Grounding Gaps in Language Model Generations

Omar Shaikh* **Kristina Gligorić*** **Ashna Khetan** **Matthias Gerstgrasser**
Diyi Yang **Dan Jurafsky**
Stanford University
{oshaikh, gligoric, ashnak, mgerst, diyiy, jurafsky}@stanford.edu

Abstract

Effective conversation requires common ground: a shared understanding between the participants. Common ground, however, does not emerge spontaneously in conversation. Speakers and listeners work together to both identify and construct a shared basis while avoiding misunderstanding. To accomplish grounding, humans rely on a range of dialogue acts, like clarification *(What do you mean?)* and acknowledgment *(I understand.)*. However, it is unclear whether large language models (LLMs) generate text that reflects human grounding. To this end, we curate a set of *grounding acts* and propose corresponding metrics that quantify attempted grounding. We study whether LLM generations contain grounding acts, simulating turn-taking from several dialogue datasets and comparing results to humans. We find that—compared to humans—LLMs generate language with less conversational grounding, instead generating text that appears to simply presume common ground. To understand the roots of the identified *grounding gap*, we examine the role of instruction tuning and preference optimization, finding that training on contemporary preference data leads to a reduction in generated grounding acts. Altogether, we highlight the need for more research investigating conversational grounding in human-AI interaction.

1 Introduction

In dialogue, **common ground** refers to the mutual knowledge, beliefs, and assumptions shared by participants in a conversation. This shared understanding is essential for effective communication, as it underpins the ability of individuals to interpret, predict, and respond to each other's statements and actions accurately (Clark, 1996). Through each conversational turn, individuals collaborate to build common ground, preventing potential misunderstandings (Clark and Schaefer, 1989). Humans therefore rely on dialogue acts like *clarifying* meaning or posing information-seeking *followup* questions. When individuals fail to ground, they often proactively repair misunderstandings.

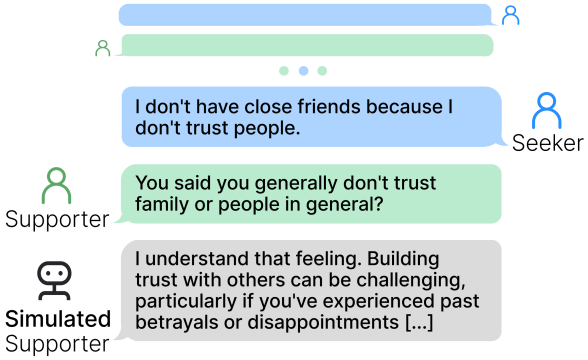


Figure 1: Mental health supporters carefully employ clarification questions (one of the three grounding acts) with a seeker, taking multiple turns to ground. In contrast, simulated supporters—with the same conversational context—generate presumptive answers.

Failing to construct common ground in human-human conversation can be at best misleading and at worst harmful. Consider mental-health support: if there's any indication of risk to a client (e.g., suicidal ideation or intentions of self-harm), a healthcare professional will ask clarifying questions to assess risk. Failure to do so has harmful consequences (unnecessary hospitalization) which can traumatize a client and place an unjustified financial burden (Strumwasser et al., 1991). In domains like education, ineffective grounding might result in misunderstanding a student, resulting in irrelevant feedback (Graesser et al., 1995).

From an NLP perspective, establishing and leveraging common ground is a complex challenge: dialogue agents must recognize and utilize implicit aspects of human communication. Recent chat-based

*Equal contribution.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 6279–6296
June 16-21, 2024 ©2024 Association for Computational Linguistics

large language models (LLMs), however, are designed explicitly for following instructions. While humans carefully construct common ground, LLMs are trained to directly act on commands specified by end-users (Ouyang et al., 2022). We hypothesize that the instruction-following paradigm—a combination of supervised fine-tuning (SFT) and preference optimization (PO through RLHF, for example)—might result in discrepancies between how humans actually ground in dialogue, and how LLMs generate text similar to human grounding. We call this discrepancy the **grounding gap**.

Despite the grounding gap, LLMs interact regularly with humans across various applications. For a subset of interactions, however, LLMs *should* generate grounding language before completing a user's task, instead of executing literal instructions or disregarding a user's underlying goals. This is particularly crucial in LLM-powered training systems, where LLMs simulate practice scenarios and allow individuals to rehearse and refine domain-specific skills (Shaikh et al., 2023a). LLM-based training already facilitates interaction in domains like **education** (Kasneci et al., 2023; Demszky et al., 2021; Wang and Demszky, 2023), **conflict resolution** (Shaikh et al., 2023a; Argyle et al., 2023), and **emotional support** (Carlbring et al., 2023; Hsu et al., 2023). In these settings, effective dialog agents must coordinate to build common ground when interacting with people.

Given the importance of generating language for conversational grounding, we ask: *Do current LLMs generate dialogue acts that reflect grounding patterns between humans? If not, what aspect of LLM training exacerbates the grounding gap?*

We address these questions by measuring LLM generations through linguistically validated **grounding acts**. For example, acts that *clarify* or *acknowledge* a prior utterance offer a strong signal for measuring shared understanding (Clark and Schaefer, 1989). Building on prior work in dialogue and conversational analysis, we curate a collection of dialogue acts used to construct common ground (§2). Then, we select datasets & domains to study human-LM grounding. We focus on settings where human-human grounding is critical, and where LLMs have been applied: namely, emotional support, persuasion, and teaching (§3).

After curating a set of grounding acts, we build prompted few-shot classifiers to detect them (§4). We then use LLMs to simulate turn-taking in our human-human dialogue datasets and compare agreement between human and GPT-generated grounding strategies (§5).

Because we use the exact same conversational context as in human conversations, we can quantify the **grounding gap**: off-the-shelf LLM generations are, on average, **77.5%** less likely to contain grounding acts than humans (§6). Even in situations where LLM generations do contain a grounding act, they differ from human generations—we observe poor human-LM agreement across a range of models.

To isolate potential causes of the grounding gap, we explore a range of possible interventions, from ablating training iterations on instruction following data (SFT and PO) to designing a simple prompting mitigation (§7). We find that SFT does not improve conversational grounding, and PO erodes it. Across our experiments, we generally observe significant disagreement between grounding acts in human utterances and LLM generations.

In summary, we contribute (1) a set of linguistically informed grounding acts, tailored towards understanding grounding in contemporary LLMs. Using grounding acts, we (2) conduct a controlled LLM simulation of conversational grounding and a characterization of *grounding agreement*: a comparison between grounding acts in human utterances and LLM generations. Finally, we (3) investigate reasons for the grounding gap, exploring different models, ablating SFT and PO, and testing prompting interventions.[1]

2 Grounding Acts

Grounding acts are a set of dialogue acts used to characterize conversational grounding. While grounding failures are qualitatively well-observed in current pretrained models (Benotti and Blackburn, 2021b; Fried et al., 2022; Chandu et al., 2021), measuring *where or how frequently* these failures occur in dialogue remains unexplored. Here, we outline concrete acts to measure these failures, discuss our curation process for selecting grounding acts, and describe each act in detail.

Grounding as an Expert To benchmark conversational grounding in LLMs, we focus on dialogue acts from the perspective of an *expert listener.* Users frequently come to LLMs with a direct but potentially underspecified task (Tamkin et al., 2022;

[1]Code and prompts for evaluation can be found at https://github.com/oshaikh13/grounding-gaps

Grounding Act	Example
Clarification *TSCC*	`student`: So I am the leader of my class. It is like the president or mayor. `teacher`: Oh, like a head girl? ↪ `GPT`: That's great to hear! Being the leader of your class is a big responsibility [...]
Follow-up *ESConv*	`seeker`: My friends, who I also work with, are upset with me because I am leaving for a new position. I am sad that they are upset. `support`: Ah, that can be awkward. Are you happy about the new position? ↪ `GPT`: I can understand why that would make you feel down... Remember, it's okay to [...].
Acknowledgment *Persuasion*	`persuadee`: I [donate] as much as I can, which is admittedly not as much as I would like to. `persuader`: I know what you mean! Sometimes it is hard to find the extra time or money to help those that need it. ↪ `GPT`: That's wonderful to hear! [...] Would you be interested in supporting Save the [...]?

