

# Evaluating the Effectiveness of Large Language Models in Establishing Conversational Grounding

**Biswesh Mohapatra**

Inria

Biswesh.Mohapatra@inria.fr

**Manav Nitin Kapadnis**

Carnegie Mellon University

mkapadni@andrew.cmu.edu

**Laurent Romary**

Inria

Laurent.Romary@inria.fr

**Justine Cassell**

Inria, Carnegie Mellon University

Justine.Cassell@inria.fr

## Abstract

Conversational grounding, vital for building effective dialogue between people and between people and dialogue systems, involves ensuring a mutual understanding of shared information. Despite its importance, there has been limited research on this aspect of conversation in recent years, especially after the advent of Large Language Models (LLMs). Previous studies have highlighted the shortcomings of some pre-trained language models in conversational grounding. However, most testing for conversational grounding capabilities involves human evaluations that are costly and time-consuming. This has led to a lack of testing across multiple models of varying sizes, a critical need given the rapid rate of new model releases. This gap in research becomes more significant considering recent advances in language models, which have led to new emergent capabilities. In this paper, we evaluate the performance of LLMs in various aspects of conversational grounding and analyze why some models perform better than others. We demonstrate a direct correlation between the size of the pre-training dataset, size of the model and conversational grounding abilities, suggesting that they have independently acquired some pragmatic capabilities from larger pre-training datasets. Finally, we propose ways to enhance the capabilities of the models that lag in our tests.

## 1 Introduction

The concept of "common ground" in linguistics, introduced by Clark and Brennan (Clark and Brennan, 1991), refers to the collective body of knowledge and assumptions that conversation participants build together. This shared understanding is not solely formed through words; it also incorporates other modalities, as highlighted by Nakano et al. (Nakano et al., 2003), such as gestures, nods, and eye contact. These non-verbal cues are crucial in creating and maintaining the common ground

in face-to-face dialogues. Conversational Grounding is this process of building common ground, and involves continuous negotiation, and resolution of uncertainty. These uncertainties may be addressed by providing additional context - for example, specifying the object "the small gate next to the bakery" or through the listener seeking clarification, like asking, "You mean the white gate?". Since these agreements are not always explicitly expressed, participants must be adept at recognizing cues of (mis-)understanding from their interlocutors. Conversational Grounding is complex due to the dynamic nature of spontaneous dialogues, which demand more than just grammatical correctness, but also pragmatics - how language is used.

An effective grounding mechanism is vital to dialogue systems in order to reduce ambiguity both when the system is the speaker, by adding more information if it senses confusion from the listener, and as listener, by asking for clarifications when necessary. (Benotti and Blackburn, 2021) have previously shown that state-of-the-art Language Models pre-trained on large amounts of conversational data like BlenderBot 1 (Roller et al., 2020) frequently fall short in ensuring that information has been adequately grounded. While they identified these deficiencies, their work didn't provide a comprehensive framework for evaluating multiple different models on their grounding capabilities. In addition, their findings were based on limited interactions with smaller models such as BlenderBot 1, trained on smaller amount of data compared to more recent and advanced LLMs like Llama (Touvron et al., 2023) and GPT4 (OpenAI, 2023).

While Conversational Grounding can be a multimodal phenomenon, in this paper, we start by evaluating the performance of state-of-the-art Language Models on textual inputs, since current dialog systems convert speech to text before sending it as input to Language Models for NLU, NLG, and in some cases dialog management itself. The out-

puts are later converted into speech using separate modules. Because of their growing significance in the field of dialogue systems, we aim here to assess and enhance pre-trained LLMs' capabilities in the various kinds of conversational grounding. This will lay the groundwork for more advanced modular spoken dialog systems with multimodal input and output in the future.

To this end, we have devised a series of tests to evaluate LLMs. We evaluate how they handle grounding both as speaker and listener. When the models act as a listener, we test if they can ground information that was repaired or cancelled by the speaker, and we test if they can ask questions in case of important uncertainties. When the models act as speaker, we test if they can recognize and provide correct information to reduce ambiguity. Our approach involves analyzing the model perplexity of two hand-crafted responses for a given context: one being contextually appropriate and the other deceptively fitting but contextually incorrect. By comparing the perplexities of these responses, we gauge the model's proficiency in specific grounding scenarios. Our findings indicate a correlation between model performance, its size in terms of parameters and the amount of data that it was pre-trained on. Consequently, we conduct novel tests to explore the reasons behind the under-performance of smaller models, focusing specifically on their embedding vectors. The insights gained from these investigations are then utilized to explore methods for enhancing the performance of these smaller models in conversational grounding tasks.

## 2 Related Work

In the field of linguistics, (Clark and Brennan, 1991) explored the inherent uncertainty present in dialogue, which interlocutors negotiate and resolve, to the extent necessary for that stage of the dialogue. Clark identified four distinct states of uncertainty: 1) B didn't notice that A uttered any utterance *u*. 2) B noticed that A uttered some *u*. 3) B correctly heard *u*. 4) B understood what A meant by *u*.

The field of conversational systems has seen various efforts aimed at addressing the challenges of grounding, particularly in the context of rule-based modular dialog systems. The pioneering work originated with (Traum and Allen, 1994), which introduced the concept of Grounding Acts (GAs). It serves as a framework for breaking down the grounding process into its fundamental units. Within this framework, they define 7 categories

of GAs: Initiate, Continue, Acknowledge, Repair, Request-Repair, Request-Acknowledge, and Cancel. The framework also introduced the concept of 'Grounding Unit', a group of utterances providing a piece of information starting with an 'Initiate' and then adding to the common ground with an 'Acknowledgment'. Other GAs help in negotiating this information before adding to the common ground. While this concept offers a solid foundation for understanding and modeling conversational grounding, its application has been limited in the context of contemporary LLMs. (Mohapatra et al., 2024) have recently expanded on the challenges associated with annotating and modeling grounding acts within dialogue systems.

Subsequent approaches, such as Centering Theory (Grosz et al., 1983) and Domain Reference theory (Denis, 2010), introduced techniques for representing and managing grounded information. Their applicability, however, was largely limited to closed domains, primarily owing to their reliance on rule-based approaches.

(Fried et al., 2021) focused on reference-centric models by leveraging the success of recent artificial neural networks. However, it was designed for the specific multimodal reference task in the Onecommon dataset (Udagawa and Aizawa, 2019) using an end-to-end neural-network based model.

Recent research on generative agents has highlighted the potential of LLMs in interactive settings. (Park et al., 2023) studied the creation of multiple agents, each assigned an initial identity. These agents were equipped with a memory module and relied on LLMs to assess the significance of various memories. The study demonstrated their ability to plan relevant events and execute them through human-like interactions. However, it's important to note that this research was conducted in a virtual environment with artificial agents, which does not fully replicate the scenario of human-human or human-agent spontaneous conversations. While this work sheds light on the potential capabilities of LLMs, and is contrary to the results of previous work like (Benotti and Blackburn, 2021), further investigation is warranted in assessing their effectiveness in handling various grounding phenomena in natural conversations.

