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

Anonymous ACL submission

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 - as our results show - correlates better with human judgement. Further, we explore the enhancement of pragmatic abilities in LLMs, proposing the use of preference optimization (PO) over supervised finetuning (SFT) since there's no "gold" answer in responding to a social situation. Our results indicate that preferential tuning significantly outperforms and proves more robust than SFT across pragmatic phenomena, and offers a near-free launch to enhance models' pragmatic ability without compromising generic abilities. Lastly, we delve into LLMs' internal space and demonstrate that the substantial boost of the model's pragmatic reasoning capabilities is linked to deeper layer representation, mirroring human's high-level thinking. Our experiments span multiple pragmatic and social reasoning data sources, covering diverse phenomena, as well as a image referential game requiring multimodal theory of mind (ToM). With our refined paradigms for evaluating and enhancing pragmatic inference, this paper offers key insights for developing more socially aware language models. [1]

1 Introduction

Social-pragmatic inference is a key aspect of human communication, requiring the ability to understand and respond to the implied meanings, intentions, and emotional states behind literal utterances (Horn, 1972; Grice, 1975; Green, 1998; Carston, 2004) along with shared social conventions (Goffman, 1959). This type of inference covers a range of phenomena including implicatures, irony, humor, and metaphor, as well as high-level cognitive thinking such as theory of mind (ToM) (Premack and Woodruff, 1978), which are all essential for interpreting non-literal language and context-dependent messages. For instance, a friend's statement, "*It's chilly in here*" that might be a polite request to close a window rather than a mere observation about temperature demonstrates pragmatic inference.

The importance of social-pragmatic intelligence in human communication underscores the need for large language models (LLMs) to possess similar capabilities to interact more naturally with users. Current approaches to addressing pragmatic abilities in LLMs face two lines of limitations:

1) On the evaluation front, typical evaluation methods measure 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). However, even if a model chooses the correct option label, it might still fail to respond by itself in a pragmatic way to a social scenario. For example (see Fig. 1), a model might correctly choose an appropriate answer in an MCQA setup without truly grasping the social intricacies of *changing the subject*. Furthermore, real-life social interactions rarely have a single "gold" answer, therefore judging by the accuracy of selecting the provided **fixed** response undermines the assessment of a model's true pragmatic capability in flexible generations.

2) On the pragmatic-ability-improvement front, 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 increase LLMs' pragmatic test results, little effort has been made to explicitly invoke the model's internal social pragmatic intelligence, so that it learns to generate social-pragmatically appropriate answers en-

[1] Our code will be made publicly available.

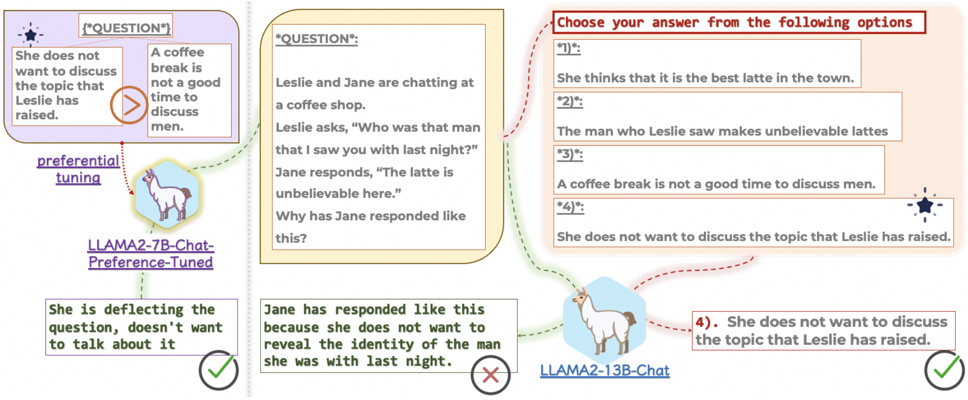

Figure 1: An example of LLMs' outputs when queried about a social-pragmatic scenario taken from Hu et al. (2023). On the right-hand side, a `LLAMA2-`**`13B`**`-Chat` (Touvron et al., 2023) model correctly selects the gold response ID when given the question and all the candidate answers in a multiple choice question answering (MCQA) format, whereas it **fails** to grasp the true underlying pragmatic meaning of the scenario when asked to generate its own response to the question. The left-hand side is the open-ended response of a smaller `LLAMA2-`**`7B`**`-Chat` model preference-tuned on the contrast of the gold answer to other less pragmatic options. Its response is equally good and pragmatically sound as the provided "gold" answer.

tirely on its own.

In this paper, we propose paradigm shifts on both fronts.

