cite it as below.

Cameron Robert Jones, Sean Trott, and Ben Bergen. 2023. EPITOME: Experimental protocol inventory for theory of mind evaluation. In *First Workshop on Theory of Mind in Communicating Agents*.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Le Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. Fantom: A benchmark for stress-testing machine theory of mind in interactions. *Preprint*, arXiv:2310.15421.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

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

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

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Benjamin Lipkin, Lionel Wong, Gabriel Grand, and Joshua B Tenenbaum. 2023. Evaluating statistical language models as pragmatic reasoners. *arXiv preprint arXiv:2305.01020*.

Andy Liu, Hao Zhu, Emmy Liu, Yonatan Bisk, and Graham Neubig. 2023. Computational language acquisition with theory of mind. *Preprint*, arXiv:2303.01502.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024a. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Jie Liu, Zhanhui Zhou, Jiaheng Liu, Xingyuan Bu, Chao Yang, Han-Sen Zhong, and Wanli Ouyang. 2024b. Iterative length-regularized direct preference optimization: A case study on improving 7b language models to gpt-4 level. *arXiv preprint arXiv:2406.11817*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Junru Lu, Jiazheng Li, Siyu An, Meng Zhao, Yulan He, Di Yin, and Xing Sun. 2024. Eliminating biased length reliance of direct preference optimization via down-sampled kl divergence. *arXiv preprint arXiv:2406.10957*.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *arXiv preprint arXiv:2405.14734*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*.

Shima Rahimi Moghaddam and Christopher J Honey. 2023. Boosting theory-of-mind performance in large language models via prompting. *arXiv preprint arXiv:2304.11490*.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

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

Peng Qi, Nina Du, Christopher Manning, and Jing Huang. 2023. PragmatiCQA: A dataset for pragmatic question answering in conversations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6175–6191, Toronto, Canada. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Joshua Robinson and David Wingate. 2023. Leveraging large language models for multiple choice question answering. In *International Conference on Learning Representations*.

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. *Preprint*, arXiv:2210.14986.

Maarten Sap, Ronan LeBras, Daniel Fried, and Yejin Choi. 2023. Neural theory-of-mind? on the limits of social intelligence in large lms. *Preprint*, arXiv:2210.13312.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. SocialIQA: Commonsense reasoning about social interactions. In *EMNLP*.

Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. 2023. Minding language models' (lack of) theory of mind: A plug-and-play multi-character belief tracker. *Preprint*, arXiv:2306.00924.

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

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2023. Clever hans or neural theory of mind? stress testing social reasoning in large language models. *Preprint*, arXiv:2305.14763.

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. 2023. Distilling reasoning capabilities into smaller language models. *Preprint*, arXiv:2212.00193.

Dan Sperber and Deirdre Wilson. 1986. *Relevance: Communication and cognition*, volume 142. Harvard University Press Cambridge, MA.

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. *Preprint*, arXiv:2401.07078.

Ece Takmaz, Nicolo' Brandizzi, Mario Giulianelli, Sandro Pezzelle, and Raquel Fernández. 2023. Speaking the language of your listener: Audience-aware adaptation via plug-and-play theory of mind. *Preprint*, arXiv:2305.19933.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, et al. 2023a. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. *arXiv preprint arXiv:2306.05087*.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023b. How far can camels go? exploring the state of instruction tuning on open resources. *Preprint*, arXiv:2306.04751.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*.

Pei Zhou, Aman Madaan, Srividya Pranavi Potharaju, Aditya Gupta, Kevin R. McKee, Ari Holtzman, Jay Pujara, Xiang Ren, Swaroop Mishra, Aida Nematzadeh, Shyam Upadhyay, and Manaal Faruqui. 2023. How far are large language models from agents with theory-of-mind? *Preprint*, arXiv:2310.03051.

Hao Zhu, Graham Neubig, and Yonatan Bisk. 2021. Few-shot language coordination by modeling theory of mind. *Preprint*, arXiv:2107.05697.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A GPT4-Judge Prompt Templates

Here's our prompt template for querying GPT-4 (`gpt-4-1106-preview`) to score the model's free-form answer in relation to the provided "gold" answer (§2.2). To mitigate position bias, we query GPT-4 twice with the reversed order of the model's and the "gold" answer. For the reversed order query, we simply rearrange the following prompt to have the "gold" answer come first.