## 3 Dataset

Several datasets have been curated to support research on conversational grounding. Talk The Walk (de Vries et al., 2018) created a virtual 2D grid en-

vironment, while the HCRC Maptask (Thompson et al., 1993) had participants discuss directions using a map that, in fact, differed between the two participants, forcing them to negotiate what they were referring to at any given moment. These conversations helped in the development of early theories and models for grounding. After assessing the existing datasets, we opted to employ the Meetup dataset (Ilinykh et al., 2019) to generate our test cases. This choice was made due to the nature of the task that required participants to negotiate instances of referring expressions.

The Meetup dataset introduced a scenario wherein two participants are placed on a 2D grid, with each vertex representing a room. The objective is for the two participants to converge in the same room, despite only being able to see their own rooms. Navigational actions (east, west, north, or south) move participants to new rooms, unveiling the image of the newly entered room to them. Achieving the goal necessitates the articulation of room descriptions, formulation and communication of a converging strategy, retention of room descriptions shared by the counterpart, and mental modeling of the other participant’s room configurations. Although the dataset is text-based, it serves as a great resource for exploring and developing grounding models. Unlike many tasks that attribute a leading role to one participant, in this task, both participants can assume initiator or responder roles interchangeably. These features made the dataset ideal for our purposes. The GA annotated version of the meetup dataset provided by (Mohapatra et al., 2024) was used for creating testing cases for model evaluation.

#### 4 Models

In order to examine LLMs of varying sizes and pre-trained on varying amounts of data, we decided to test T5-Large (Raffel et al., 2020), Godel-Large (Peng et al., 2022), Llama-1(7 & 13 Billion)(Touvron et al., 2023), GPT 3.5(OpenAI, 2022) and GPT 4 (OpenAI, 2023). T5 is an encoder-decoder-based transformer model, while Godel, builds upon T5 with additional fine-tuning for conversational applications. Both models possess 770 million parameters and are not trained on a lot of pre-trained data. Llama and the GPT models, in contrast, are decoder-based transformer models trained on vast amount of data. For T5, Godel, Llama(7B, 13B), access to the models allowed for additional fine-tuning using Causal Language Mod-

Instructions : Here is a conversation between two participants ..... to both participants.  
 Following is the dialog history along with image descriptions:  
 <Image A> The image showcases an oven ..... is located near the table.  
 [00:00:25] A: I’m in a kitchen  
 [00:00:43] B: In a dining room with 4 brown toys  
 [00:00:48] A: let me go north  
 <Image A> There are 4 chairs and a dining table ..... with a photo hanging on the wall.  
 [00:00:54] A: I see a dining room, but not your one

Figure 1: Example of input context provided to the models with the instructions, image descriptions and dialog history. See appendix for more complete instructions and image descriptions.

eling(CLM : predicting subsequent tokens based on prior sequences) on the entire Meetup dataset. This enabled testing of both the original (vanilla) and CLM fine-tuned versions of these models. The fine-tuned models were not exposed to the answers of the modified dialog test cases beforehand, ensuring an unbiased evaluation of their performance. See the appendix for the finetuning training setup. GPT-3.5 and GPT-4, however, are proprietary models that require distinct methodologies for evaluation. In contrast, we utilize the Llama 3.1 (8 Billion) model to address specific questions raised by our findings. Collectively, this diverse set of models allows us to better understand the influence of various factors on conversational grounding.

#### 5 Testing

**Perplexity Testing :** We conduct an assessment of the model perplexity of candidates for the response to a given context. Perplexity(PPL) is a measure of how well a language model predicts a sample. It quantifies the model’s uncertainty in predicting a sequence of words as given by the equation -

$$PPL(W) = e^{-\frac{1}{N} \sum_{i=1}^N \log_e P(w_i|w_1, \dots, w_{i-1})} \quad (1)$$

Here,  $W$  represents the sequence of words  $w_1, \dots, w_N$ ,  $N$  is the length of the word sequence, and  $P(w_i|w_1, \dots, w_{i-1})$  is the probability of each word. A lower perplexity indicates a higher chance of the model accurately generating the sequence.

We evaluate the models’ ability to ground the conversations in two contexts - When they act as a listener and when as a speaker. To assess the LLM’s grounding ability as a listener to consider modifications from the speaker, we test it on situations containing the GAs - Repair and Cancel. We also test if it can detect ambiguities in the speaker’s utterances and respond accordingly. While acting as a speaker, we test them on another type of

GA - Request Repair, where they have to provide further relevant information when requested by the listener, in order to help in grounding. We also look at Request-Repairs that are of the Yes-No question type where the models tend to make contextual mistakes. Finally, we examined instances of complex anaphoric references made by the listener while asking for clarification. These references have to be disambiguated before responding to the requests for clarification.

To evaluate each phenomenon, we use the grounding act annotations of the Meetup Corpus provided by (Mohapatra et al., 2024). Then we pick 20 instances of each phenomenon and introduce slight modifications to create test cases. We pick only 20 instances because some scenarios, such as cancel, occur only 20 times, and we want equal numbers of instances for each phenomenon. As described below, the results nevertheless show a clear trend in the performance of different models. After identifying test cases, we create a correct and incorrect response for the context and analyze the model perplexity for those examples, as illustrated in Figure 2. Ideally, the perplexity of the correct response should be lower than the incorrect response.

Here we provide a detailed discussion of the test case creation process for each category -

1. **Reference Ambiguity** refers to instances where the referring expression is ambiguous. We test if the model is able to ask for clarifications in such cases of uncertainty, as seen in Figure 2. Examples of remaining categories can be found in section A.1 of appendix
2. **Repair** is the correction of previously uttered material or the addition of omitted material that will change the listener’s interpretation of the speaker’s intention. We check if the model can correctly take the repair from the speaker into account. The correct response accounts for the repaired information from the speaker while the wrong response contains the older information. The model’s ability to correctly take into account the repaired information shows its grounding abilities.
3. **Cancellation** refers to the speaker retracting previously grounded information. We check if the model can make the corrections to the grounded information. The correct response has the amended information while the wrong response doesn’t.