1) For evaluation, we argue for an open-ended evaluation protocol that directly assesses a model's own response to a social scenario. We introduce ***length-normalized relative score*** ($LNRS$) that directly rates the model's free-form response in reference to the provided "gold" answer with GPT4 [2] (OpenAI, 2023) as judge and further debiased for reducing length gameability (Dubois et al., 2024; Galambosi, 2024). Supported by human evaluation, our open-ended metric $LNRS$ is better correlated with human preferences than the MCQA accuracy.

2) For enhancing LLM's pragmatic inference, we regard the not-selected answer options in existing MCQA-formatted datasets not as incorrect, but as a less pragmatically grounded answer in comparison to the "gold" response. We use **preference optimization (PO)** objectives such as DPO (Rafailov et al., 2024) to finetune an LLM so that it grasps the subtle nuances of pragmatic preference. We empirically demonstrate that preferential tuning yields a much better performance boost on an LLM than typical **supervised finetuning (SFT)** across pragmatic phenomena, and induces less impact on other abilities inherited from the base LLM. When transferring to the multimodal setting of image referential game (Corona et al., 2019; Zhu et al., 2021; Liu et al., 2023) that requires the captioning model to have a theory of mind (ToM) (Premack and Woodruff, 1978), the PO objective also results in a more capable ToM-aware image captioner, which further illustrates the superiority of PO over SFT for imparting models with pragmatic abilities.

To develop a deeper understanding of how the internal components of a transformer (Vaswani et al., 2017)-based LLM are most responsible for invoking social-pragmatic abilities, we further experimented with controlling different trainable layers. The results suggest that pragmatic understanding is clearly associated with **deeper-down** transformer layers, which hints at a potential similarity with how human pragmatic inference also relies on **higher-level** cognitive processes.

Overall, the main contributions of this paper are:

• Proposing open-ended assessment of models' free-form responses instead of MCQA classification for evaluating social-pragmatic understanding, which correlates better with human judgement;

• Proposing preference optimization (PO) over supervised finetuning (SFT) for the enhancement of LLMs' pragmatic capacity without harming other inherited model abilities, which is effectively proved by experiments across pragmatic data sources and multimodal theory of mind (ToM);

• Providing empirical analyses of how only training deeper layers of an LLM can invoke pragmatic performance gains, which potentially mirrors human's high-level cognitive thinking.

[2]GPT4 is the sole model available performing with high robustness and human-likeness in most social pragmatic studies (Gandhi et al., 2023; Sap et al., 2023; Zhou et al., 2023; Ruis et al., 2023; Kosinski, 2023)

2 Evaluating Pragmatic Abilities

2.1 Existing Evaluation: *MCQA Accuracy*

Existing works mostly assess a language model's pragmatic intelligence in the form of multiple (or even binary) choice question answering (MCQA) tasks, where for a given social scenario, a set of answer options is provided, from which the model being evaluated needs only choose one as its response (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 correctly selecting the annotated "gold" answer is used as the indicator of a model's pragmatic abilities ($MCQA$-Acc). In recent studies, the way to elicit a model's choice among the set of provided answer options can be divided into two methods:

• *Metalinguistic*[3] *Probing*: The model is directly prompted the instruction to choose from a set of answers associated with symbolic indicators (alphabetic letters like `A|B|C|D` (Le et al., 2019; Sravanthi et al., 2024; Robinson and Wingate, 2023) or index digits like `1|2|3|4` (Hu et al., 2023)). The model then generates the symbolic indicator of the option it chooses.

• *Probability Probing*: The model is prompted the scenario and question text (context, $\mathbf{x}$). We then calculate the model's likelihood of generating each one of the answer options $\mathbf{y}_i$ conditioned on the input context. The option with the highest probability is deemed the answer the model chooses in the sense that it is most likely to be generated by the model. For the probability calculation, there can again be variations in the normalization technique (Brown et al., 2020; Robinson and Wingate, 2023; Holtzman et al., 2021) that lead to different formulations:

• Without normalization: $P\left(\mathbf{y}_i \mid \mathbf{x}\right)$;

• With length normalization over j tokens in $\mathbf{y}_i$: $\frac{\sum_{j=1}^{\ell_i} P\left(y_i^j | \mathbf{x}, \mathbf{y}^{1\cdots j-1}\right)}{\ell_i}$;

• Normalization by unconditional answer probability[4]: $\frac{P(\mathbf{y}_i|\mathbf{x})}{P(\mathbf{y}_i|\mathbf{x}_{\text{uncond}})}$

The problems with these accuracy-based MCQA tests are multi-fold:

1) This task format deviates far from real-life social interactions, where there's no fixed answer to select. Even the provided "gold" answer in these benchmarks may not be the best response to the given scenario. For instance, the preference-tuned model's response in Fig. 1 (left-hand part) is equally sound in its social and pragmatic sense.

2) As also pointed out in Robinson and Wingate (2023), different models have different levels of proficiency binding an option to its symbol (*multiple choice symbol binding, MCSB*), which is an ability potentially conflated with true pragmatic intelligence, especially with the *metalinguistic probing* approach.