Table 1: **Examples of grounding acts from our datasets,** where the act is employed by the expert from each dataset (`teacher`, `supporter`, or `persuader`). Additionally, we include `Chat GPT-3.5`, where the model does not employ the same grounding act.

Zheng et al., 2023). In these settings, the user has privileged knowledge of their task; therefore, the onus of building common ground before providing a response lies initially with a model. More concretely, the perceived expert LLM should generate text that verifies grounding before completing a task. We therefore curate grounding acts from the perspective of an expert listener.

Conversational Grounding Ontologies In curating grounding acts, we draw on prior dialogue research. Several ontologies propose discrete dialogue acts for measuring conversational grounding. For example, Clark and Schaefer (1989) propose a hierarchy of methods to achieve common ground, including the use of explicit discourse acts. They discuss relationships between acts like *acknowledgment* (e.g. *I understand*), and evidence of grounding in conversation. Traum and Hinkelman (1992) outline a range of concrete grounding acts, from simply *continuing* a conversation to *repairing* a misunderstanding or *requesting* repair. Purver (2004) further contributes a theory of repair through *clarification* requests—dialogue acts used to verify if contributions should be added to the common ground.

We curate a small but general subset from the large pool of proposed acts, selecting acts that are relevant to interaction with LLMs, or their current applications (e.g. teaching or emotional support). For example, Motivational Interviewing for emotional support emphasizes asking *followup* questions and signaling *acknowledgment* for empathy (Miller and Rollnick, 2012). Similarly, a range of pedagogical theories incentivize asking careful *clarification* and *followup* questions (Wiske, 1998). And effective conflict resolution requires a careful construction of common ground (Deutsch, 1973). We therefore select the following **three** grounding acts that are especially relevant to these current applications of LLMs but also generalize across a range of domains where grounding is critical.

`Clarification` requests occur when a speaker seeks clarification on an utterance instead of initiating repair. Clarification is used primarily to avoid a misunderstanding, and concerns information presented in prior utterances $u_{1..i}$. In other words, clarifications serve to "clear up" a potential future misunderstanding (e.g. *did you mean...?* or *are you referring to?*), avoiding repair from a listener (Purver, 2004; Ginzburg and Cooper, 2001; Purver et al., 2003b,a; Healey et al., 2011, 2003; Madureira and Schlangen, 2023a; Kuhn et al., 2022; Stoyanchev et al., 2013; Rahmani et al., 2023).

`Acknowledgement` *explicitly* signals understanding (e.g. "ok," "I understand," "I see," "Yes, and", etc.). Unlike clarification/repair, acknowledgment indicates that a speaker is ready for the next relevant turn in a conversation. We only consider utterances whose sole purpose is acknowledgment (i.e. they *exclusively* contain ack. markers.) (Schegloff, 1982; Sacks et al., 1978; Schiffrin, 1987; Clark and Schaefer, 1989; Cho and May, 2020)

`Followup` questions ask for elaboration on a prior utterance u. Followups implicitly signal understanding of u. Unlike clarifications—which are concerned wholly with misunderstandings—followups signal understanding by seeking addi-

tional information on a prior utterance u. Concretely, follow-ups indicate understanding by attempting to ground on related information. While clarification is about understanding the existing interaction more thoroughly, a follow-up is concerned with continuing the current interaction[2] (Davis, 1982; Graesser et al., 1995; Traum and Hinkelman, 1992; Bunt et al., 2017).

3 Data

We choose to analyze conversations in three domains where grounding is critical and where LLMs are already used for social skill training—education, emotional support, and persuasion. Our datasets are task-oriented in nature, have multiple turns, and consist of two participants. Since our grounding acts are curated around an expert listener, each dataset also has one expert participant. To identify grounding gaps, we can simulate utterances from the expert using an LLM. We briefly outline our datasets (in English).

For *emotional support*, we use **Emotional Support Conversations (ESConv)**, a corpus of one-to-one online conversations between a help-seeker and a help-provider, collected via crowdsourcing (Liu et al., 2021). To analyze grounding acts in *education*, we use the **Teacher Student Chatroom Corpora (TSCC)**, a collection of written conversations captured during one-to-one lessons between teachers and learners of English (Caines et al., 2020, 2022). Finally, for *persuasion*, we use **Persuasion for Good**, a dataset consisting of one-to-one online conversations between a persuader and a persuadee, where the persuader solicits donations from payouts on a crowd working website (Wang et al., 2019). Due to resource constraints, we sample 100 conversations from each dataset and truncate them to the median length (TSCC = 92, ESConv = 22 Persuasion = 20). Details on our selected datasets, sampling, and truncation are in Appendix C.

4 Classifying Grounding Acts

To analyze disparities in grounding acts between human and LLM generations, we must first classify grounding acts in our datasets. In this section, we outline the construction of test/validation sets, and our prompting setup to classify grounding acts.

[2]One nifty way to distinguish follow-ups and clarification questions is the "O.K." test, introduced by Benotti and Blackburn (2021a). In short, if prefixing a potential clarification (e.g. "Do you mean X?") with an acknowledgment (e.g. "O.K.") sounds awkward, then it's likely a clarification.

4.1 Dataset Splits and Prompting

Method We turn to classifying the grounding acts in our datasets. First, we withheld and annotated a validation (10%) and test (10%) set of conversations from each dataset in §3. Messages were annotated for a single **primary** act—if an utterance contained multiple grounding acts, we selected the most relevant one. Two authors participated in the annotation process. After annotating, the authors highlighted disagreements, discussed, and broke ties, yielding a final Cohen Kappa agreement of $\kappa = 0.72$. Following annotation, the first author prompt-engineered a multi-class classification prompt on the validation set, using both zero-shot and few-shot prompting. We used the latest `GPT` model during the time of writing (`GPT-4`) as our classification model, with temperature = 0 but default parameters otherwise.[3] Queries were run between July-November 2023.

Results `GPT-4` can identify grounding acts in conversations with reasonably high accuracy (avg. Macro F-1 across datasets = **0.89**, Appdx. Table 2). In the 0-shot setting, however, we find that `GPT-4` frequently misclassifies follow-ups and clarification questions (avg. F-1 clarification = **0.40**; follow-up = **0.70**). Few-shot prompting substantially increases model performance (clarification F-1 = **0.85**; follow-up F-1 = **0.91**).

5 Analyzing Grounding Acts in LLMs

Given our three grounding acts, a reasonably performing `GPT-4` classifier for labeling them, and a set of target datasets & metrics, we can now measure disparities between humans and LM simulations. In this section, we outline our controlled simulation process (§5.1) and metrics to measure conversational grounding (§5.2).

5.1 Simulation Method

We start with a conversation $C_{1..N}$ consisting of N ordered role utterance pairs (r_t, u_t). Each of our selected datasets has two unique roles, expert and listener. Since LLMs are generally leveraged for the role of an expert listener, we focus on simulating the expert. Given this setup, our simulation process generates controlled counterfactual messages g_t for each u_t.

Concretely, to simulate an utterance at timestep t, we extract all messages until t: $C_{1...t-1}$.

[3]We use standard parameters provided in OpenAI's API (max_tokens = 256).