Template for GPT4-judge

[Scenario]:
`{QUESTION}`

[Model's Answer]:
`{ANSWER_MODEL}`

[Gold Answer for Reference]:
`{ANSWER_GOLD}`

[System]:
We request your evaluation of the AI model's answer in relation to the provided scenario and the gold answer. Assess the responses based on the following criteria:
1. Social Understanding: How well does the model's answer grasp the social dynamics or pragmatic nuances of the scenario?
2. Appropriateness: Is the model's answer appropriate and contextually fitting for the scenario?
3. Insightfulness: Does the answer demonstrate a deep understanding of the underlying intentions, implicature, deceit, irony, sarcasm, humor, metaphor, etc.?
4. Completeness: How comprehensive is the model's response in capturing the essential elements of the scenario?

Please first output a single line containing only two numeric values representing scores for the model's answer and the gold answer respectively, on a scale of 1 to 10, where a higher score indicates better performance. The two score values should be separated by a space. The gold answer is for reference only and should not strictly limit the evaluation.
In the next line, provide a comprehensive explanation of your evaluation, discussing each of the criteria mentioned. This explanation should avoid any potential bias and ensure that the judgment is solely based on the response's merits in the context of the scenario and the gold answer for reference.

B Human Evaluation Instruction

Instructions for Human Evaluators

We request your ranking evaluation of different answers to the provided scenarios and questions. Please assess the answers based on the following criteria:
1. Overall Appropriateness: Is the answer suitable and contextually fitting for the scenario?
2. Social Understanding: How well does the answer grasp the social dynamics or pragmatic nuances of the scenario?
3. Conversational Insightfulness: Does the answer demonstrate a deep understanding of the underlying intentions, implicature, deceit, irony, sarcasm, humor, metaphor, etc.?

Rank the answers based on their qualities. Place the best answer first, the second-best second, and so on.
Do NOT let the length of the answers bias your judgment. A longer answer may better capture the scenario, or it may be unnecessarily verbose.
Disregard minor format variations such as ending with or without a period, extra quotation marks, or differences in upper/lower cases.

Feel free to include any additional comments at the end of the questionnaire.

Any data you submitted remains anonymous and will be used for research purposes only.

C Implementation Details

Tab. 5 provides the detailed finetuning hyperparameters for the pragmatic question answering task discussed in §4.1.

Tab. 6 provides the detailed finetuning hyperparameters for the image referential game discussed in §4.2. Since our focus is on how the VLM generates captions (*i.e.*, how it arranges the wording), we do not finetune the VLM's image-encoder module, allowing it to maintain a stable and robust image embedding space throughout the experiments.

D Qualitative Examples of Model Responses in Pragmatic Question Answering

To provide more fine-grained analyses and better illustrate one of our key motivations – "the human-annotated 'gold' answer might not always be the best response" (§1) – we analyze qualitative examples from the model's generations in the pragmatic QA task discussed in §4.1.

In Tab. 7, we present examples where the responses generated by our models under DPO tuning are judged by GPT-4 as **even better than the reference "gold" answer**. These examples illustrate how our PO-tuned models handle nuanced contextual cues across a variety of social-pragmatic phenomena. In many cases, the model's responses provide more detailed and clearer messages than the "gold" answer. For instance, in *metaphor* comprehension, the preference-tuned models use more descriptive words with better details, facilitating easier communication. Similarly, in scenarios involving *social norms*, the PO-tuned models generate responses that capture richer sentiments beyond the "gold" response (*e.g.*, *sad because of the inability to go out*) or provide more in-depth reasoning (*e.g.*, *trying to change the subject*).

However, we also acknowledge certain limitations with current PO techniques, such as verbosity (Meng et al., 2024; Lu et al., 2024; Liu et al., 2024b), which exactly motivates the length-normalization aspect of our proposed $LNRS$ metric (§2.2).