3) Being able to classify the correct answer option does not necessarily mean that a model really understands the social scenario and can respond in a socially and pragmatically grounded way on its own (see right-hand part of Fig. 1), which is the actual ability desired for more natural human-LLM interaction in real-life applications.

Therefore, we argue for a paradigm shift in evaluating machine pragmatics towards **open-ended** assessment of the model's autonomous response, while still keeping the use of the annotated "gold" answer as reference.

2.2 Open-Ended Evaluation: *Length-Normalized Relative Score*

We introduce ***Length-Normalized Relative Score*** ($LNRS$) to quantitatively measure how well the model's own response is when compared to the provided "gold" answer. Instead of providing the model with options for choice, we directly obtain the model's own response to the pragmatic question describing a social scenario. Then we ask the most advanced GPT4 (OpenAI, 2023) to score the model's own response in reference to the provided "gold" answer.

GPT4-Judge. We use GPT4 as judge, for it is the sole LLM available that has been most consistently shown to perform robustly at a human-matching level across various social-pragmatic studies (Gandhi et al., 2023; Sap et al., 2023; Zhou et al., 2023; Ruis et al., 2023; Kosinski, 2023). Also, GPT4 has been commonly applied in numerous settings, *e.g.,* in typical instruction-following evaluation (Chiang et al., 2023; Li et al., 2023; Dubois et al., 2024, 2023; Wang et al., 2023a), and even as a "teacher" to guide other LLMs in reasoning tasks (Shridhar et al., 2023; Hsieh et al., 2023). In line with prior work using GPT4-judge, we also randomly permute the order of the model's answer and the provided "gold" answer to allevi-

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

[4] *domain conditional point-wise mutual information* in Holtzman et al. (2021)'s term.

ate potential position bias. Specifically, we query GPT4 twice with reversed order of the model's and the "gold" answer. Our prompt template for querying GPT4 (`gpt-4-1106-preview`) to score the model's free-form answer in reference to the provided gold answer is given in Appx.A.

After parsing each of GPT4's responses as a pair of scores, we then compare the average scores of the model's answer to the average scores of the gold answer. For all the questions from the test set T, we first calculate the ***relative score*** (RS) of the model's response a_{model} in reference to the "gold" answer a_{gold} as $RS = \frac{\sum_{q \in T} \mathrm{JS}(a_{model})}{\sum_{q \in \mathrm{T}} \mathrm{JS}(a_{gold})}$, in which JS denotes the judge's score. This intuitively measures the degree to which the model's answers are as good as (or even better than) the "gold" responses throughout the test set, which directly indicates if the model's understanding – as manifested in its own free-form answer – aligns with nuanced social norms and pragmatic rules.

Length Normalization. Inspired by recent advancements in LLM evaluations such as AlpacaEval-2.0 (Dubois et al., 2024; Galambosi, 2024), we also carefully reduce the influence of length bias that may affect GPT4's judgment (termed *length gameability* in Dubois et al. (2024)) in our pragmatic evaluation. We adopted the *logistic length normalization* technique (Galambosi, 2024; Dubois, 2024) [5] to our open-ended pragmatic evaluation. Specifically, *length-normalized relative score* ($LNRS$) normalizes the RS by a temperature-weighted sigmoid function of the differences between the length of model's and the "gold" response:

$$\begin{aligned} LNRS = {} & \frac{\sum_{q \in T} \mathrm{JS}(a_{model})}{\sum_{q \in T} \mathrm{JS}(a_{gold})} \\ & \cdot \sigma\Big(\frac{1}{\tau \cdot T}\Big(\sum_{q \in T} \mathrm{Len}(a_{gold}) - \sum_{q \in T} \mathrm{Len}(a_{model})\Big)\Big) \end{aligned} \tag{1}$$

in which τ betokens a temperature hyperparameter, and JS and Len denotes the judge score and token length respectively.

In §4.1, we empirically demonstrate the superiority of the open-ended $LNRS$ over current $MCQA$-Acc, the former of which correlates better with real user preferences in **human evaluation**.

[5] The *length control* method used in AlpacaEval-2.0 (Dubois et al., 2024) cannot be transferred to our evaluation setting without prior win-rate data. So we turned to *length normalization* that has only a close performance gap to *length control*.

3 Improving Pragmatic Abilities

On top of establishing an open-ended evaluation paradigm that matches real-life scenarios more closely, we also set out to investigate how the social-pragmatic inference of LLMs can be intrinsically improved. Different from previous works (§5) that are more inclined to apply external modules for better cognitive abilities (Sclar et al., 2023; Takmaz et al., 2023) or few-shot prompt engineering (Moghaddam and Honey, 2023; Ruis et al., 2023), we are concerned about aligning the model's **intrinsic representation** towards a more social-pragmatically grounded distribution.