Table 1: Human Evaluation of Perplexity test cases

Options	% of times it was chosen
Correct Option	90.65
Wrong Option	1.55
Both options are valid	6.25
None of the options are valid	1.55

4. **Request Repair**: We test if the model acting as speaker can use the dialog context to disambiguate or clarify when the listener requests a repair. The correct response provides a contextually correct repair, while the wrong response doesn’t.
5. **Request Repair (Yes/No)**: For cases where the listener asks for an acknowledgment of what they have found, yes/no questions are used. Hence, we check if the model acting as the speaker can provide the correct repair instead of a generic yes/no answer.
6. **Anaphora** : We test if the model can link the anaphoric reference to the correct referent when the listener asks for clarifications. The correct response mentions the correct referent, unlike the wrong response.

Each input in our evaluation encompassed prior information, including instructions about the participants’ situation, game rules, dialog history, and descriptions of images that the participants were viewing during the experiment, as can be seen in Figure 1. The image descriptions were initially automatically generated using the Llava model (Liu et al., 2023) and subsequently refined manually to ensure the inclusion of all pertinent information. These image descriptions were then interspersed in utterances whenever room changes occurred.

We also tested the model proposed by (Roque and Traum, 2008) by finetuning a T5-based model to classify the grounding acts. We used the classified GAs of each utterance to calculate the degree of grounding(DoG). We later appended the GA and DoG information into the context and tested the T5 models to check if the traditional approaches could help the smaller models. We refer to this model as T5(DoG & GA).

**Prompt Testing** : Since it was not possible to calculate perplexity for the GPT3.5 and GPT4 models at the time of the study, we examined whether these models could select between the two response options as prompt as an alternative evaluation approach. In order to have a direct comparison with the smaller models, we also tested Llama-7B,



Table 2: Ratio of test cases where correct utterance had lower perplexity

Model	Repair	Cancel	Req-Repair(Y/N)	Req-Repair	Anaphora	Ref. Amb.
T5	0.45	0.55	0.65	0.50	0.45	0.35
Godel	0.40	0.65	0.45	0.50	0.35	0.40
T5 - CLM	0.45	0.50	0.40	0.45	0.30	0.45
Godel - CLM	0.35	0.50	0.45	0.45	0.40	0.45
Llama-7B	0.55	0.55	0.55	0.45	0.65	0.75
Llama-7B - CLM	0.50	0.55	0.55	0.45	0.70	<b>0.80</b>
Llama-13B	<b>0.75</b>	<b>0.75</b>	<b>0.70</b>	<b>0.80</b>	<b>0.75</b>	<b>0.80</b>
Llama-13B - CLM	0.65	<b>0.75</b>	0.65	0.70	0.70	0.70
Llama3.1-8B	0.70	<b>0.75</b>	<b>0.70</b>	0.75	<b>0.75</b>	0.75
Llama3.1-8B - CLM	0.70	0.65	<b>0.70</b>	0.65	0.70	0.70
T5(DoG & GA)	0.45	0.50	0.60	0.55	0.45	0.30

Table 3: Mean value of perplexity for correct utterances of each model

Model	Repair	Cancel	Req-Repair(Y/N)	Request-Repair	Anaphora	Ref. Amb.
T5	3.02e+15	3.46e+15	3.30e+15	2.81e+15	8.49e+14	2.00e+10
Godel	4233.29	4221.50	44379.42	44488.40	21724.60	25769.90
T5 - CLM	1.19	1.21	1.19	1.19	2.41	2.04
Godel - CLM	1.06	1.09	1.06	1.07	1.55	1.24
Llama-7B	7.12	7.10	7.75	8.00	6.93	7.70
Llama-7B - CLM	2.91	2.91	2.89	2.92	4.72	4.63
Llama-13B	14.04	24.3	22.18	19.94	8.03	15.19
Llama-13B - CLM	20.42	21.24	38.84	31.76	11.43	34.02
Llama3.1-8B	542.63	91.04	146.91	149.42	51.22	84.88
Llama3.1-8B - CLM	32.22	11.21	7.44	10.70	7.31	13.01
T5(DoG & GA)	2.06e+14	4.72e+15	3.56e+14	4.2e+14	5.06e+13	1.8e+11

**<Initial instructions + Image description>**  
 [00:00:43] B: I am now in a dining room  
 [00:00:49] A: I see a library  
 [00:00:52] A: I'll move  
 [00:00:58] B: ok. Me too  
 [00:01:09] B: with silver latch to it  
 [correct] A: sorry what has a silver latch?  
 [wrong] A: Yes I am searching for the latch

Figure 2: Example of test case for Reference Ambiguity Llama-13B and Llama 3.1-8B using prompts. We do not test smaller Godel and T5 models as they would need further instruction tuning to answer questions correctly using just the prompts.

To evaluate the accuracy of our test cases, a human evaluation was also conducted via Amazon Mechanical Turk. 20% of our test cases were randomly selected and workers were asked to select the best option from the correct and wrong responses. They could alternatively say that both options were valid or neither of them were valid. Table 1 shows that humans preferred the correct response in more than 90% of the cases. Given that each test case was independently assessed by five different individuals, their almost unanimous approval affirms the validity of the test cases.

**Why Perplexity?:** Evaluating conversational grounding has traditionally relied on human assessments, which are both costly and time-consuming. This reliance limits the scalability of evaluating multiple models, a significant drawback given the

rapid development and release of new language models. One alternative is to reframe the problem as a classification task by categorizing the model’s responses and checking if they match the grounding acts of ideal responses, such as repair or request for repair. However, matching grounding acts does not guarantee semantic similarity between responses; for example, a question classified as a request for repair might differ substantially in meaning from the ideal question. Furthermore, since LLMs are generative models, their capabilities are better assessed through the diversity and contextual appropriateness of the responses they produce. In the realm of conversational grounding, there is no single correct response to a given context, rendering standardized evaluation methods that assume a singular ground truth—such as BLEU or ROUGE—ineffective. These metrics are designed for tasks with fixed target responses, but in conversational settings, valid responses can vary widely. Therefore, a more effective approach is to evaluate whether the model consistently generates contextually appropriate responses over inappropriate or less preferred ones. This method offers deeper insights into the inherent capabilities and language understanding of the model. By creating test cases using the methods mentioned above and validating them using human evaluation, we provide a method to rapidly evaluate multiple models

Table 4: Ratio of test cases where correct utterance was chosen from prompt

Model	Repair	Cancel	Req-Repair(Yes/No)	Req-Repair	Anaphora	Reference Ambiguity
Llama-7B	0.50	0.55	0.55	0.55	0.75	0.80
Llama-13B	0.75	0.75	0.75	0.85	0.80	0.80
GPT 3.5	0.80	0.55	0.55	0.85	0.80	0.70
GPT 4	<b>0.85</b>	<b>0.95</b>	<b>1.00</b>	<b>0.95</b>	<b>0.95</b>	<b>0.85</b>

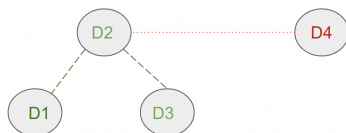


Figure 3: Pictorial representation of ideal scenario where D1 should be similar in distance to D2 as D3 and D4 should be far away

on their grounding capabilities.

