Frame of Reference: Addressing the Challenges of Common Ground Representation in Situational Dialogues

Anonymous ACL submission

Abstract

Common ground plays a critical role in situated spoken dialogues, where interlocutors must establish and maintain shared references to entities, events, and relations to sustain coherent interaction. For dialog systems, the ability to correctly ground conversational content in order to refer back to it later is particularly crucial in situated contexts. Prior studies have demonstrated that LLMs are capable of performing grounding acts such as requesting clarification or producing acknowledgments, yet relatively little work has investigated how common ground can be explicitly represented and stored for later use. Without such mechanisms, it remains unclear whether acknowledgment or clarification behaviors truly reflect a grounded understanding. In this work, we introduce an evaluation benchmark `IndiRef` to assess a model's ability to establish and exploit common ground through relational references to entities within the shared context in a situational dialogue. We test multiple baselines for representing common ground in situated dialogues and further propose approaches to improve both the establishment of common ground and its subsequent use in dialogue.

1 Introduction

In dialogue, common ground refers to the shared knowledge, beliefs, and assumptions that accumulate between participants as a conversation progresses (Clark and Schaefer, 1989). This shared understanding renders conversations more efficient and more effective. The process of building and maintaining this shared understanding is known as conversational grounding. Its importance is particularly pronounced in the domain of human-agent interaction, and especially for Embodied Conversational Agents (ECAs) (Cassell et al., 2001) and Robots (Chen et al., 2025) that operate in dynamic and shared environments. To become truly effective collaborators, these agents must move beyond simple command-response behaviors and engage in situational dialogues, demonstrating an awareness of their environment and the ongoing interaction.

Figure 1: An example of people referring back to the common ground.

Recent advancements in Large Language Models (LLMs) have shown promise in performing what has been called grounding acts (Traum and Allen, 1994) within local context, like asking for clarifications, understanding modifications of previous information, and providing acknowledgment (Khebour et al., 2024; Mohapatra et al., 2024b). However, a significant research gap remains in how to effectively represent and store established common ground for later use. This is a critical limitation, as people expect conversation to be a rather continuous process where context is built and referred back to over time.

To develop algorithms capable of building a robust common ground for later use, it is essential to first establish a method for measuring its effectiveness. Inspired by (Kruijt and Vossen, 2022), who demonstrate that people frequently refer to established common ground through relational references to their shared context rather than explicit repetition, we propose that a system's ability to resolve these references (Krahmer and van Deemter, 2012) can serve as a measure of its grounding capabilities over time. The difficulty of this task is significantly amplified in situated dialogues (Seminck, 2023), where users often rely on what we call

"relational references" as entities may be referred to in several different ways in relation to another entity or event. For instance, in Figure 1, the user refers to the phone via the shop where it was bought. Such entities in situated dialogue lack unique references and could equally be described temporally (e.g., first shop) instead of spatially (e.g., next to the church). Such multiple relational references necessitate a sophisticated grasp of context to map multiple linguistic forms to their correct real-world entities. This natural tendency for relational references poses a substantial challenge for current dialog systems.

Beyond representing entities, common ground representation must also incorporate the beliefs and perspectives of its participants. This shifts the focus from simple state tracking (Jacqmin et al., 2022) to belief tracking (Varonina and Kopp, 2021), where the system models the user's belief state, such as their prior knowledge or perspective. Such interplay of perspectives becomes prominent in situated dialogues that often consist of multiple deictic instances (Shopen, 1985). For example, the referent in the phrase "*the second car that I bought*" is different from "*the second car that you bought*".

Building on these challenges, this paper seeks to address the critical gap in representing and effectively utilizing common ground in spoken situated dialogues. In this work, we first analyze two situated dialogue datasets to provide an empirical understanding of how people leverage their common ground during a conversation, particularly through relational references. Informed by this analysis, we introduce a new evaluation benchmark `IndiRef` , designed to evaluate a model's ability to establish and exploit the common ground by resolving various relational references to it. Using this benchmark, we test multiple baseline models and algorithms for representing common ground. We then propose and evaluate novel approaches aimed at improving both the establishment of common ground and its subsequent use. To facilitate the training of these approaches, we also introduce a novel framework to generate synthetic data for situational dialogues. Our findings thus provide a path towards developing dialog systems with more robust conversational grounding capabilities. The `IndiRef` test data, synthetic data generation code, and evaluation code will be open-sourced.

2 Related Work

(Clark and Brennan, 1991) argued that interlocutors iteratively establish what must be sufficiently mutually understood for the current stage of a conversation. (Traum and Allen, 1994) formalized this process by introducing the concepts of Grounding Acts (GAs), that represent fundamental conversational moves like acknowledging, repairing, or requesting clarifications that are used to negotiate shared understanding, and Grounding Units (GUs), that structure the utterances into chunks that commit information to the common ground. (Mohapatra et al., 2024b) showed that modern LLMs can successfully perform these GAs when the required context is limited to the last few utterances, for instance, by appropriately asking for clarification or acknowledging new information. However, this demonstrated ability to perform grounding acts may not guarantee actual availability of the information later in the dialogue. The practical challenges of connecting these local grounding decisions to the long-term maintenance of a coherent common ground remain significant (Mohapatra et al., 2024a), and there has been limited investigation into whether an LLM's ability to engage in these acts leads to effective storage and integration of information for subsequent use, a critical gap our work aims to address.