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

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

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

Preference Optimization (PO). In social contexts, however, there is no definitive <u>right</u> answer. For example, in the MCQA-formatted data sources like in Fig. 1, we do not consider *e.g.*, `option 3)` a <u>wrong</u> answer. It is just not as socially and pragmatically appropriate in common sense as `option 4)` in the described context. Such nuanced understanding – weighing the possible responses in terms of pragmatic soundness and social appropriateness – is exactly what we want to develop in the model.

We thus turn to the preference optimization (PO) paradigm with the simplified *direct preference optimization (DPO)* objective (Rafailov et al., 2024), which does not solely rely on maximizing the likelihood of a given answer but rather focuses on optimizing the model parameters θ to reflect a preference for more desired answers over less desired ones. Among different answer options to q, we construct pairwise triples (q, a_{gold}, a_{other}), where given a question q, a_{gold} is the provided "gold" answer and thus the preferred response over any other answer option a_{other}. For a data source $\mathcal{D}$, the PO objective can be formulated as:

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

where σ is the sigmoid function, β is a hyperparameter.

4 Experiments

4.1 Pragmatic Question Answering

Setup. We experimented with 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). They cover a wide range of pragmatic phenomena including implicature, metaphor, irony, and various social norms. Tab. 6 summarizes the dataset details. We used three versions of base LLM across different pretraining data and model sizes: `PYTHIA-6.9B-Tulu` (Wang et al., 2023b), `LLAMA2-7B-Chat` and `LLAMA2-13B-Chat` (Touvron et al., 2023).[6] Our detailed training configurations can be found in Tab. 4.

Human Evaluation. To further support our advocate for open-ended assessment of pragmatic abilities, we recruited 12 voluntary human participants from top educational institutions to judge the quality of different responses. Given a social-pragmatic context and question, the human evaluator is presented with randomly ordered four types of responses (the dataset-annotated "gold" option, the base LLM's responses, the PO-tuned and the SFT-tuned models' generations). Then we ask the evaluator to rank the responses in terms of their pragmatic understanding and fitness to the context scenario. Appx.B gives the detailed instructions we employed for this user study. The ranking of the four responses is transformed into scores, with the first place receiving 4 points and the last place receiving 1 point. In total, we randomly sampled 192 samples coupled with the four responses, and randomly assigned 16 data points to each evaluator for assessment.

Results. Fig. 2, Fig. 3, and Tab. 1 shows the performance of LLMs finetuned with different paradigms (PO *v.s.* SFT) – evaluated respectively in the open-ended framework (§2.2), the MCQA format[7] (§2.1), and user study (see above). From the results, we observe the following patterns:

1) Across almost all configurations of base models, training data, test sets as well as evaluation paradigms (MCQA/open-ended/human-eval), the PO-tuned LLMs significantly outperforms the SFT-trained counterparts, boosting the pragmatic inference over the base model by a substantial margin. There are very few exceptions such as the negligibly lower *LUDWIG_Test* $LNRS$ score of the `PYTHIA-6.9B-Tulu` DPO-tuned on *PUB* in contrast to SFT. Additionally, under the MCQA setup, the DPO-tuned `LLAMA2-13B-Chat` performs worse than SFT on *PRAGMEGA_Test*, which however strongly contrasts human users' judgement (Tab. 1) that ranks the PO-version of `LLAMA2-13B-Chat` as having the best response quality.

2) The open-ended evaluation paradigm correlates better with human judgement than the MCQA results. Tab. 1 reveals the clear human preference for responses generated by PO-tuned models, which claims the **best** place (even better than the annotated "gold" answer) for both `LLAMA2` models and second only to the "gold" answer for `PYTHIA`. In contrast, the SFT-ed models is even lower rated than its base LLMs, showing that SFT can even hurt pragmatic performance. These human evaluation results resonate with the $LNRS$ comparisons Fig. 2, where we observe similar patterns of PO's superiority and SFT's potential harm on model pragmatics.

3) The PO objective enables a more robust transfer to "out-of-domain" pragmatic phenomena. We intentionally designed our test sets to consist of both "in-domain" (*i.e.*, same data source and similar phenomena with train sets, *e.g., SOCIAL-IQA_Train/_Test*) and "out-of-domain" (*i.e.*, different data source and phenomena from the train sets) data. We sometimes observe even larger performance gains of PO on different data sources. For instance, when tested on *SOCIAL-IQA_Test*, `LLAMA2-13B-Chat` DPO-finetuned on *PUB* (impli-

[6] We only adopted already instruction-tuned chat models as baseline in order to start with a decent instruction-following ability for our models, especially because the social-pragmatic data is relatively scarce and might not be sufficient for general-purpose alignment tuning.