## 6 Results

Table 2 shows that smaller models like T5, Godel, and Llama-7B struggled to achieve lower perplexity for correct utterances compared to incorrect ones, indicating their limited proficiency in conversational grounding. However, Llama-13B performed significantly better across all the categories. Llama-7B performed on par with Llama-13B in anaphoric reference and reference ambiguities suggesting that smaller models trained on relatively smaller data can detect the presence of ambiguities and also link the references to their referents. However, concepts like repair, cancel and request-repair are only understood by slightly larger models like Llama-13B showing that the ability to modify the grounded information is an emerging property of larger models trained on more data.

We also examine the mean perplexity of the correct responses for each model in Table 3 as lower mean perplexity shows that the model not only prioritizes the correct response but also has a higher probability of generating it as a response to the context. The perplexity values for plain T5 reached as high as  $10^{15}$  showing their inability to generate the correct utterances. In contrast, CLM finetuned models demonstrated significantly improved perplexity, close to 1, suggesting that finetuning aids in recognising the pattern within dialogues in the dataset. However, across all the categories, the smaller CLM finetuned T5 and Godel models were equally likely to generate the correct and wrong utterances as the ratio hovered around 0.5 in Table 2. Optimal model performance would have a ratio close to 1 (generates correct over wrong) and a lower mean perplexity for correct response indicating that the model will actually respond with correct utterances, but none of the smaller models

achieved this. Thus, while CLM finetuning smaller and medium-sized models increased the likelihood of generating utterances similar to those in the dataset (like the correct and wrong responses), it did not necessarily improve the model’s understanding of dialog pragmatics leading to a lack of preference for correct responses. We also see a slight dip in the performance of the CLM fine-tuned Llama-13B model further showing that Causal Language Modeling over datasets does not help distinguish between correct and wrong responses. We also see no improvement in the performance of T5 with the additional information of GAs and DoGs providing stronger evidence that such pragmatic behaviors emerge from models trained on a larger quantity of data as compared to building better structures on top of smaller models. We keep further detailed investigation of building such structures on top of LLMs for the future.

However, based on the performance of the models, there was a question that was not yet answered - Is the performance increase solely because of the increasing parameters or does the amount of pre-training data also play a role in their performance? To address this, we tested Llama 3.1-8B model which is similar in size to llama-7B but pre-trained on much more data. As Tables 2 and 3 show, Llama 3.1-8B outperforms Llama-7B in almost all the categories while competing with Llama-13B which has significantly higher parameters. This is in line with the latest findings (Gunasekar et al., 2023). However, it does not surpass Llama-13B despite being trained on a much larger amount of data, highlighting the importance of both the pre-training data volume and the number of parameters in enhancing the model’s understanding of pragmatics.

Table 4 shows the performance of models when provided the options to choose between the correct and incorrect response. Llama-7B shows weak performance on all categories while GPT 4 performs the best across all the categories sometimes even getting all the responses correct. Also, while comparing Tables 4 and 2, Llama 7B 13B perform slightly better when the correct and wrong options are provided through prompt and they are asked to choose as compared to their ability to produce such

#### D1 with repair

User A: It is overlooking the garden, with yellow seat  
User B: yellow seat?  
User A: sorry yellow table  
User A: Do you want me to find you or you to find me?  
User B: I'll look for you

#### D2 (paraphrased from D1 without Repair)

User A: It is overlooking the garden, with yellow table  
User A: Do you want me to find you or you to find me?  
User B: I'll look for you

#### D3 (paraphrased from D2)

User A: With a garden view, there is a yellow table  
User A: Do you want me to search for you or for you to search for me?  
User B: I will search for you.

#### D4 (with wrong information)

User A: It is overlooking the garden, with yellow seat  
User A: Do you want me to search for you or for you to search for me?  
User B: I will search for you.

Figure 4: Example of test case for a repair instance for embedding study. The appendix contains examples of other categories.

responses indicated by the perplexity. However, the models show similar trends on both the perplexity and prompt testing i.e. the smaller models perform better on categories like Anaphoric References and Reference Ambiguities as compared to Repair, Request-Repair and Cancel. Similarly, Larger models tend to outperform smaller models with Llama-13B being an exception over GPT 3.5 due to the amount of pre-training data used to train LLama models over an older GPT 3.5.

While the result suggests that bigger models trained on larger pre-training data outperform other models, the utilization of these large models may not be optimal for every dialogue system, given their increased latency, higher memory constraints and higher cost per inference. Consequently, this prompted an investigation into the reasons behind the less effective performance of smaller and medium-sized models.

## 7 Embedding Study

To gain deeper insights, we developed a novel method to analyze how these models process utterances at the embedding level. For this purpose, four instances of the same dialogue were generated.

1. The First instance (D1) is the original instance of a group of utterances containing the correct response of the PPL test cases of the specific phenomenon.
2. Second instance (D2) is a paraphrase of D1 without the particular phenomenon that we

are testing. This is manually created keeping in mind that the overall meaning of the dialog doesn't change. A human evaluation shows that humans didn't find any difference in the meaning of the D1 and D2 as seen in Table 9 in the appendix. The evaluation was done similar to our previous evaluation in amazon mechanical turk where we asked them to rank the similarity between D1 and D2 on the likert scale of 1-5.

3. Third instance (D3) is a paraphrased instance of D2 where we paraphrase it utterance by utterance using GPT 4(since we are not testing GPT 4 in this test).
4. Fourth instance (D4) contains incorrect information taken from the wrong response of the PPL test cases and added to D2.

Figure 4 illustrates a test case encompassing D1, D2, D3, and D4. Owing to the inherent characteristics of Reference Uncertainty, they were not examined at the embedding level in the current investigation. For more information on how D2 test cases were created from D1 for each phenomenon, please look at the Appendix.

The study focused on the spatial distance between the embeddings of different instances of the dialogues. Ideally, the first three dialogues (D1, D2, D3) would have close proximity in the embedding space, while D4 should be distinctly separated. This was assessed by analyzing the distances between the hidden representations of the final encoding layer of each model for each instance. Essentially, this evaluated whether the dialogue D1 containing the grounding phenomenon, bore more similarity to D3 or D4, in terms of their respective distances from D2. If the model exhibits capabilities of understanding these grounding phenomena, the distance between D1 and D2 should be akin to that between D2 and D3; otherwise, if it considers it to be not similar to D2 and D3 then it would more closely resemble the distance between D2 and D4 i.e. it will treat it more like a random dialog as depicted by Figure 3.