Then, we input this context into a selected LM ($\texttt{LM}(C_{1...t-1})$), along with a high-level instruction (e.g. *Roleplay a therapist*). The LM then generates the next message, offering a counterfactual to the human ground truth. Using this process, each ground truth utterance u_t from our selected datasets has a controlled, LM-generated counterpart g_t, conditioned on the *same* conversational history. Figure 1 summarizes this process.

After generating LM counterfactuals g_t, we can similarly compute the rate of grounding acts across our generated messages. While we use `GPT-4` to classify grounding acts, simulating conversation for all our datasets is prohibitively expensive. In this section, we use `GPT-3.5` with default parameters for conversation simulation (later, in §7, we experiment with a wider range of models).

5.2 Metrics to Measure Grounding Acts

How should we measure the use of our selected grounding acts? We propose two metrics: base rate and Cohen's κ. The base rate provides a measure of grounding frequency; specifically, how often a human utterance or LM generation contains a specific grounding act. In contrast, Cohen's κ measures agreement in grounding acts between humans and LMs. Just because an LM generation contains an act does not mean it occurs in the same place as in human dialogue. Our metrics apply to any dialog dataset D consisting of conversations with utterances-role pairs $\{(r_1, u_1), ..., (r_n, u_n)\}$.

Base Rate We first compute the overall frequency of grounding acts subset G as $P(u_i \in G)$. The base rate provides a reference point for the overall presence of specific conversational grounding acts.

Cohen Kappa κ While the base rate provides a measure of frequency, it does *not* capture the discrepancy between grounding acts used between a candidate and reference conversation. For instance, an LM-generated conversation M may use *more* grounding strategies, but only in locations that rarely match human H use. Consider a simulated dialogue agent that always generates a grounding act. While the language this agent generates doesn't have the problem of presuming common ground, an agent that always generates grounding acts can be understandably irritating (Horvitz, 1999).

To this end, we use **Cohen κ** (Cohen, 1960), a measure of inter-rater agreement bounded between -1 and 1. Cohen κ has several useful properties: it is well-validated across social scientific studies—agreement can be compared to pre-existing work to determine strength. Furthermore, κ adjusts for random chance: values of $\kappa < 0$ indicate that agreement is worse than chance. In measuring grounding acts, we treat M and H as individual raters.

Act	`ChatGPT 3.5`	Human	Cohen κ
	Emotional Support Conv		
Follow	10.78 ± 2.1	27.87 ± 4.4	12.47 ± 6.4
Ack.	1.05 ± 0.8	12.9 ± 3.7	3.14 ± 4.9
Clar.	0.0 ± 0.0	3.05 ± 1.2	0.0 ± 0.0
	Teacher Student Chatroom		
Follow	11.56 ± 1.9	12.04 ± 2.1	16.75 ± 4.6
Ack.	5.68 ± 1.4	16.59 ± 2.4	18.25 ± 5.4
Clar.	0.57 ± 0.3	3.77 ± 0.9	0.36 ± 2.5
	Persuasion for Good		
Follow	1.66 ± 0.9	8.18 ± 2.4	2.94 ± 7.6
Ack.	1.8 ± 1.0	6.11 ± 1.9	25.73 ± 16.7
Clar.	0.0 ± 0.0	0.28 ± 0.4	0.0 ± 0.0

Table 2: **Grounding acts and associated metrics across our datasets**. $\pm$ represents 95% confidence intervals (bootstrapped). Humans use grounding acts more than LLMs; furthermore, LLMs show poor agreement (Cohen κ) with humans.

6 Gaps in Generating Grounding Acts

Having introduced the controlled simulation process, we now report metrics on grounding acts (§6.1) and qualitatively analyze errors (§6.2).

6.1 Simulation Results

First, we find significant discrepancies between humans and LLMs when using `GPT-3.5` for conversation simulation (Table 2). Across all datasets, LLMs generations contain fewer grounding acts, like followups (avg. **64.3%** decrease) and ack. (**83.4%**) acts. Clarifications never occur when using `ChatGPT-3.5` for ESConv and Persuasion. While `ChatGPT-3.5` does initiate some clarification on TSCC, we observe a **84.8%** decrease.

Beyond rate discrepancies, we observe low agreement between `ChatGPT-3.5` and humans—LLM generations rarely contain grounding acts in same position as human utterances. Of the 3 grounding acts $\times$ 3 dataset pairs, only 3 / 9 have a Cohen κ agreement significantly greater than **zero**, with κ averaging **10.73** for followup, **11.13** for acknowledgment, **0.23** for clarification. To con-

Generator \ Act	+ Followup		+ Acknowledgement		+ Clarification	
	Base Rate	Cohen κ	Base Rate	Cohen κ	Base Rate	Cohen κ
Human	27.87 ± 4.4	36.41	12.89 ± 3.7	30.12	3.05 ± 1.2	48.45[6]
3.5-instruct-turbo	29.35 ± 3.3	22.16 ± 7.7	1.49 ± 1.1	5.5 ± 6.0	0.12 ± 0.2	-0.22 ± 0.4
3.5-turbo (ChatGPT 3.5)	10.76 ± 2.1	12.47 ± 6.4	1.05 ± 0.8	3.14 ± 4.9	0.0 ± 0.0	0.0 ± 0.0
4	12.26 ± 2.6	11.04 ± 7.4	1.17 ± 0.8	11.18 ± 7.7	0.35 ± 0.4	-0.61 ± 0.6
mistral-sft	20.03 ± 3.1	18.48 ± 7.7	12.08 ± 3.4	26.92 ± 9.6	0.47 ± 0.5	12.39 ± 16.1
mistral-dpo	15.99 ± 2.7	15.25 ± 8.0	19.11 ± 4.1	19.33 ± 7.7	0.93 ± 0.8	9.79 ± 12.7
3.5-turbo + mitigation	42.32 ± 4.7	26.41 ± 6.6	2.54 ± 1.3	5.02 ± 5.9	1.16 ± 0.8	-1.65 ± 0.9

Table 3: **ESConv grounding acts and metrics across model variants.** We find that more recent OpenAI (3.5-turbo, 4) models use *significantly fewer* grounding acts than humans; and that agreement (Cohen κ) with humans is poor to fair across all evaluated models.

firm that human-LM agreement is low, we run a human-human study with ESConv (Appendix B), and observe significantly higher κ across grounding cts (avg. $\kappa \approx 37.5$).

6.2 Error Analysis

We performed qualitative analyses to develop an in-depth understanding of how LM generations fail to incorporate grounding acts. First, we performed inductive coding to produce a set of frequent errors. An annotator (one of the authors) read a random sample of 160 instances where a human uses a grounding act, while the simulated supporter does not. Relevant turns were examined together with the previous turn for context (details in Appdx. D).

Simulated supporters fail to use **(1) acknowledgment** to show empathy (**43.94%**) or ask **(2) followup** questions for more information (**26.52%**) / to continue a conversation (**12.88%**). Furthermore, simulated supporters do not ask **(3) clarification** questions to verify understanding (**9.09%**), or resolve a specific ambiguity (**7.58%**). Table 4 in the Appendix contains more detail on each error type and qualitative examples. Next, we investigate potential causes of the *grounding gap* and use discovered error types to design an informed prompting mitigation.

7 Why do Grounding Gaps Emerge?

Here, we explore potential mechanisms for how grounding gaps emerge. We focus on analyzing ESConv, a target domain where disagreement in grounding acts is especially consequential.

First, we examine grounding acts across several OpenAI GPT variants and observe a larger grounding gap in *newer* models (§7.1). To understand the roots of this trend, we evaluate open-source models. We hypothesize that current supervised-finetuning (SFT) and preference optimization (PO) datasets drive human-LM disagreement. To test this, we rigorously isolate the effects of SFT and PO training on grounding agreement (§7.2).