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

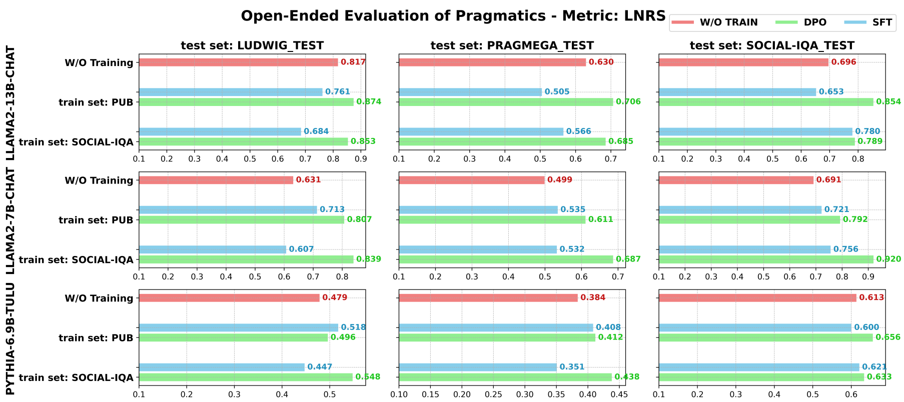

Figure 2: $LNRS$ comparisons across models, data sources and training paradigms (PO *v.s.* SFT).

catures, presuppositions, etc.) even outperforms the version DPO-ed on the same social norm dataset.

4) The PO objective exerts little influence on other abilities inherited from the base LLMs. In Tab. 3, across almost all benchmarks including professional examination (Hendrycks et al., 2020; Zhong et al., 2023; Clark et al., 2018), math (Cobbe et al., 2021), reading comprehension (Mihaylov et al., 2018), the models DPO-ed on our pragmatic data always outperforms their SFT counterparts, frequently by a large margin. This strongly shows that despite being finetuned on pragmatic datasets, the preference-optimized version offers a **near-free launch** of pragmatic abilities, while even **improving** the various other abilities learnt by the base models at the same time. The SFT-tuned alternatives, however, performs far worse in terms of retaining these inherited abilities.

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

Table 1: Average human evaluation scores elicited from our user study ranking different responses (§4.1). Best and second results are highlighted.

4.2 Image Referential Game with ToM

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

Task Formulation. The image referential game encompasses two interlocutors – a speaker and a listener: Given an image i_{target}, the speaker generates a descriptive caption $c_{speaker}$, based on which the listener tries to choose the target image i_{target} out of a set of images containing both the one described by the speaker i_{target} and several distraction images $i_{distractor} \in I_{distractor}$. ToM is vividly present in this task, because the speaker has to be able to take the listener's understanding into account when arranging the wording of its caption, so that the listener makes the correct choice of the target image. In line with §4.1, we improve the speaker VLM's intrinsic ToM via the same SFT and PO objectives as in §3 and §4.1, with additional visual conditions represented as the image encodings.

Setup. We implement the base VLM-speaker as `LLaVA-1.5-7B` (Liu et al., 2024). For the listener, we use the discriminative `OpenCLIP-ViT-B/32` (Ilharco et al., 2021) to match the target image i_{target} with the given caption from the speaker $c_{speaker}$ based on image-text similarity. Detailed finetuning configurations and are provided in Tab. 5. Our image referential game data source is the widely-adopted *COCO-CAPTION* (Lin et al., 2014) containing 5 captions for each image. We follow the Karpathy-split[8], using *COCO-Karpathy-Train* for training and *COCO-Karpathy-Val* as the test set. To build the preferential data pairs `{preferred caption, dispreferred caption}` for PO, we use a pretrained CLIP (Ilharco et al., 2021) to calculate the similarity scores between an image and its corresponding five captions, among which the caption with the highest text-image similarity is

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

taken as the preferred option. We then randomly sample another caption as the dispreferred one. We assess the speaker-VLM's ToM with two metrics according to the image referential game setting:

• *CLIP-Score Win Rate*: We compare different models' captions in terms of their similarity to the target image implemented as CLIP-Score (Hessel et al., 2021), and decide on the winner. This win rate metric indicates if a model's output is superior in terms of its absolute fidelity to the target image.

• *Target Image Retrieval Recall*: We calculate the recall rate of the target image among all distractions, given the caption generated by the speaker. This metric directly simulates the listener's choice among a set of distraction images.

Fig. 4 demonstrates our data curation, preferential tuning, and evaluation pipeline.

Results. Tab. 2 presents the evaluation results of the base `LLaVA-1.5-7B` speaker, together with the SFT and PO finetuned versions in terms of both *CLIP-Score Win Rate* and *Target Image Retrieval Recall*. For the win rate, we compare each pair among the three models. For the recall metric, we set R@k with $k \in \{1, 5, 10\}$ indicating the number of retrieved candidates. The results show:

1) Similar to the pure-text results (§4.1), the PO-finetuned speaker also outperforms both the base VLM and the SFT-trained counterpart across metrics here in our multimodal experiment. The `+PO` version of `LLaVA` wins both the base and `+SFT` speaker in the absolute caption-image CLIP-score similarity and it leads to the highest retrieval success on the listener's part, directly indicating the best image referential game success.