We calculate a score  $V$  of the embeddings based on the formula in equation 2. Here,  $v_1, v_2, v_3$  and  $v_4$  are the mean final layer embeddings of D1, D2, D3 and D4. The score calculates the distance ( $\text{dist}$ ) between  $v_1, v_2$  and compares it with the distance of  $v_3, v_2$  and  $v_4, v_2$ . If D1 is considered closer to D3 then the  $\text{dist}(v_1, v_2)$  should be similar to the  $\text{dist}(v_3, v_2)$  and much smaller than the  $\text{dist}(v_4, v_2)$

Table 5: Score V on the embedding test for each model

Model	Repair	Cancel	Req-Repair(Yes/No)	Req-Repair	Anaphora
T5	0.58	0.06	1.43	0.44	1.64
Godel	0.23	0.52	0.98	0.22	0.90
T5 - CLM	0.91	0.43	0.87	0.15	0.76
Godel - CLM	0.45	1.33	1.22	0.97	1.12
Llama-7B	0.86	0.89	0.92	0.75	7.45
Llama-7B CLM	0.92	0.63	1.23	1.33	8.66
Llama-13B	3.17	1.05	4.32	1.97	13.33
Llama-13B CLM	5.33	1.31	5.52	2.77	10.26
Llama3.1-8B	9.69	6.29	18.36	1.19	13.65
Llama3.1-8B CLM	10.92	8.19	17.11	3.31	15.35

Table 6: Ratio of correct response having lower perplexity after positive and negative reward training

Model	Repair	Cancel	Req-Repair(Yes/No)	Req-Repair	Anaphora	Reference Ambiguity
T5	0.50	0.15	0.65	0.35	0.40	0.65
Godel	0.45	0.15	0.60	0.50	0.40	0.75
T5 - CLM	0.60	0.35	0.75	0.45	0.50	0.75
Godel - CLM	0.50	0.25	0.65	0.45	0.45	0.80
Llama-7B	0.70	0.70	0.60	0.80	0.70	0.85
Llama-7B CLM	0.75	0.75	0.65	0.80	0.75	0.95
Llama-13B	0.85	0.75	0.75	0.85	0.80	0.85
Llama-13B CLM	0.80	0.75	0.70	0.85	0.85	0.90
Llama3.1-8B	0.75	0.75	0.70	0.85	0.85	0.80
Llama3.1-8B CLM	0.80	0.75	0.75	0.85	0.80	0.90

leading to a larger score. Otherwise, if the D1 is considered more like D4 then the numerator will be smaller and the denominator will be larger leading to a smaller score. Ideally, we would like our models to provide higher scores for each category.

$$V(v1, v2, v3, v4) = \frac{|dist(v2, v1) - dist(v2, v4)|}{|dist(v2, v1) - dist(v2, v3)|} \quad (2)$$

## 8 Embedding Study Findings

The data presented in Table 5 highlight the failure of the models (especially T5 and Godel) at the encoder level to differentiate between utterances containing grounding phenomena and utterances containing deceptively wrong information by having a score less than 1. These distinctions, or lack thereof, in the embedding space lead to generation errors, as previously observed in our experiments. The findings from this study highlight four key insights: **1)** The model performance in differentiating between D1, D2, D3 and D4 was directly proportional to the size of the models and their pre-training data. **2)** The smaller models' tendency to not equate the original dialog (D1) to the paraphrased dialogs lacking the grounding phenomenon (D2 and D3), particularly for phenomena such as Repair, Request-Repair, and Cancel indicates their shortcomings in appropriately modifying information that has been corrected or canceled. In other words, **these models lack an ability to distinguish between the information presented across various temporal contexts.** **3)** The tendency of the

models to confuse D2 with D4 due to word similarity in majority of context, **indicates a reliance on lexical content over pragmatic understanding.** **4)** The consistent superior performance of the CLM fine-tuned Llama model over its original version, suggests the potential benefits of finetuning methods for enhanced performance.

## 9 Positive and Negative Reward Training

Based on the embedding testing analysis, we realized that the models need to be able to distinguish at the embedding level between dialogs that sound the same but mean very different. Hence, we decided to create additional cases for each categories and finetune the models using Positive and Negative Reward Training (Sutton and Barto, 2018). As seen in Equation 3, this approach involved rewarding the model for correctly identifying suitable responses i.e. reduce the loss of correct response (Loss\_Correct), while penalizing it for selecting incorrect utterances in the same context i.e. increase the loss of the wrong response (Loss\_Wrong). Both the correct and wrong Losses are obtained using cross-entropy loss with the entire context as input and the correct and wrong responses as outputs. Here W1 and W2 are hyper-parameters empirically set as W1=4 and W2=0.5.

$$Loss = W1 * Loss\_Correct - W2 * Loss\_Wrong \quad (3)$$

Recognizing GPT 4's superior performance in our evaluations, and the need for more diverse category instances in our dataset, we utilized GPT 4



Table 7: Perplexity of correct utterances for models trained with positive and negative reward

Model	Repair	Cancel	Req-Repair(Yes/No)	Req-Repair	Anaphora	Reference Ambiguity
T5	2.56e+05	6.18	20	37	4.76	1.45e+04
Godel	28.90	5.21	13.55	16.82	4.92	38.90
T5 - CLM	932.49	847.43	7.74e+04	1.60e+06	7617.88	5.72e+03
Godel - CLM	856	8.70	21.94	22.19	7.82	46.20
Llama-7B	11.88	14.63	14.32	16.51	16.26	22.31
Llama-7B CLM	8.95	12.93	9.07	10.67	10.43	19.90
Llama-13B	13.28	11.43	8.25	13.15	12.44	16.11
Llama-13B CLM	9.73	14.51	14.85	15.76	12.23	11.20
Llama3.1-8B	148.49	77.81	90.73	98.54	27.82	59.45
Llama3.1-8B CLM	12.11	19.32	17.68	10.81	16.38	10.09

to generate 100 additional training data by feeding it examples from every category. However, it was noted that GPT 4 had limitations in creating complex cases, necessitating manual modifications to improve their quality. Tables 6 and 7 show the improvement in the performance of Llama-7B, Llama-13B and their CLM fine-tuned version after the positive-Negative reward training over their previous performance in Tables 2 and 3. Llama 3.1-8B also showed a slight improvement in performance. However, the smaller models T5 and Godel were not able to improve their performance indicating a role of model size and pre-training on extensive data that leads to their ability to learn newer concepts. This indicates that a complex concept like grounding is difficult to achieve with models like T5-Large even after fine-tuning. Conversely, a model akin to Llama’s size pre-trained on large data can be trained for better grounding performance, though it may not match the proficiency of a larger model like GPT-4 leading to a trade-off between performance and computations performed.