Tab. 8 shows examples where the model's response is overly verbose. In these cases, the DPO-tuned models produced responses that, while containing the correct intent, were excessively verbose, weakening the intended *humor* (first example) or *ironic* messages (second example). Addressing these non-ideal cases will be a promising avenue for future work.

Data Source	Phenomena	#Train	#Test
SOCIAL-IQA[a]	various social norms	$33,410$	$2,224$
PRAGMEGA[b]	deceits, indirect speech, irony, maxims, metaphor, humor	0	130
LUDWIG[c]	implicature	0	718
PUB[d]	implicature, presupposition, reference, deixis	$18,627$	0

Table 4: Details of the data sources for experimenting with our evaluation and tuning methods. If #Train is 0, it means that we do not use this data source for training – because of the data's scarcity.

[a] `https://allenai.org/data/socialiqa`. We keep the original train/dev/test splitting.

[b] This is an ongoing project at `https://osf.io/6abgk/?view_only=42d448e3d0b14ecf8b87908b3a618672`. We used the data provided by `https://github.com/jennhu/lm-pragmatics` and discarded the binary classification "Coherence" task.

[c] `https://huggingface.co/datasets/UCL-DARK/ludwig`.

[d] `https://huggingface.co/datasets/cfilt/PUB`. We combined the original train/dev as our training split. We also discarded the task instances made easier with hints. The testing questions rely too much on the MCQA selection format, so we choose not to use its test set.

Method	Parameter	Value
SFT, DPO	batch size	64
SFT, DPO	learning rate	$5.0e-07$
SFT, DPO	max gradient norm	10.0
SFT, DPO	optimizer	`RMSprop` (Hinton, 2014)
SFT, DPO	warmup iterations	150
SFT, DPO	training epochs	1
SFT, DPO	max sequence length	512
SFT, DPO	max prompt length	256
SFT, DPO	label smoothing	0
DPO	DPO beta	0.1

Table 5: Pragmatic question answering base LLMs' finetuning hyperparameters.

Method	Parameter	Value
SFT, DPO	LoRA (Hu et al., 2021) r	128
SFT, DPO	LoRA (Hu et al., 2021) alpha	256
SFT, DPO	batch size	16
SFT, DPO	learning rate	$1.0e-07$
SFT, DPO	optimizer	`AdamW` (Loshchilov and Hutter, 2017)
SFT, DPO	learning rate schedule	`Cosine`
SFT, DPO	weight decay	0
SFT, DPO	warmup ratio	0.03
SFT, DPO	training epochs	1
SFT, DPO	max sequence length	2048
DPO	DPO beta	0.1

Table 6: Hyperparameters for finetuning the base speaker VLM `LLaVA` in the image referential game.