2) We also find that the SFT training could even result in a slight decrease in performance compared to the base pretrained VLM under both evaluation protocols. The `+SFT` speaker wins the base `LLaVA-1.5-7B` less than 50% of times and its resulting retrieval recall is worse than the base speaker across candidate numbers k. This further proves how forcing just one correct answer may even hurt a model's ToM that requires flexibility in the face of dynamic social scenarios as well as the listener's knowledge space.

4.3 Layer Depth

Human social reasoning and pragmatic predictions with ToM are integral to high-level cognitive processes (Sperber and Wilson, 1986; Bara, 2011). Inspired by this fact, in this section, we explore the relationship between the network layers and the pragmatic reasoning abilities in a Transformer (Vaswani et al., 2017) -based LLM.

Setup. Following §4.1, we conducted DPO on *SOCIAL-IQA_Train* as an example train set and took the `LLAMA2-7B-Chat` (Touvron et al., 2023) with 32 transformer layers as a demonstrative model. We controlled trainable layer_id [9] combinations with 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. 6) using the open-ended assessment metric $LNRS$ (§2.2).

Results. From Fig. 5, we observe an overall clear decrease in performance as the depth of trained LLM layers becomes shallower. While DPO-tuning deeper layers leads to a marked improvement in pragmatic inference compared to the non-finetuned base model `LLAMA2-Chat`, training shallower layers produces limited effects and can even degrade performance. This underscores the necessity of engaging deeper network layers for effective pragmatic learning. Approximately from the middle of all transformer stacks, the LLM's ability to learn pragmatic inference degrades severely. After about the `21`th layer, the finetuning yields few performance gains, as demonstrated by the almost flat lines of metric scores' change. The best performance is achieved by training the deep-down `5-` or `9-32` layers. It also seems that skipping the training of the `5-8`th layer even leads to a slightly better $LNRS$ score, which however does not account for a significant difference.

This contrast between the effectiveness of preferential tuning in deeper versus shallower transformer layers suggests a possible correspondence with the pattern observed in human cognitive processes. Just as high-level cognitive abilities in humans such as social-pragmatic inference rely on deep cognitive strategies, our experimental results (Fig. 5) similarly demonstrate that deeper layers in an LLM significantly enhance pragmatic performance, while shallower layers have a negligible impact.

5 Related Work

Machine Pragmatics. With theoretical underpinning in linguistics (Grice, 1975; Austin, 1962; Searle, 1975; Sperber and Wilson, 1986), pragmat-

[9] Layer_id starts from 1.

(a) CLIP-Score Win Rate

	LLaVA-1.5-7B	(+ SFT)	(+ DPO)
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$_{\downarrow 1.3}$
31.9$_{\uparrow 0.9}$	**58.0**$_{\uparrow 1.1}$	**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-trained (`+SFT`) and PO-trained (`+PO`) `LLaVA` model.

ics within the machine learning communities 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). The theory of mind (ToM) (Premack and Woodruff, 1978) abilities have been tested in false-belief tasks (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 an LLM itself to expand on ToM evaluation samples, whose results showed `GPT4` (OpenAI, 2023) as the sole LLM matching human capabilities whereas all other LLMs struggle. To improve LLM's ToM inference, Moghaddam and Honey (2023) employed 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 specific challenge of image referential game, approaches that explicitly build a simulated ToM-listener have been proposed to externally model ToM that 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 typically serves to better align these models with human requests (*i.e.*, instructions) and human-like conversation. **Supervised finetuning (SFT)** – sometimes also referred to as instruction tuning – follows the language modeling loss on `{human instruction, response}` data to directly trains the LLMs to follow human instructions and respond like the given "gold" `response`. Instruction-tuned LLMs typically become "chatbots" in that they follow user inquiries and carry on with dialogues in a more natural way. For instance, the instruction-tuned `InstructGPT` (Ouyang et al., 2022) outperforms `GPT3` (Brown et al., 2020) in terms of conversation with users. **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 to interpret human feedback, which then guides the LLM's outputs to align with the human preferences under a Reinforcement Learning framework. In the face of RLHF's limitations in its implementation complexity and unstable training process, recent works (*e.g.*, DPO (Rafailov et al., 2024), SimPO (Meng et al., 2024), and *etc.*) greatly improve the training efficiency of RHLF by alleviating the requirements for a reward model or reference model.