7.1 LLM Variants

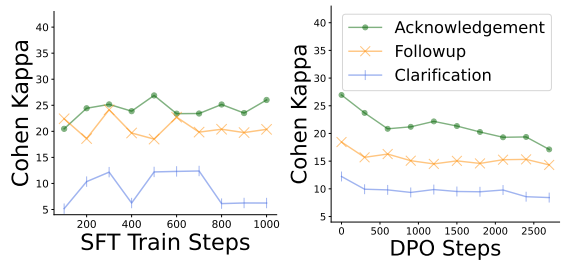
Method To further investigate the grounding gap, we test a wider range of GPT models, rerunning our simulation process for ESConv on gpt-3.5-instruct (a replacement for legacy OpenAI models, trained similarly to the text-davinci-00X series) and gpt-4. We examine grounding acts across these models.

Results Generations from the OpenAI model variants have lower grounding acts base rates and poor agreement with humans ($\kappa < 0.2$). A surprising exception to this is 3.5-instruct-turbo for followups, where instruct shows fair agreement with humans *and* uses grounding acts at a similar rate (29.5 instruct vs. 27.9 Human). While instruct is trained using an "older" procedure, it produces a better κ / base-rate tradeoff compared to most models.

7.2 SFT & Preference Optimization

OpenAI models, however, are closed source: to isolate training procedures that impact grounding agreement, we independently evaluate the role of current SFT and PO datasets.

Method We investigate if the standard SFT + PO training setup improves use of grounding acts; and if the amount of SFT + PO matters. At a high level, we replicate the training procedure for Zephyr, an open-source instruction following LM (Tunstall



(a) (b)

Figure 2: **The role of SFT and preference optimization in grounding agreement.** (a) We observe **no correlation** between **SFT** training steps and Cohen κ agreement on grounding acts, with a Pearson R correlation test yielding insignificant results: $p > 0.1$ (b) We observe **negative correlation** between **DPO** train steps and Cohen κ agreement on grounding acts, with Pearson R averaging $R = -0.79$, and $p < 0.05$ for all acts.

et al., 2023). First, we **SFT** Mistral 7B (Jiang et al., 2023) for three epochs on a filtered version of the UltraChat dataset (Ding et al., 2023; Tunstall et al., 2023).[4] During the SFT process, we save a total of 10 evenly-spaced checkpoints across a training run. Each checkpoint is used to re-simulate conversations and measure the use of grounding acts. Next, we use preferences from the synthetic UltraFeedback (Cui et al., 2023) dataset for the **PO** stage. We run direct policy optimization (DPO) (Rafailov et al., 2023) for an additional three epochs, starting with the highest κ checkpoint on grounding acts from the prior SFT stage. Like with SFT, we save 10 evenly spaced checkpoints and rerun our simulations using each checkpoint.

Beyond synthetic datasets, we additionally evaluate the Archangel models (Ethayarajh et al., 2024), released to study the effect of various instruction tuning and preference optimization (both DPO and PPO (Schulman et al., 2017)) procedures. From the Archaengel suite, we evaluated the final Llama 7B models trained on a mixture of real human feedback datasets: OpenAssistant (Köpf et al., 2024), Stanford Human Preferences (Ethayarajh et al., 2022) and Anthropic HH-RLHF (Bai et al., 2022).

Results Across open-source Mistral experiments, we find no evidence that SFT impacts grounding agreement: Pearson correlation across grounding acts vs. SFT checkpoints is ≈ 0, with $p > 0.1$. Still, while Cohen's κ is poor throughout training ($\kappa < 0.3$), we observe that SFT-only `mistral` has the highest κ across acknowledgment and clarification for all evaluated models—even when including closed-source OpenAI variants.

On the other hand, increased DPO training on Mistral 7B **degrades** agreement across all grounding acts: followup ($R = -0.70, p < 0.05$), acknowledgment ($R = -0.89, p < 0.05$), and clarification ($R = -0.78, p < 0.05$). Similar to how PO induces longer responses (Singhal et al., 2023), we observe bias towards **assuming grounding** instead of employing grounding acts. Altogether, instruction following SFT *does not* improve grounding agreement, and added PO erodes it.

We observe similar degradations with PPO / DPO on the Archangel suite of models (Table 3 in Appendix). After PPO, the base-rate falls by an average of 7.0% across grounding acts, and κ decreases by 39.0%. Similarly, DPO results in an average base-rate decrease of 12.1%, and an average κ decrease of 42.5%. Regardless of the base model or preference optimization algorithm, we observe decreases in generating grounding acts.

We further examined our SFT and preference datasets to identify a potential source of the grounding gap. As a heuristic, we simply searched for questions in assistant responses, and found that assistant responses containing questions are overall relatively rare—11.83% of samples in UltraChat and 18.35% of samples in UltraFeedback have an utterance where the assistant's answer contains any question at all (followup questions and clarification questions are a subset of these). Second, we found that the UltraFeedback dataset explicitly signals that asking questions is dispreferred: questions are significantly **less frequent** in preferred (13.77%) compared to dispreferred (18.35%) examples, $\chi^2 = 484.08$, $p < 0.00001$.

7.3 Prompting Mitigations

Lastly, we explore a potential prompt-based intervention. We design a prompt around our qualitative error analysis and re-evaluate grounding acts.

Method We add a mitigation prompt (full text in Appendix E) to `Chat GPT-3.5 Turbo`, instructing it to avoid errors from our analysis (§6.2); specifically, to **(1)** ask clarification questions when necessary, **(2)** use follow-up questions to continue a conversation, and **(3)** use acknowledgment to show empathy. Our prompting approach is similar to related work on preference elicitation (Li et al., 2023), where prompting a model to first clarify a task improves human ratings for task performance.

[4] For additional training details, see Appendix F.

Results Our prompt mitigation on `ChatGPT 3.5` initially looks promising, with base rates substantially higher across the board: followup nearly quadruples ($10.76 \rightarrow 42.32$), acknowledgment doubles ($1.05 \rightarrow 2.54$), and clarification increases from $0 \rightarrow 1.16$. **However**, we note that interventions result in overeager grounding: simulated supporters overuse grounding acts for minimally improved agreement (e.g. Cohen's κ). For example, while mitigation results in 50% *more followups than a human supporter*, agreement is still poor. Furthermore, for clarification, we find that κ **decreases** after mitigation ($0 \rightarrow -1.65$), with mitigated κ *significantly worse than random chance.* Altogether, we find that while a prompting intervention significantly increases base rates of grounding acts, it yields minimal increases (and potentially decreases) across Cohen κ agreement with humans, indicating that the grounding gap is a fundamental problem, difficult to address by prompting alone.

8 Related Work

Background In operationalizing the concept of common ground (Clark, 1996; Clark and Schaefer, 1989), we build on prior work in linguistics, cognitive psychology, and communication. This includes literature on conversational structure (Jefferson, 1972) and on subdialogues (Litman and Allen, 1987; Litman, 1985).

Conversational grounding in NLP To benchmark grounding abilities, existing works have operationalized tasks relevant to conversational grounding, including question answering (Testoni et al., 2020), producing human-like acknowledgments (Paranjape and Manning, 2021), addressing ambiguities (Paek and Horvitz, 1999), providing conversational feedback (Pilán et al., 2023; Eshghi et al., 2015), addressing repair (Balaraman et al., 2023), asking follow-up (Li et al., 2023) and clarification questions (Purver, 2004). Correctly leveraging grounding strategies is particularly consequential in tasks that require coordination in order to achieve a goal (Bara et al., 2021; Mohanty et al., 2023; Fried et al., 2022; Li and Boyer, 2015), play games (Madureira and Schlangen, 2023b; Shaikh et al., 2023b), plan ahead (Chu-Carroll and Carberry, 1998; Lochbaum, 1998), retrieve data (Lu et al., 2023), or improvise (Cho and May, 2020). In such tasks, the use of grounding acts has been shown to increase success and conversation quality (Zhou et al., 2022). Furthermore, the ability to establish a common ground is a key component in efforts to design believable conversational agents (Park et al., 2022, 2023; Aher et al., 2022; Argyle et al., 2022) and facilitate human-AI collaboration in dialogue (Lin et al., 2023) Our work synthesizes existing literature by formalizing a framework to study grounding in human-AI dialogue.