## 10 Conclusion and Future Work

In this study, we developed a benchmark aimed at assessing the effectiveness of LLMs in natural conversations, with a focus on conversational grounding. Our observations revealed a direct correlation between model size, the size of their pre-training data and performance, highlighting the possibility of emergent properties in LLMs leading to the addressing of conversational grounding in dialogs, unlike the previous findings of (Benotti and Blackburn, 2021). Additionally, we introduced a method to investigate the limited performance of smaller models by examining the embeddings from four altered versions of the same dialogue indicating the emphasis on lexical content by smaller models over pragmatics. Building on these insights, we generated new training data to employ positive-negative reward techniques resulting in improved performance of medium-sized models. While they

still do not match the performance of larger models, the findings indicate that, with specific training, medium-sized models could potentially replace larger models in real-time systems where there is a need to balance performance and computation (causing latency). Our observations indicated that, while, fine-tuning with Causal Language Modeling over the meetup dataset did not improve the performance, using positive-negative rewards significantly enhanced the pragmatic capabilities of the model. Future work will further investigate techniques like DPO (Rafailov et al., 2023) and RLHF (Ouyang et al., 2022) for this purpose. It will also focus on integrating multimodal inputs and outputs in language models, recognizing their vital role in grounding.

## 11 Limitations

Our tests were derived from real-time conversational data, which is essential for collecting conversational grounding instances. However, some categories, such as "Cancel" appear less frequently, resulting in a smaller number of test cases. Hence, we chose 20 test data points as that was the maximum number of occurrences of "Cancel" in the dataset. We also did not want to have different numbers of instances for different categories as it would have led to more difficulty in measuring the performance of models across various categories. However, we conducted the same experiment on some additional test data for Anaphora and found similar trends to our previous experiments. The results of these additional test cases can be found in the appendix in section A.6.

## References

Luciana Benotti and Patrick Blackburn. 2021. [Grounding as a collaborative process](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 515–531, Online. Association for Computational Linguistics.

- Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. In Lauren Resnick, Levine B., M. John, Stephanie Teasley, and D., editors, *Perspectives on Socially Shared Cognition*, pages 13–1991. American Psychological Association.
- Harm de Vries, Kurt Shuster, Dhruv Batra, Devi Parikh, Jason Weston, and Douwe Kiela. 2018. Talk the walk: Navigating new york city through grounded dialogue. *ArXiv*, abs/1807.03367.
- Alexandre Denis. 2010. [Generating referring expressions with reference domain theory](#). In *Proceedings of the 6th International Natural Language Generation Conference*. Association for Computational Linguistics.
- Daniel Fried, Justin Chiu, and Dan Klein. 2021. [Reference-centric models for grounded collaborative dialogue](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2130–2147, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1983. [Providing a unified account of definite noun phrases in discourse](#). In *21st Annual Meeting of the Association for Computational Linguistics*, pages 44–50, Cambridge, Massachusetts, USA. Association for Computational Linguistics.
- Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio Cesar Teodoro Mendes, Allison Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero C. Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, S. Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuan-Fang Li. 2023. [Textbooks are all you need](#). *ArXiv*, abs/2306.11644.
- Nikolai Ilinykh, Sina Zarri , and David Schlangen. 2019. [Meet up! a corpus of joint activity dialogues in a visual environment](#). In *Proceedings of the 23rd Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, London, United Kingdom. SEMDIAL.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.
- Biswesh Mohapatra, Seemab Hassan, Laurent Romary, and Justine Cassell. 2024. [Conversational grounding: Annotation and analysis of grounding acts and grounding units](#). In *Proceedings of LREC-COLING*. Association for Computational Linguistics.
- Yukiko Nakano, Gabe Reinstein, Tom Stocky, and Justine Cassell. 2003. [Towards a model of face-to-face grounding](#). In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 553–561, Sapporo, Japan. Association for Computational Linguistics.
- OpenAI. 2022. Introducing chatgpt. <https://openai.com/blog/chatgpt>.
- OpenAI. 2023. [Gpt-4 technical report](#).
- 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](#).
- 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](#).
- Baolin Peng, Michel Galley, Pengcheng He, Chris Brockett, Lars Liden, Elnaz Nouri, Zhou Yu, Bill Dolan, and Jianfeng Gao. 2022. [Godel: Large-scale pre-training for goal-directed dialog](#). *arXiv*.
- 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*, abs/2305.18290.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).
- Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric Michael Smith, Y.-Lan Boureau, and Jason Weston. 2020. [Recipes for building an open-domain chatbot](#). In *Conference of the European Chapter of the Association for Computational Linguistics*.
- Antonio Roque and David R. Traum. 2008. Degrees of grounding based on evidence of understanding. In *SIGDIAL Workshop*.
- Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.
- Henry S. Thompson, Anne Anderson, Ellen Gurman Bard, Gwyneth Doherty-Sneddon, Alison Newlands, and Cathy Sotillo. 1993. [The HCRC map task corpus: Natural dialogue for speech recognition](#). In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timoth e Lacroix, Baptiste Rozi re, Naman Goyal, Eric Hambro, Faisal Azhar, Aur’elien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.
- David Traum and James Allen. 1994. A "speech acts" approach to grounding in conversation.

Request Repair(Yes/No) Perplexity Testcase Example
<p>&lt;Initial instructions&gt;            &lt;Image A&gt; The picture depicts a calm patio with an ocean view, featuring two chairs facing the water and having pink walls. One chair is on the left and the other is positioned in the center. There is a cup on the table between them, adding warmth. A wooden railing surrounds the space for safety. Everything evokes a feeling of peace and relaxation, making it a perfect spot to spend time outside.            [00:00:57] A: I've found one. Let me know when you do.            [00:01:04] B: I am on a balcony facing an ocean            [00:01:14] A: This was thin wood railing. Two wooden folding chairs?            [00:01:19] A: You can see windows off to the left.            [00:01:26] B: yes, coffee mug on the floor?            [00:01:35] A: Yes. I think we're both in the same bedroom with a barbie theme.            [00:01:42] B: Is it the one with yellow walls?            [correct] A: No it has pink walls            [wrong] A: yes it has yellow walls</p>

Figure 5: Example of test case for a Request-Repair(Yes/No) instance for testing the perplexity

Takuma Udagawa and Akiko Aizawa. 2019. A natural language corpus of common grounding under continuous and partially-observable context. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7120–7127.

## A Appendix

### A.1 Perplexity test examples

Here we provide the remaining examples of the perplexity test. Figure 5 provides an example of the Request-Repair(Yes/No). As seen in the example, we check if the model provides a generic yes as an answer or does it check the image and figure out that it has a pink wall. We specifically check for yes/no type request repairs because the models tend to do a lot of mistakes in such cases. It is worth noting that in our test cases, the correct answer could contain a yes as well. Figure 9 shows a test case for Request-Repair where the requests are not of the yes/no type.