6 Conclusion

This paper addresses two lines of challenges with regard to the social-pragmatic abilities in LLMs. We first advocate for shifting from MCQA to open-ended assessment that directly measures the soundness of the model's own answer to a social scenario. Then we propose to enhance the LLM's intrinsic pragmatic abilities via preference optimization (PO) over supervised finetuning (SFT), where a model learns to capture the subtle nuances between preferred and dispreferred social interactions. Our experiments on multiple pragmatic data sources coupled with human evaluation, and the image referential game, effectively demonstrate both the advantages of our free-form evaluation protocol and the superiority of PO over SFT in pragmatic scenarios. We also reveal the impact of trainable layer depth on the model's pragmatic performance gains, which potentially mirrors human's high-level social thinking.

Limitations

Under our proposed paradigm of open-ended evaluation, this paper employed GPT4 (OpenAI, 2023) as judge to score the models' generation, which, though effective, is based on API that allows limited control over the judge's assessment. Future work should look into more transparent and controllable methodologies for quantifying the quality of free-form outputs.

The benefits of preference optimization (PO) for improving machine pragmatics is both intuitively motivated by our insight of the non-existence of a "gold" answer and empirically proved by our experiments across modalities. Nevertheless, the exact numeric mechanism underlying the pronounced impact of PO on social-pragmatic inference remains to be explored.

Furthermore, as demonstrated by our layer-control studies (§4.3), LLMs' social-pragmatic abilities are linked to deeper representation, which possibly resonates with how human pragmatic reasoning is also governed by high-level cognitive processes. This potential synergy between machines' deep understanding and humans' high-level cognition should inspire future work on 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 do recognize that our finetuned models may generate unsafe contents, and we advise all users of careful verification before putting our 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. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

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

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, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

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 GPT4 (`gpt-4-1106-preview`) to score the model's free-form answer in relation to the provided "gold" answer (§2.2). We query GPT4 twice with reversed order of the model's and the "gold" answer:

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. 4 is our detailed finetuning hyperparameters for pragmatic question answering task (§4.1).:

Tab. 5 is our detailed finetuning hyperparameters for image referential game (§4.2). Note that since we are concerned with how the VLM "speaks" (*i.e.*, how it arranges the caption wording), we do not finetune the VLM's image-encoder module, which then provides a robust and stable embedding space of images throughout our experiments.Since we are concerned with how the VLM "speaks" (*i.e.*, how it arranges the caption wording), we do not finetune the VLM's image-encoder module, which then provides a robust and stable embedding space of images throughout our image referential game experiments.

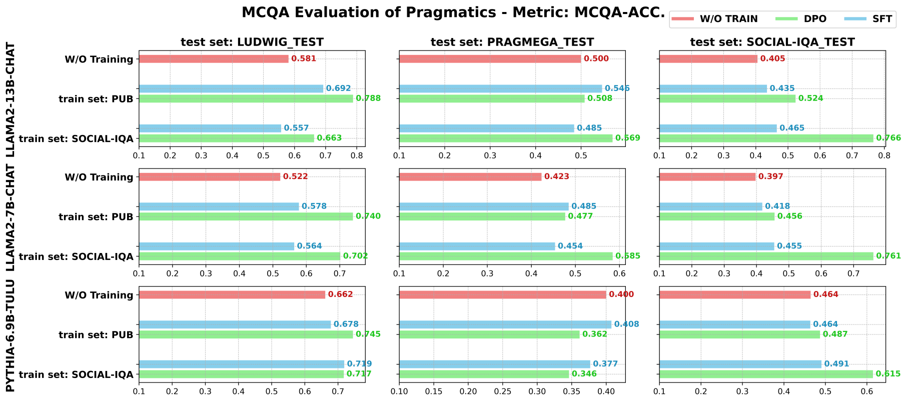

Figure 3: *MCQA-ACC.* comparisons across models, data sources and training paradigms (PO *v.s.* SFT).

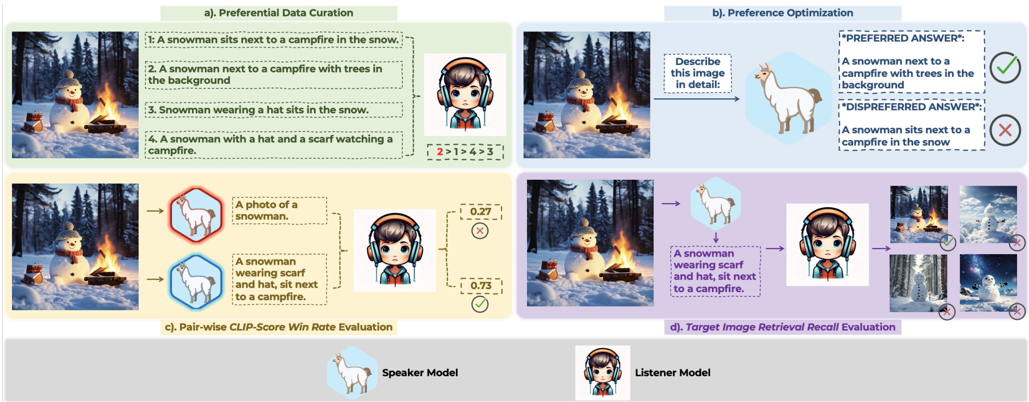

Figure 4: Illustrations of our image referential game experiment with the preferential tuning objective DPO (Rafailov et al., 2024): **a)** Data curation of paired preferential captions; **b)** DPO-finetuning a base speaker VLM; **c)** Evaluating different output captions in terms of *CLIP-Score Win Rate*; **d)** Evaluating caption's *Target Image Retrieval Recall*.