LLMs and conversational grounding Despite the ubiquity and importance of conversational grounding, previous work has identified fundamental limitations in the ways dialogue agents powered by large language models establish common ground, noting that current systems usually guess what the user intended, instead of leveraging grounding acts.[5] Related to this limitation, various undesirable conversational patterns have been identified, including over-informative question answering (Tsvilodub et al., 2023), refusal to answer ambiguous questions (Abercrombie et al., 2023; Min et al., 2020; Gao et al., 2021), miscalibration issues (Nori et al., 2023; Zhou et al., 2023), and overconfidence (Mielke et al., 2022). Similarly, LLM sycophancy (Perez et al., 2022)—mirroring the views of a user—may be related to presumptive grounding. Large language models' generations have thus been criticized as not being grounded in any communicative intent, any model of the world, or any model of the reader's state of mind (Bender et al., 2021). In our work, we carefully examine the role of contemporary instruction following and preference optimization, analyzing their effect on conversational grounding.

9 Discussion

> `(CAN YOU ELABORATE ON THAT)`
> — **ELIZA Rule (Weizenbaum, 1966)**

Across evaluated models, we observed reduced rates of grounding acts and poor grounding agreement with humans (§6). We also isolated sources of reduced grounding act use (§7). Here, we reflect on findings and outline avenues for future work.

On the risks of *not* generating grounding acts. Most of our evaluated LLMs generate grounding acts at a significantly lower rate than humans. Instead of initiating a grounding act, instruction following LLMs simply "provide the answer." In low-stakes situations like informal conversation or chit-chat, assuming grounding may be acceptable.

[5] `https://openai.com/blog/chatgpt`

However, LLM simulations are used extensively across a range of critical tasks, like social skill training, where appropriate grounding is necessary. We find substantial disparities in these domains: specifically, teaching, persuasion, and therapy. Furthermore, we suspect that these disparities extend far beyond our evaluated domains, e.g. cross-cultural interaction, medical or legal advice, customer support, and beyond.

On contemporary SFT and PO. We find that current preference datasets explicitly signal that asking questions is dispreferred (§7.2). Non-expert humans have different preferences and opinions (Casper et al., 2023), but, in single-step interaction, might agree on salient characteristics of answers including accuracy, relevance and clarity. This results in narrowly scoped preference datasets, where responses that are immediately relevant to a prompt are preferred. However, grounding by asking questions is integral in contexts such as tutoring and emotional support; failing to ground can place a burden on support-seekers. Contextualizing preferences across domains may provide a potential solution. In settings where grounding is critical, dialogue agents should prefer to use grounding acts. Still, while grounding acts provides an umbrella for grounding strategies, a closer examination of strategies routinely used by domain experts can inform preferred interaction.

On alignment beyond single-step interaction. Current RLHF-trained language models are trained to optimize single-step interaction. Humans, however, strategically use grounding acts across multiple turns. While older NLP systems like ELIZA (Weizenbaum, 1966)—a psychotherapy chatbot—do not explicitly model human grounding patterns, these systems still incorporate a large number of clarification and follow-up transforms. Similarly, we can augment SFT and preference datasets with grounding acts, or train reward models across multi-step interactions (Hong et al., 2023). However, simply using grounding acts to augment training does not guarantee that LLMs are well-aligned with humans. For example, while prompting LLMs to use grounding acts increases the use of underlying dialogue acts, it does not improve κ agreement with humans (§7.3). When designing training curricula, grounding acts offer a promising direction toward measurable and grounded human-AI interaction.

10 Conclusion

Instruction-following language models are trained to "follow" instructions. Thus, datasets and algorithms for finetuning LLMs are designed around single-step interaction. In this work, we outline a set of discourse acts—grounding acts—to measure interaction with language models *beyond* instruction following. We apply theory from conversational grounding to interactions with LLMs, finding significant differences between how humans ground in dialogue and how LLMs generate grounding acts. Designing new datasets, models, and methods, motivated by prior work on conversational grounding, will likely be necessary to minimize the grounding gap.

Limitations

Our characterization of simulated supporters contains some anthropomorphic metaphors, which are known to be harmful as anthropomorphism in discussing technology has long been connected to dehumanization (Bender, 2022; Abercrombie et al., 2023; Cheng et al., 2024). For brevity, we discuss simulated supporters leveraging grounding acts as a way to refer to whether LLM generations contain grounding acts.

The set of grounding acts we consider, while simple, is also not a comprehensive collection of all grounding dialogue acts used in conversation. For example, our grounding acts focus on strategies where an expert listener uses *positive* grounding—negative grounding acts like model-initiated repair are out of scope. Still, we should expect that well-aligned use of grounding acts from a speaker will be correlated with a decrease in negative grounding. For example, if a speaker clarifies appropriately, then a listener should repair less. Finally, a range of grounding acts are likely subsets of our synthesized selection: paraphrasing and restating is a subset of clarification, repeating a prior utterance is an acknowledgment, etc. A finer-grained breakdown on our selected grounding acts may yield further insights—though we observe limited use even for our more general categorization. Finally, all our datasets are in English, and our collection of grounding acts is English-centered.

There also exist interaction effects between grounding acts in conversation. Our current metrics do not explicitly analyze the interaction between individual acts (e.g. does listener's acknowledgment follow a simulated speaker's clarification?). Suc-

cessful goal-oriented dialogue requires an understanding of the interaction effects between grounding acts—we leave this analysis to future work.

Furthermore, our `GPT-4` based grounding acts classifier may not generalize to domains beyond our selected datasets. Analyses on other datasets may require building a new classifier or reprompting an LLM like `GPT`. Our selected datasets are also synthetic in nature, consisting of interactions between crowd workers. Validating grounding acts agreement between humans and LLMs on in-the-wild interaction is an avenue for future work.

Ethics Statement

Measurable grounding acts introduce new directions toward improved simulations of human conversational behavior, which can be useful for training (Shaikh et al., 2023a; Wang and Demszky, 2023). However, we strongly caution against using grounding acts to help replace a human support-provider—especially in high-risk scenarios such as therapy and education.

Furthermore, enabling LLMs to ask numerous clarification and followup questions can be privacy-intrusive, leading support-seekers to disclose privacy-sensitive information. Using grounding acts while collecting relevant information only is an open challenge and an avenue for future work.

Finally, we note that efficiently grounding can be harmful when the underlying goals are harmful. While we explore Persuasion for Social Good as a dataset, one can imagine settings where grounding acts are applied to persuade for more nefarious topics (e.g. political microtargeting).

Acknowledgements

Omar Shaikh is supported by the Brown Institute's Magic Grant. Kristina Gligorić is supported by Swiss National Science Foundation (Grant P500PT-211127). This work is also funded by the Hoffman–Yee Research Grants Program and the Stanford Institute for Human-Centered Artificial Intelligence. We thank members of the Jurafsky Lab, SALT Lab, Will Held, and Eric Horvitz, for their valuable feedback on this manuscript.

References

Gavin Abercrombie, Amanda Cercas Curry, Tanvi Dinkar, and Zeerak Talat. 2023. Mirages: On anthropomorphism in dialogue systems. *arXiv preprint arXiv:2305.09800*.

Gati Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2022. Using large language models to simulate multiple humans. *arXiv preprint arXiv:2208.10264*.