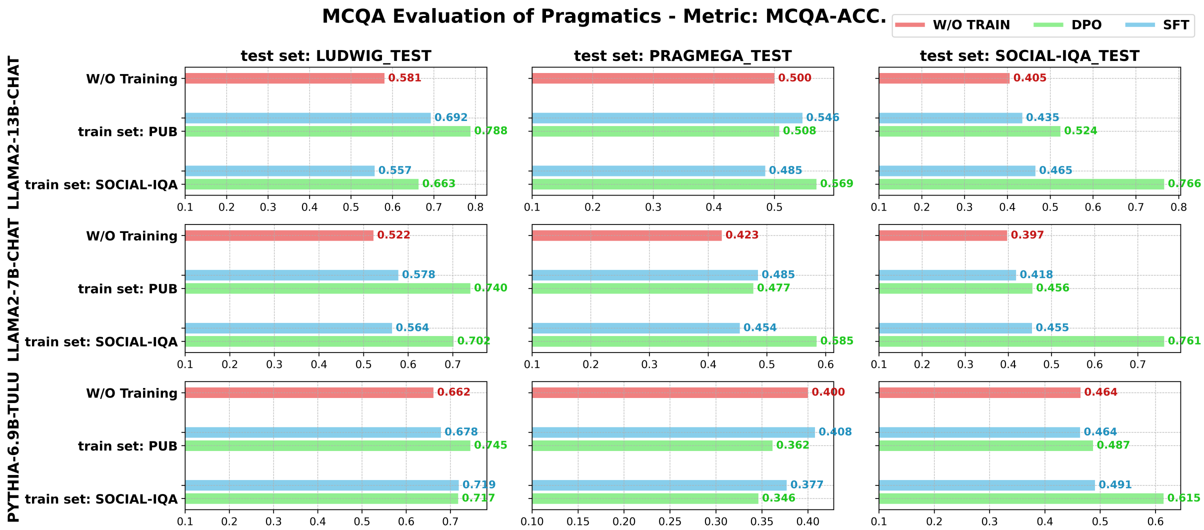

Figure 5: *MCQA-ACC* comparisons across models, data sources, and training paradigms (PO *v.s.* SFT).

Question	"gold" Answer	Model's Answer	Model Version	Judge Scores	Phenomenon
Mike and Melissa are arguing about the state of secondary education. Mike states that standardized test scores are up. Melissa responds sarcastically about other unrelated factors. Why?	Melissa thinks that statistical information is useless in this debate.	Melissa has responded to highlight the limitations and flaws in the standardized test.	`LLAMA2-7B-Chat` after PO	Gold: 6 Model: 6.5	Maxims
Cindy got her paints ready but has a meeting in 10 minutes. Her dad doubts it's the best time for painting. What is he implying?	He does not want Cindy to start painting.	She should prioritize her meeting over painting.	`LLAMA2-13B-Chat` after PO	Gold: 7 Model: 8	Indirect Speech
Janet said the dancers were butterflies. What did she mean?	The dancers flitted around the stage.	Janet meant that the dancers were graceful and light, like butterflies.	`LLAMA2-7B-Chat` after PO	Gold: 7.5 Model: 8	Metaphor
The teacher told Sam: "You have too much on your plate." What did the teacher mean?	Sam has too many things to take care of.	The teacher means that Sam has too many tasks or responsibilities and is overwhelmed.	`LLAMA2-13B-Chat` after PO	Gold: 8 Model: 9	Metaphor
Melissa's friend convinces her to a nightclub said to be "super hot". When cold inside, Melissa says, "The heat is really on in here." What did she want to convey?	It is too cold.	Melissa wanted to convey that the nightclub is freezing cold, not hot.	`LLAMA2-7B-Chat` after PO	Gold: 8.5 Model: 9	Irony
Austin got a really bad sunburn and decided to stay inside. How would Austin feel afterwards?	Burned.	Sad that he couldn't be out in the sun again.	`PYTHIA-6.9B-Tulu` after PO	Gold: 4 Model: 5.5	Social Norms
Riley beat around the bush when asked about crimes. What will she want to do next?	Lie more.	Try to avoid answering questions, probably, or try to change the subject.	`LLAMA2-7B-Chat` after PO	Gold: 5.5 Model: 7	Social Norms

Table 7: Examples where the model's response is judged better than the annotated "gold" answer.

Question	**“gold” Answer**	**Model’s Answer**	**Model Version**	**Judge Scores**	**Phenomenon**
A woman decides to prepare her will. She tells her lawyer she had two final requests. First, she wanted to be cremated, and second she wanted her ashes scattered all over the local mall. "Why the mall?" asked the lawyer. She replied:	"Then I’ll be sure my daughter visits me twice a week."	"Well, you see, I’ve spent so much time there over the years, I figure it’s only fitting that a part of me should be there forever. Plus, I think it would be kind of funny to see my ashes scattered all over the food court and watch the looks on people’s faces as they try to figure out where the smell is coming from!"	`LLAMA2-13B-Chat` after PO	Gold: 9 Model: 4	Humour
The Jones’ had hired a new gardener. After a week the lawn was looking particularly unkempt. John remarked to his wife, "He does such a good job!" What did John want to convey?	He is unhappy with the new gardener.	This is a classic lateral thinking puzzle. The answer is "John was being sarcastic." The statement "He does such a good job!" is unlikely to be sincere if the lawn is looking unkempt after only a week. It is more likely that John is joking or being ironic about the gardener’s work.	`LLAMA2-7B-Chat` after PO	Gold: 8.5 Model: 7.5	Irony

Table 8: Examples of verbosity in model responses.