The computational representation of common ground has been a long-standing goal in dialog systems research. Earlier attempts to represent and manage the common ground, such as Domain Reference theory (Denis, 2010), were limited in nature due to their reliance on rule-based approaches. Foundational work has largely focused on structured, symbolic approaches for task-oriented dialogues. Methods like Dialog State Tracking (DST) (Jacqmin et al., 2022) and frame-based systems (Blache and Houlès, 2021) represent conversational context as slots and values, capturing the user's goals and constraints. While highly interpretable and effective for well-defined tasks (Kim et al., 2023), these approaches often lack the flexibility to handle the dynamic and unpredictable nature of situated interactions.

To overcome these limitations, more flexible representations have been proposed, such as using dynamic knowledge graphs (KGs) (Zhang et al., 2019; Walker et al., 2022; Bu et al., 2025) to model entities and their relationships as they emerge in conversation. KGs allow for a richer, more interconnected model of context and can naturally handle entity-based references by linking mentions to nodes in the graph. However, a significant chal-

Phenomena	Dataset	
	Meetup	Spot the difference
Temporal Reference	1	4
Spatial Reference	34	72
Comparative Reference	0	6
Attributive Reference	33	8
Inferred grounding	177	67

Table 1: Frequency of different reference types

lenge arises in situated dialogues where entities often lack a single, stable referring expression (e.g. 'park next to the restaurant' or 'duck park') or relation (e.g., 'the room that has a TV' and 'TV is next to a table' implies 'room has a table'), making it difficult to determine what each node and relation should represent. Their dynamic nature makes it difficult to provide specific rules for building nodes and relations. Prior work on these references has largely focused on resolving grounding within the immediate conversational frame (Kelleher and Kruijff, 2006; Shridhar et al., 2020; Padmakumar et al., 2022). (Wu et al., 2025) offers a dataset for multi-session dialogues but lacks a focus on situational contexts. While (Hosseini et al., 2023) provide a valuable dataset for indirect references, their focus is on using existing common ground to disambiguate entities, rather than establishing the common ground from the dialogue history itself. (Williams et al., 2019; Madge et al., 2025) address referential entities in situated dialogues but overlook long-term representation and references.

In parallel, contemporary neural memory architectures (Henaff et al., 2016; Rajendran et al., 2019; Wu and Yu, 2024) encode dialogue history in distributed vector representations that can learn to resolve some references through attention mechanisms. Yet they often lack an explicit, persistent representation of common ground. This "black-box" nature makes it difficult to verify whether their conversational behaviors, such as acknowledgments, stem from a truly grounded understanding or pattern matching, especially, when there is an absence of an evaluation metric.

Beyond tracking facts, a robust common ground representation requires modeling participant beliefs and perspectives (Traum and Larsson, 2003; Schlangen and Skantze, 2009), necessitating a shift from state tracking to belief tracking. Frameworks like those proposed by (Di Maro et al., 2021; Khebour et al., 2024; Sclar et al., 2024; Qiu et al., 2024) improve reliability by verifying mutual understanding through the tracking of shared facts, evidence, and unresolved information. (Li et al., 2025) provide a benchmark for theory of mind from egocentric video frames. However, they still fail to address the core representational challenges of situated dialogues: handling entities with ambiguous references and resolving their mentions over time. Indeed, most common ground research has been conducted on the immediate conversational frame as compared to grounding over multiple events.

With the increasing use of LLMs in spoken dialog systems, their finite context window poses a constraint on maintaining long-term conversational memory. A common workaround is dialogue summarization, where models generate concise summaries of past exchanges to be injected into future contexts, enabling a personalized experience (Ramprasad et al., 2024). A more scalable solution involves Retrieval-Augmented Generation (RAG) (Lewis et al., 2020), which offloads the entire dialogue history to external memory. During conversation, relevant excerpts are retrieved and fed into the LLM's context, a technique used to maintain long-term persona memory in conversational agents (Shuster et al., 2021). However, they face similar challenges as other methods in situated dialogues due to various relational references. Consequently, an effective common ground representation must maintain information that is efficiently retrievable, dynamically updatable, and capable of tracking participant perspectives.

3 Spoken Dialog Analysis

To measure grounding beyond superficial acknowledgment patterns, we turned to how humans leverage common ground. Inspired by (Kruijt and Vossen, 2022), who showed that people frequently use relational references to access shared knowledge, we propose that a model's ability to resolve such references can serve as a robust metric for a model's true grounding capabilities, one that can deal with the increasingly important situated dialogues, where entities may not have a single, unambiguous referring expression.