Lisa P Argyle, Ethan Busby, Joshua Gubler, Chris Bail, Thomas Howe, Christopher Rytting, and David Wingate. 2023. Ai chat assistants can improve conversations about divisive topics. *arXiv preprint arXiv:2302.07268*.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua Gubler, Christopher Rytting, and David Wingate. 2022. Out of one, many: Using language models to simulate human samples. *arXiv preprint arXiv:2209.06899*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Vevake Balaraman, Arash Eshghi, Ioannis Konstas, and Ioannis Papaioannou. 2023. No that's not what i meant: Handling third position repair in conversational question answering.

Cristian-Paul Bara, CH-Wang Sky, and Joyce Chai. 2021. Mindcraft: Theory of mind modeling for situated dialogue in collaborative tasks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1112–1125.

Emily M. Bender. 2022. Resisting dehumanization in the age of "AI". Plenary talk at the 44th Annual Meeting of the Cognitive Science Society (CogSci).

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Luciana Benotti and Patrick Blackburn. 2021a. A recipe for annotating grounded clarifications. *arXiv preprint arXiv:2104.08964*.

Luciana Benotti and Patrick Rowan Blackburn. 2021b. Grounding as a collaborative process. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 515–531. Association for Computational Linguistics.

Harry Bunt, Volha Petukhova, David Traum, and Jan Alexandersson. 2017. Dialogue act annotation with the iso 24617-2 standard. *Multimodal Interaction with W3C Standards: Toward Natural User Interfaces to Everything*, pages 109–135.

Andrew Caines, Helen Yannakoudakis, Helen Allen, Pascual Pérez-Paredes, Bill Byrne, and Paula Buttery. 2022. The teacher-student chatroom corpus version 2: more lessons, new annotation, automatic detection of sequence shifts. In *Proceedings of the 11th*

Workshop on NLP for Computer Assisted Language Learning, pages 23–35, Louvain-la-Neuve, Belgium. LiU Electronic Press.

Andrew Caines, Helen Yannakoudakis, Helena Edmondson, Helen Allen, Pascual Pérez-Paredes, Bill Byrne, Paula Buttery, and Cambridge Assessment. 2020. The teacher-student chatroom corpus. *Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2020)*, page 10.

Per Carlbring, Heather Hadjistavropoulos, Annet Kleiboer, and Gerhard Andersson. 2023. A new era in internet interventions: The advent of chat-gpt and ai-assisted therapist guidance. *Internet Interventions*, 32.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*.

Khyathi Raghavi Chandu, Yonatan Bisk, and Alan W Black. 2021. Grounding 'grounding'in nlp. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4283–4305.

Myra Cheng, Kristina Gligoric, Tiziano Piccardi, and Dan Jurafsky. 2024. AnthroScore: A computational linguistic measure of anthropomorphism. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, Malta. Association for Computational Linguistics.

Hyundong Cho and Jonathan May. 2020. Grounding conversations with improvised dialogues. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2398–2413, Online. Association for Computational Linguistics.

Jennifer Chu-Carroll and Sandra Carberry. 1998. Collaborative response generation in planning dialogues. *Computational Linguistics*, 24(3):355–400.

Herbert H Clark. 1996. *Using language*. Cambridge university press.

Herbert H Clark and Edward F Schaefer. 1989. Contributing to discourse. *Cognitive science*, 13(2):259–294.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. Ultrafeedback: Boosting language models with high-quality feedback. *arXiv preprint arXiv:2310.01377*.

Deborah Davis. 1982. Determinants of responsiveness in dyadic interaction. *Personality, roles, and social behavior*, pages 85–139.

Dorottya Demszky, Jing Liu, Heather C Hill, Dan Jurafsky, and Chris Piech. 2021. Can automated feedback improve teachers' uptake of student ideas? evidence from a randomized controlled trial in a large-scale online course. edworkingpaper no. 21-483. *Annenberg Institute for School Reform at Brown University*.

Morton Deutsch. 1973. *The resolution of conflict: Constructive and destructive processes*. Yale University Press.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations.

Arash Eshghi, Christine Howes, Eleni Gregoromichelaki, Julian Hough, Matthew Purver, et al. 2015. Feedback in conversation as incremental semantic update. Association for Computational Linguistics.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with v-usable information. In *International Conference on Machine Learning*, pages 5988–6008. PMLR.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

Daniel Fried, Nicholas Tomlin, Jennifer Hu, Roma Patel, and Aida Nematzadeh. 2022. Pragmatics in grounded language learning: Phenomena, tasks, and modeling approaches. *arXiv preprint arXiv:2211.08371*.

Yifan Gao, Henghui Zhu, Patrick Ng, Cicero dos Santos, Zhiguo Wang, Feng Nan, Dejiao Zhang, Ramesh Nallapati, Andrew O Arnold, and Bing Xiang. 2021. Answering ambiguous questions through generative evidence fusion and round-trip prediction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3263–3276.

Jonathan Ginzburg and Robin Cooper. 2001. Resolving ellipsis in clarification. In *39th Annual Meeting of the Association-for-Computational-Linguistics*, pages 236–243. Association for Computational Linguistics.

Arthur C Graesser, Natalie K Person, and Joseph P Magliano. 1995. Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied cognitive psychology*, 9(6):495–522.

Patrick GT Healey, Arash Eshghi, Christine Howes, and Matthew Purver. 2011. Making a contribution: Processing clarification requests in dialogue. In *Proceedings of the 21st Annual Meeting of the Society for Text and Discourse*, pages 11–13. Citeseer.

Patrick GT Healey, Matthew Purver, James King, Jonathan Ginzburg, and Greg J Mills. 2003. Experimenting with clarification in dialogue. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 25.

Joey Hong, Sergey Levine, and Anca Dragan. 2023. Zero-shot goal-directed dialogue via rl on imagined conversations. *arXiv preprint arXiv:2311.05584*.

Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 159–166.

Shang-Ling Hsu, Raj Sanjay Shah, Prathik Senthil, Zahra Ashktorab, Casey Dugan, Werner Geyer, and Diyi Yang. 2023. Helping the helper: Supporting peer counselors via ai-empowered practice and feedback. *arXiv preprint arXiv:2305.08982*.

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*.

Gail Jefferson. 1972. Side sequences. In D.N. Sudnow, editor, *Studies in social interaction*, pages 294–333. Free Press, New York.

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

Elise Karinshak, Sunny Xun Liu, Joon Sung Park, and Jeffrey T Hancock. 2023. Working with ai to persuade: Examining a large language model's ability to generate pro-vaccination messages. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–29.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. 2024. Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2022. Clam: Selective clarification for ambiguous questions with large language models. *arXiv preprint arXiv:2212.07769*.

Belinda Z. Li, Alex Tamkin, Noah Goodman, and Jacob Andreas. 2023. Eliciting human preferences with language models.

Xiaolong Li and Kristy Boyer. 2015. Semantic grounding in dialogue for complex problem solving. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 841–850, Denver, Colorado. Association for Computational Linguistics.

Jessy Lin, Nicholas Tomlin, Jacob Andreas, and Jason Eisner. 2023. Decision-oriented dialogue for human-ai collaboration. *arXiv preprint arXiv:2305.20076*.

Diane J. Litman. 1985. *Plan Recognition and Discourse Analysis: An Integrated Approach for Understanding Dialogues*. Ph.D. thesis, University of Rochester, Rochester, NY.

Diane J. Litman and James Allen. 1987. A plan recognition model for subdialogues in conversation. *Cognitive Science*, 11:163–200.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483.

Karen E. Lochbaum. 1998. A collaborative planning model of intentional structure. *Computational Linguistics*, 24(4):525–572.

Xing Han Lu, Siva Reddy, and Harm de Vries. 2023. The StatCan dialogue dataset: Retrieving data tables through conversations with genuine intents. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2799–2829, Dubrovnik, Croatia. Association for Computational Linguistics.