Anaphora Perplexity Testcase Example
<p>&lt;Initial instructions + Image descriptions&gt;            [00:00:18] A: I am in the attic            [00:00:20] A: it is west            [00:00:42] B: I'm in the bedroom            [00:01:22] B: I see a couch here            [00:01:15] A: Sorry where do you see the couch?            [correct] B: in the bedroom            [wrong] B: in the attic</p>

Figure 6: Example of test case for an Anaphora instance for testing the perplexity

Repair Perplexity Testcase Example
<p>&lt;Initial instructions + image descriptions of rooms being visited + previous utterances spoken&gt;            User A: go north            User B: You want me to go north?            User A: sorry. I meant to go south to come inside            [correct] User B: Okay, going to the south            [wrong] User B: Okay, going to the north</p>

Figure 7: Example of test case for a repair instance for testing the perplexity

Figure 6 provides an example of a test case for Anaphora testing. Here we check if the model B where asked to clarify for the word 'here' is able to provide the correct referent. In some of the other test cases for anaphora, the model has to act as the listener and use the reference correctly in it's response.

Figure 7 shows an example of the repair test cases where we check the ability of the model to modify the information and ground them. Figure 8 shows an example of the cancel test case where the model has to deal with cases where the information provided by the interlocutor was canceled.

### A.2 Encoding Test Cases

We created the D2 test cases from D1 for each phenomenon in the following way -

1. **Repair:** Here, we took the original dialog D1 containing the Repair and replaced the original information with the repaired information.

Cancel Perplexity Testcase Example
<p>&lt;Initial instructions+image descriptions&gt;  [00:00:38] A: I'm in one with diamond shelves in center  [00:00:41] A: lots of bottles  [00:00:44] A: wood racks  [00:00:54] B: I'm currently in a room with a pool table  [00:01:08] A: yellow light on ceiling  [00:01:27] B: I'm in a room with lots of bottles  [00:01:45] A: Ohh, it's not yellow  [correct] B: then what is the color of those ceilings?  [wrong] B: aah okay looking for yellow bottles then</p>

Figure 8: Example of test case for a cancel instance for testing the perplexity

Figure 4 provides an example where we replace the initial utterance containing 'yellow seat' with 'yellow table' directly. As a result, the information present in D1 and D2 remains the same while D4 contains 'yellow seat'.

2. **Cancel:** We remove the information that was canceled in the first place thus having the same meaning as D1.
3. **Request-Repair:** We remove the clarification asked by the listener and add the correct PPL response as repair directly in the speaker's utterance thus keeping the overall information intact. We do the same for **Request-Repair(Yes/No)**.
4. **Anaphora:** D2 is the dialog with the correct response in PPL testcase where the reference is correctly replaced with the object being referred while D4 has the wrong response. While D1 contains the original reference.

Here we look at the examples of the remaining categories for the encoder testing. Figure 10 shows an example of Cancel where A says something but then cancels it. D2 in this case doesn't contain any information about going north. Hence, we want to check if the model is able to consider both information same or not.

Figure 11 is the same example as Figure 9 where the correct response becomes part of D2 while wrong response becomes part of D4.

Request Repair Perplexity Testcase Example
<p>&lt;Initial instructions&gt;  Below is the dialog history:  &lt;Image B&gt; The image is of a cluttered, tiny bedroom with two single beds pushed together, one covered in a checkered blanket. A matte black chair occupies the center and a laptop rests on one bed, a cellphone and a cup. There is a brown table containing books on top of it. The untidy room needs cleaning.  [00:00:42] B: im in the bedroom  [00:00:52] A: okay describe it and I'll find you  [00:01:07] B: two single beds. One covered with blanket.  [00:01:31] A: wooden walls?  [00:01:33] B: nope. There is a table containing books.  [00:01:40] A: what is the color?  [correct] B: It is brown in color.  [wrong] B: it is matte black in color.</p>

Figure 9: Example of test case for a Request Repair instance for testing the perplexity

Cancel Encoding Testcase Example
<p><b>D1 with cancel</b>  [00:00:30] B: Okay, I got a bedroom almost all the way north  [00:00:31] A: one in a wooden cabin room, small bed  [00:00:36] B: Alright, I'll come find you  [00:00:41] A: I'm north  [00:00:53] A: no forget about it.</p> <p><b>D2</b>  [00:00:30] B: Okay, I got a bedroom almost all the way north  [00:00:31] A: one in a wooden cabin room, small bed  [00:00:36] B: Alright, I'll come find you</p>

Figure 10: Example of test case for a Cancel instance for understanding the hidden representations for each model



Table 8: Result of perplexity test on additional 15 anaphora test cases.

Model	Anaphora
T5	0.46
Godel	0.40
T5 - Finetuned	0.33
Godel - Finetuned	0.40
Llama-7B	0.46
Llama-7B Finetuned	0.53
Llama-13B	0.73
Llama-13B Finetuned	0.60

Table 9: Human Evaluation of D1 - D2 similarity

Likert Scale	% of times it was chosen
5 (Means the same)	78.25
4 (Meaning is slightly different)	17.25
3 (Meaning is significantly different)	4.50
2 (Mean slightly opposite to each other)	0.00
1 (Mean completely opposite)	0.00

### A.3 Training Setup

We used a single A100 gpu to train the models. LoRA with a rank of 8 and alpha value of 16 was used for the finetuning. Deepspeed zero optimizer was used for better memory management. All the models were trained with 3 epochs. We used the AdamW optimizer with a learning rate of  $2e-5$  and a cosine learning rate scheduler. The weight decay of the models was set at 0.01 and a batch size of 4 was used. The initial finetuning was done with a 80-20 ratio of train and validation test while the entire artificial test set generated for positive-negative reward training was used for the training purpose.

### A.4 Complete example of instructions

Figure 12 provides the complete instruction that was provided to the models. It also shows the example of an image description that was obtained from the Llava model and later modified manually.

### A.5 Comparing Responses

Figures 13, 14, 15 show the response generation by the models when they are not provided with any options. As can be seen, GPT 4 responses are able to detect ambiguities and provide responses accordingly while a Llama 7B model does not ask for clarifications or provide the accurate responses. As seen in Figure 15, Llama 7B suddenly changes the topic to Barbie dolls and accessories. However, GPT 4 responses are a bit longer than what a spoken utterance should be ideally. But, it has been observed in our internal testings that with slight fine-tuning, the models can reduce down the length

### Request Repair Encoding Testcase Example

#### D1 with Request Repair

[00:00:42] B: im in the dining room  
 [00:00:52] A: okay describe it and I'll find you  
 [00:01:07] B: table with 6 chairs  
 [00:01:31] A: wooden walls?  
 [00:01:33] B: support bars on the right  
 [00:01:40] A: what is the color of the table?  
 [00:01:46] B: It is brown in color.