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 marked.

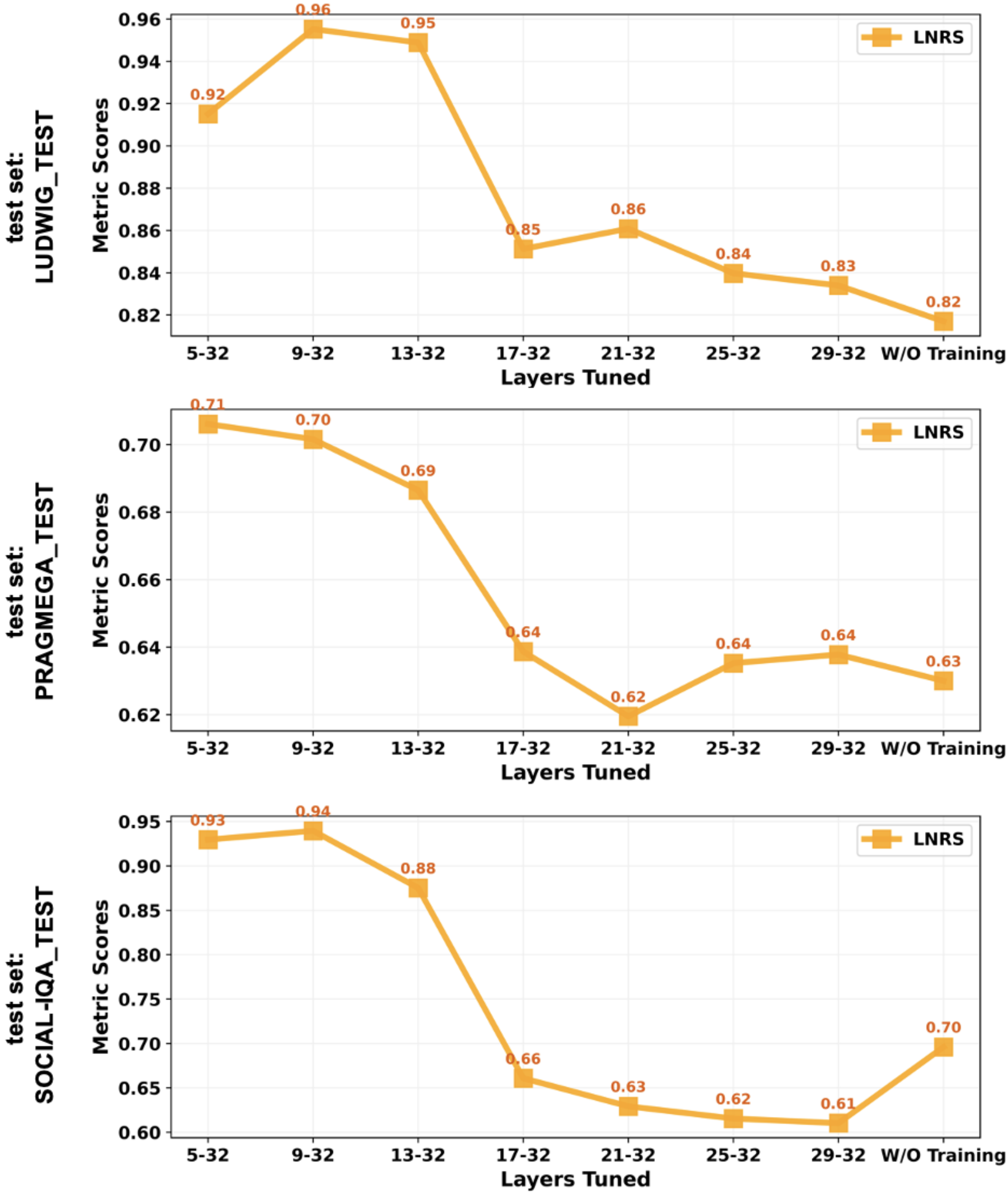

Figure 5: Effects of trainable LLAMA2-7B transformer layer depth on the outcome PO-tuned pragmatic performance.

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 4: 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 5: Hyperparameters for finetuning the base speaker VLM LLaVA in the image referential game.

Data Source	Phenomena	#Train	#Test
SocialIQA[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 6: 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.