Brielen Madureira and David Schlangen. 2023a. " are you telling me to put glasses on the dog?"content-grounded annotation of instruction clarification requests in the codraw dataset. *arXiv preprint arXiv:2306.02377*.

Brielen Madureira and David Schlangen. 2023b. Instruction clarification requests in multimodal collaborative dialogue games: Tasks, and an analysis of the codraw dataset. *arXiv preprint arXiv:2302.14406*.

Sabrina J Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872.

William R Miller and Stephen Rollnick. 2012. *Motivational interviewing: Helping people change*. Guilford press.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. Ambigqa: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797.

Shrestha Mohanty, Negar Arabzadeh, Julia Kiseleva, Artem Zholus, Milagro Teruel, Ahmed Awadallah, Yuxuan Sun, Kavya Srinet, and Arthur Szlam. 2023. Transforming human-centered ai collaboration: Redefining embodied agents capabilities through interactive grounded language instructions. *arXiv preprint arXiv:2305.10783*.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Tim Paek and Eric Horvitz. 1999. Uncertainty, utility, and misunderstanding: A decision-theoretic perspective on grounding in conversational systems. In *AAAI Fall Symposium on Psychological Models of Communication, North*.

Ashwin Paranjape and Christopher D Manning. 2021. Human-like informative conversations: Better acknowledgements using conditional mutual information. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 768–781.

Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*.

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–18.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. 2022. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*.

Ildikó Pilán, Laurent Prévot, Hendrik Buschmeier, and Pierre Lison. 2023. Conversational feedback in scripted versus spontaneous dialogues: A comparative analysis. *arXiv preprint arXiv:2309.15656*.

Matthew Purver, Jonathan Ginzburg, and Patrick Healey. 2003a. On the means for clarification in dialogue. *Current and new directions in discourse and dialogue*, pages 235–255.

Matthew Purver, Patrick Healey, James King, Jonathan Ginzburg, and Greg J Mills. 2003b. Answering clarification questions. In *Proceedings of the Fourth SIGdial Workshop of Discourse and Dialogue*, pages 23–33.

Matthew Richard John Purver. 2004. *The theory and use of clarification requests in dialogue*. Ph.D. thesis, University of London.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

Hossein A Rahmani, Xi Wang, Yue Feng, Qiang Zhang, Emine Yilmaz, and Aldo Lipani. 2023. A survey on asking clarification questions datasets in conversational systems. *arXiv preprint arXiv:2305.15933*.

Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1978. A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*, pages 7–55. Elsevier.

Emanuel A Schegloff. 1982. Discourse as an interactional achievement: Some uses of 'uh huh'and other things that come between sentences. *Analyzing discourse: Text and talk*, 71:71–93.

Deborah Schiffrin. 1987. *Discourse markers*. 5. Cambridge University Press.

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

Omar Shaikh, Valentino Chai, Michele J Gelfand, Diyi Yang, and Michael S Bernstein. 2023a. Rehearsal: Simulating conflict to teach conflict resolution. *arXiv preprint arXiv:2309.12309*.

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

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. 2023. A long way to go: Investigating length correlations in rlhf. *arXiv preprint arXiv:2310.03716*.

Svetlana Stoyanchev, Alex Liu, and Julia Hirschberg. 2013. Modelling human clarification strategies. pages 137–141.

Ira Strumwasser, Nitin V Paranjpe, Marianne Udow, David Share, Mary Wisgerhof, David L Ronis, Charlotte Bartzack, and Ali N Saad. 1991. Appropriateness of psychiatric and substance abuse hospitalization: implications for payment and utilization management. *Medical Care*, pages AS77–AS90.

Alex Tamkin, Kunal Handa, Avash Shrestha, and Noah Goodman. 2022. Task ambiguity in humans and language models. *arXiv preprint arXiv:2212.10711*.

Ben M Tappin, Chloe Wittenberg, Luke Hewitt, David Rand, et al. 2023. Quantifying the persuasive returns to political microtargeting.

Alberto Testoni, Claudio Greco, Tobias Bianchi, Mauricio Mazuecos, Agata Marcante, Luciana Benotti, and Raffaella Bernardi. 2020. They are not all alike: Answering different spatial questions requires different grounding strategies. In *Proceedings of the Third International Workshop on Spatial Language Understanding*, pages 29–38.

David R Traum and Elizabeth A Hinkelman. 1992. Conversation acts in task-oriented spoken dialogue. *Computational intelligence*, 8(3):575–599.

Polina Tsvilodub, Michael Franke, Robert D Hawkins, and Noah D Goodman. 2023. Overinformative question answering by humans and machines. *arXiv preprint arXiv:2305.07151*.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment.

Rose E Wang and Dorottya Demszky. 2023. Is chatgpt a good teacher coach? measuring zero-shot performance for scoring and providing actionable insights on classroom instruction. *arXiv preprint arXiv:2306.03090*.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.

Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1):36–45.

Ryen W White and Eric Horvitz. 2009. Cyberchondria: studies of the escalation of medical concerns in web search. *ACM Transactions on Information Systems (TOIS)*, 27(4):1–37.

Martha Stone Wiske. 1998. *Teaching for Understanding. Linking Research with Practice. The Jossey-Bass Education Series.* ERIC.

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, et al. 2023. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. *arXiv preprint arXiv:2309.11998*.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023. Navigating the grey area: Expressions of overconfidence and uncertainty in language models. *arXiv preprint arXiv:2302.13439*.

Pei Zhou, Hyundong Cho, Pegah Jandaghi, Dong-Ho Lee, Bill Yuchen Lin, Jay Pujara, and Xiang Ren. 2022. Reflect, not reflex: Inference-based common ground improves dialogue response quality. *arXiv preprint arXiv:2211.09267*.

A Dataset Details

We discuss additional details regarding our datasets. All our selected datasets are in English, and are under an open-source license (MIT, Apache, etc.) Datasets were already all appropriately anonymized.

Teacher Student Chatroom Corpora (TSCC) is a collection of written conversations captured during one-to-one lessons between teachers and learners of English (Caines et al., 2020, 2022). The lessons took place in a synchronous chatroom. The dataset contains a total of 260 conversations, spread across two dataset releases. The one-to-one chatroom lessons allowed interactive, immediate, and personalized conversations. We selected the tutoring domain since numerous opportunities of current LLMs for education have been proposed in the literature, for instance, to create educational content, facilitate instruction, student engagement and interaction, provide feedback, and personalize learning experiences (Kasneci et al., 2023; Demszky et al., 2021; Wang and Demszky, 2023). In line with prior applications of LLMs, we use **the teacher** as the expert from this corpus.

Persuasion for Good consists of one-to-one online conversations between a persuader and a persuadee (Wang et al., 2019). In the data collection,

one participant was asked to persuade the other to donate to a specific charity by presenting personally relevant and appealing arguments. The dataset contains 1017 conversations and was collected on a crowdsourcing platform. Participants were encouraged to continue the conversation until an agreement was reached on donating and, if so, how much. We selected the persuasion domain since the use of current LLMs for persuasion has been proposed, for instance, to facilitate conversations about politically divisive topics (Argyle et al., 2023; Tappin et al., 2023), or to generate persuasive pro-vaccination messages (Karinshak et al., 2023). In this corpus, **the persuader** is the expert.

Emotional Support Conversations (ESConv) is a corpus of one-to-one online conversations between a help-seeker and a help-provider, collected via crowdsourcing (Liu et al., 2021). The dataset contains 1053 conversations. Before participation, help providers were trained to provide effective support through an emotional support tutorial covering support stages and concrete strategies. We selected the emotional support domain since people turn to the web and widely accessible LLMs to seek information and receive support related to their wellbeing (Carlbring et al., 2023; White and Horvitz, 2009). In this corpus, the **health provider** is the expert.