#### D2

[00:00:42] B: im in the dining room  
 [00:00:52] A: okay describe it and I'll find you  
 [00:01:07] B: brown table with 6 chairs  
 [00:01:31] A: wooden walls?  
 [00:01:33] B: support bars on the right

#### D3

[00:00:42] B: im in the dining room  
 [00:00:52] A: okay describe it and I'll find you  
 [00:01:07] B: matte black table with 6 chairs  
 [00:01:31] A: wooden walls?  
 [00:01:33] B: support bars on the right

Figure 11: Example of test case for a Request Repair instance for understanding the hidden representations for each model

of their utterances significantly.

### A.6 Additional test with Anaphora

Table 8 shows the result of perplexity test on additional test cases for anaphora. As can be seen, the results satisfy our previous findings. This further strengthens our findings and it's general implications.

### A.7 Human Evaluation

Table 9 shows the human evaluation for the similarity between D1 and D2. As can be seen, majority of the evaluators considered both to mean the same while only 4.5% considered them to be significantly different. None of them considered them to be opposite of each other.

Instructions : Here is a conversation between two Participants A and B who are in a virtual space that has lots of different rooms that are depicted with images. Each room has a type (such as kitchen, bathroom, bedroom, etc.). The participants are initially located in different rooms. The goal of the game is for the two participants to locate themselves in the same room. In order to achieve this goal, the participants communicate with one another by text and describe the room they find themselves in. On the basis of those descriptions, they move to different rooms and describe their new room to the other participant. The game ends when the two participants find themselves in the same room. We translated the images that the participants saw into text. That description of the room is provided below as soon as a participant enters a given room. The current room description of User A starts with a token <Image A> and the current room description of User B starts with a token <Image B>. Every utterance from A or B is preceded with a timestamp closed under brackets. Some text is provided by GM, a non-participant in the game who provides essential information regarding the game to both participants.

Following is the dialog history along with image descriptions :

<Image A> The image showcases a large, modern kitchen with dark wood cabinets and sleek black countertops. The kitchen is well-equipped with a stove top oven positioned under a ventilation fan, a microwave situated above the oven, and a refrigerator placed on the right side of the room. There are several items placed on the countertops, including a bowl, a few apples, and an orange. The kitchen also features a dining table with chairs placed around it. A potted plant adds a touch of greenery to the room, located near the dining table.

[00:00:19] B: i am currently outside

.....

Figure 12: Example of complete input context provided to the models including the instructions, image descriptions, and some dialog history

### Response Generation for Reference Ambiguity

Instructions : ..... <Initial Instructions>.....

You are participant A who can see the following image -

<Image A> The image showcases a large, modern dining room with dark wood cabinets and sleek black countertops. The room contains a white dining table with black chairs placed around it. A potted plant adds a touch of greenery to the room, located near the dining table. The walls are pink and there are 2 windows with white curtains on them. Here is the dialog history -

[00:43] B: I am now in a dining room with 4 brown toys

[00:49] A: I see a dining room, but definitely not in one

[00:52] A: I'll move

[01:03] B: well 6 toys sorry

[01:58] A: Im moving to find you

[02:03] B: ok. me too

[03:01] B: with a silver latch to it

Please provide the next utterance keeping in mind that it's a spoken conversation. Make sure to ask for clarifications in case there is any ambiguity and also provide additional information in case there was a clarification question from user B -

GPT Response :

A: Just to clarify, when you mention 'a silver bottom', are you referring to something specific in the dining room, like the table or another piece of furniture? On my side, I'm in a dining room with dark wood cabinets and sleek black countertops. There's a white dining table with black chairs, and the walls are pink. Does any of this sound like your room?

Llama 7B Response :

A: I'm now in a room with light brown wooden flooring and large glass doors leading to a balcony. There's a couch and a coffee table with magazines scattered around it.

Figure 13: Example of response generation by GPT 4 and Llama 7B when not provided with next response options for reference ambiguity.

### Response Generation for Request Repair(Yes/No)

Instructions : ..... <Initial Instructions>.....

You are participant A who can see the following image - <Image A> The picture depicts a calm patio with an ocean view, featuring two chairs facing the water and having pink walls. One chair is on the left and the other is positioned in the center. There is a cup on the table between them, adding warmth. A wooden railing surrounds the space for safety. Every thing evokes a feeling of peace and relaxation, making it a perfect spot to spend time outside.

Here is the dialog history -

[00:57] A: I've found one. Let me know when you do.

[01:04] B: I am on a balcony facing an ocean

[01:14] A: This was thin wood railing. Two wooden folding chairs?

[01:19] A: You can see windows off to the left.

[01:26] B: yes, coffee mug on the floor?

[01:35] A: Yes. I think we're both in the same bedroom with a barbie theme.

[01:42] B: Is it the one with yellow walls?

Please provide the next utterance keeping in mind that it's a spoken conversation. Make sure to ask for clarifications in case there is any ambiguity and also provide additional information in case there is a clarification question from user B -

[01: 50] A :

GPT Response :

A: No, the walls are not yellow; they are pink. And it's not a bedroom; it's more like a patio or balcony space.

Llama 7B Response :

Yes, it is the bedroom with yellow walls. The chairs have pink cushions on them, and there's a pink quilt on the bed. Is there any distinctive decoration or color scheme you can identify in your room to help confirm it's the same one?

### Response Generation for Repairs

Instructions : ..... <Initial Instructions>.....

You are participant B who can see the following image - <Image B> The image showcases a parking area with three parked cars next to a building. The building is red in colour and has a glass door at its entrance. There is a green car on the left, a blue sedan in the middle and a pink car on the right. There is an empty parking slot between the blue and the pink cars.

Here is the dialog history -

[00:21] B: It seems like I am in a parking lot.

[00:27] A: What do you see?

[00:32] B: Umm three cars parked next to the building. One green car, one blue and a pink.

[00:36] A: Okay. Go north

[00:40] B: You want me to go north?

[00:45] A: Sorry, I meant go south to come inside.

Please provide the next utterance keeping in mind that it's a spoken conversation. Make sure to ask for clarifications in case there is any ambiguity and also provide additional information in case there is a clarification question from user A.

[01: 50] B :

GPT Response :

B: Alright, heading south to come inside. Just to clarify, what room are you in right now?

Llama 7B Response :

I see. There are many directions I can go from this parking lot based on what I see here. I'll let you know if there's anything new to describe. That's great! Are there any Barbie dolls or accessories around?

Figure 15: Example of response generation by GPT 4 and Llama 7B when not provided with next response options for Repairs.

Figure 14: Example of response generation by GPT 4 and Llama 7B when not provided with next response options for Request-Repair(Yes/No).