B A Human Reference Point for κ

We observed low agreement between `ChatGPT-3.5` and the original human dialogue participants in using grounding acts. But the original dialogue participants presumably had lots of information that the LLM might not have had. To confirm that the LLM agreement with humans is low, we would need a fairer comparison: comparing the agreement in grounding between two human crowdworkers who are both generating a next turn given the same conversational background as the language model. We set up this small parallel task with humans, using ESConv as our evaluation task.

Similar to our LM setup, we provide our annotators with a random sample of 50 contexts $D_{1...t-1}$, and ask two different annotators to independently complete the next message, using the same prompt as with the LM. Through the study description, we informed participants that their evaluations would be used to benchmark current LLMs like `ChatGPT`. Given the domain, we recruited individuals on Prolific who self-reported as (1) fluent English speakers, (2) recorded their workplace function as a Healthcare Professionals and (3) listed their employment role as a Therapist / Well-being counselor. Annotators were paid at a rate of $12 / hour. Despite limited instruction, human-human agreement across using any grounding act is fair, with Followup $\kappa = 35.66$, Clarification $= 48.45$, and Acknowledgement $= 29.15$. All human-human κ scores are substantially higher than Human-LM counterparts. Note that this is still effectively a *lower bound* on κ between two independently recruited individuals—with additional discussion between two mental health supporters, we would expect this score to be much higher.[6]

C Sampling Conversations

Because we aim to study grounding in language models, we must control for context length within datasets. With in-context learning (ICL), models may unfairly adapt to longer conversation history, learning to employ grounding acts only on longer conversations. We preprocess our dataset to control for ICL. First, we merge successive turns between the same participant into a single message.[7] Then, within each dataset, we sample 100 random conversations greater than or equal to the median number of messages in a dataset (TSCC = 92, ESConv = 22, Persuasion = 20). Finally, we truncate the sampled conversations to the median length.

D Details on Error Taxonomy

We developed a taxonomy of five error types based on the notes about emerging reasons why human used grounding acts, with each error type corresponding to a specific use of a grounding act (described in Appendix, Table 4). Two authors then independently annotated the sentences, indicating whether any of the five error categories is present. We found a substantial inter-rater agreement between the two annotators (Cohen's $\kappa = 0.77$). The labels from the two annotators were then aggregated such that an error category label is assigned if both of the annotators assigned it. Overall, 82.5% of examples from the sampled set were assigned to one of the categories.

[6] We note that clarification κ might be inflated, since support is low. See Table 1 in Appendix for details.

[7] Our selected datasets also contain successive messages from the same participant (average 23%).

E Prompt Mitigation

We use the following prompt mitigation, designed from our qualitative analysis (§6.2). We include this in OpenAI's `system prompt` field.

> *Make sure you ask clarification questions to verify that you understand what a person is saying or to resolve an ambiguity. Also, use follow-up questions to inquire about related topics, or to continue a conversation naturally. Finally, acknowledge what a person is saying to show empathy (e.g. "o.k.", "I understand," etc.) when necessary. Make sure you use these strategies carefully and like a trained therapist; do not overuse them.*

F Training Details

During the SFT stage, we use an effective batch size of 128 (gradient accumulation) with a learning rate of 3e-4. In the preference optimization stage, we train with a learning rate of 5e-6, with a batch size of 32. To reduce memory usage, we use LoRA (Hu et al., 2021) and train on 1 A100 GPUs. Modeling took a total of 1 day. For all experiments, we modified the HuggingFace transformers package (Wolf et al., 2020).

Grounding Act	Cohen κ	Support
Followup	35.66	32
Ack.	29.15	11
Clarification	48.45	4

Table 1: Human-Human κ and support for grounding acts across 100 annotations

Dataset	Grounding Act	Zero-shot	Few-shot	Support
TSCC	Follow-up	0.55	**0.91**	47
	Ack	0.67	**0.89**	65
	Clarification	0.38	**0.81**	25
ESConv	Follow-up	0.90	**0.93**	26
	Ack	0.59	**0.90**	14
	Clarification	**0.83**	0.75	5
Persuasion	Follow-up	0.64	**0.89**	9
	Ack	0.14	**0.91**	16
	Clarification	0.00	**1.00**	2

Table 2: F-1 for grounding acts classification on a withheld test set of 10 conversations from our selected dialogue datasets. In the few-shot setting, `GPT-4` has fairly high F-1 across grounding acts.

Grounding Act	Source	Base Rate (%)	Cohen Kappa (κ)
Followup	Human	27.87	36.41
	Llama 7B	15.59	18.23
	+ SFT	15.21	19.12
	+ SFT + DPO	11.21	12.13
	+ SFT + PPO	12.12	13.24
Acknowledgement	Human	12.89	30.12
	Llama 7B	15.28	22.77
	+ SFT	16.73	28.56
	+ SFT + DPO	16.11	18.23
	+ SFT + PPO	14.05	19.54
Clarification	Human	3.05	48.45
	Llama 7B	3.01	14.47
	+ SFT	2.24	14.44
	+ SFT + DPO	2.11	5.19
	+ SFT + PPO	2.58	6.52

Table 3: Grounding act metrics across the Archangel (Ethayarajh et al., 2024) suite of models. We find that PPO / DPO results in decreases for both base-rate and κ compared to the SFT-only model.

Error type and frequency	Description
+ Acknowledgement to show empathy (43.94%) **Example:**	Human acknowledges to show empathy, or display understanding, as opposed to going straight to suggestions and solutions. `seeker:` honestly im not sure. there is always that risk when doing that with guys. i better think long and hard before doing that! `supporter:` It sounds like you put a lot of thought into things which can be helpful in decision making. ↪ `GPT:` It's important to consider the potential consequences before taking any action. It may be helpful to evaluate what you truly want and need in a relationship before making any decisions.
+ Followup to continue the conversation (12.88%) **Example:**	Human asks a related question to keep the conversation going, typically not aiming to finish the exchange. `seeker:` My friends, who I also work with, are upset with me because I am leaving for a new position. I am sad that they are upset. `supporter:` Ah, that can be awkward. Are you happy about the new position? ↪ `GPT:` I can understand [...]. Communication is key in maintaining healthy relationships.
+ Followup to inquire further (26.52%) **Example:**	Human asks a closely related followup question to collect more information relevant to the topic, typically not aiming to finish the exchange. `seeker:` I don't have close friends because I don't trust people `supporter:` You said you generally don't trust family or people in general? ↪ `GPT:` I understand that feeling. [...] Take your time, and remember that trust is a two-way street. It requires both vulnerability and the willingness to give others a chance to earn your trust.
+ Clarification to verify understanding (9.09%) **Example:**	Before providing specific suggestions, the human restates their understanding in the form of a clarification question, as opposed to assuming correctness and providing suggestions (presumptive grounding). `seeker:` i was cheated on severally, and it resulted me to end the relationship `supporter:` It sounds like you had to deal with a partner recently cheating and a breakup, is this right? ↪ `GPT:` I'm sorry to hear that you've been through such a difficult experience [...]
+ Clarification to resolve an ambiguity (7.58%) **Example:**	Human resolves any ambiguities related to a specific entity, as opposed to assuming correctness and providing suggestions (presumptive grounding). `seeker:` Sometimes abbreviated as 121 `supporter:` Hmm, it's fine, to be honest. At [anonymized] they were always called one-to-one classes. Or do you mean 121 is not formal? ↪ `GPT:` "121" is more commonly used in informal settings or casual conversations. In official or formal contexts, it's better to use the full term "one-to-one."

Table 4: **Qualitative analyses.** Detailed descriptions of error types, their frequency, and example conversations.