To better understand the mechanisms of relational reference, we analyzed two dialogue datasets previously used in grounding research. The Meetup dataset (Ilinykh et al., 2019) places two participants on a 2D grid and tasks them with finding each other despite only seeing their own locations. Success in this setting depends on describing rooms, maintaining a record of shared knowledge, modeling the partner's perspective, and clarifying ambiguities—all central challenges for grounding. In con-

LLMs	Meetup				Spot the difference			
	Temporal	Spatial	Attributive	Inferred	Temporal	Spatial	Attributive	Inferred
Gemma2 - 2B	0.20 / 0.18	0.18 / 0.16	0.24 / 0.26	0.26 / 0.16	0.18 / 0.12	0.18 / 0.10	0.20 / 0.12	0.16 / 0.10
Llama3.1 - 8B	0.38 / 0.32	0.46 / 0.38	0.46 / **0.44**	0.20 / 0.20	0.40 / **0.38**	**0.30 / 0.32**	**0.36 / 0.36**	0.26 / 0.16
Gemma2 - 9B	0.48 / 0.38	0.50 / 0.46	0.38 / 0.40	0.26 / 0.18	0.34 / 0.20	0.24 / 0.16	0.28 / 0.16	0.20 / 0.10
Gemma2 - 27B	**0.50 / 0.44**	**0.58 / 0.56**	**0.48 / 0.44**	0.28 / 0.26	0.40 / 0.30	0.28 / 0.20	0.32 / 0.26	0.26 / 0.12
Qwen QwQ - 32B	0.38 / 0.32	0.52 / 0.38	0.44 / 0.40	**0.40 / 0.40**	**0.42** / 0.34	0.26 / 0.22	0.26 / 0.20	**0.28 / 0.20**

Table 2: FEM (↑) / LLM-as-Judge (↑) accuracy for baselines where the entire dialog history is used as the context to answer the question. Best results in **bold**.

trast, the Spot the Difference (STD) (Lopes et al., 2018) dataset involves two speakers, each given slightly different images, who must coordinate over an audio channel to identify discrepancies. While Meetup provides a scenario closer to real world situations for building common ground and referring back to it, we also used STD as it provides a different context to help test the robustness of our representation technique.

We manually analyzed each utterance in the dialogue datasets, identifying all references to previously mentioned entities. This analysis yielded a taxonomy of four primary categories: 1) **Temporal**, referring to an entity by its order of appearance (e.g., "*first book*"); 2) **Spatial**, locating an entity in relation to another (e.g., "*bottle on the table*"); 3) **Attributive**, describing an entity by an attribute (is-a: "*yellow house*" or has-a: "*sofa room*"); and 4) **Comparative**, using a relative property (e.g., "*bigger chair*"). Furthermore, we noticed multiple occurances of implicit grounding mechanisms such as the "indirect acknowledgments" (e.g., 'use,' 'move on') identified by (Mohapatra et al., 2024a), the understanding of which is crucial for robust grounding. We termed them **inferred grounding**, where participants ground information that was implied rather than explicitly stated. For instance, when a navigator is asked by their partner whether they see a "*pink towel*" in their bathroom, that navigator also grounds the fact that their partner has visited a bathroom with a pink towel, later referring to it as "*the bathroom with the pink towel that you had seen*". The frequencies found from our analysis are presented in Table 1.

4 `IndiRef` data creation

Our goal here is to extend beyond prior work examining the ability of LLMs to perform basic grounding acts, such as clarifications and acknowledgments, to an evaluation of a model's ability to engage in more sophisticated strategies as described above, and to utilize the common ground in later stages of a conversation. This goal is particularly salient given the finite context size of LLMs, which requires a representation that can be efficiently stored and retrieved over time. Furthermore, a truly robust evaluation must also test a model's ability to differentiate and manage the individual perspectives that constitute this shared understanding.

To address this gap, we introduce `IndiRef` a new benchmark designed to evaluate a model's grounding ability in situational dialogues. The benchmark consists of Question/Answer (Q/A) pairs where each question seeks information about an entity that is itself identified by a relational reference. We focus on three primary reference types based on their conversational frequency in Table 1: **Temporal** (e.g., asking about the food sold in the park you went to *before going to the market*), **Spatial** (e.g., ... restaurant in the park *next to the school*), and **Attributive** (e.g., ... restaurant in the park *with the ducks*). We also designed a fourth category, **inferred grounding**, to specifically probe the model's understanding of implicit grounding. Due to the insufficient occurrence of these references in the original Meetup and STD datasets (Table 1), we manually constructed 400 Q/A pairs from the dialog contexts to ensure each question has a single, unambiguous answer. The benchmark contains 100 instances for each of the four categories, drawn equally from both datasets.

Each test case is designed to be robust against superficial heuristics; for example, by taking dialogs containing multiple entities of the same type (e.g., several 'sofas'), we force the model to use the provided contextual reference (e.g., "*color of the sofa in the tv room*" or "*color of the sofa in the room in front of the bathroom*") to disambiguate the target entity rather than relying on simple keyword matching. A correct answer thus serves as strong evidence of successful reference resolution, indicating that the model was able to ground the infor-

LLMs	Meetup				Spot the difference			
	Temporal	Spatial	Attributive	Inferred	Temporal	Spatial	Attributive	Inferred
Full Dialog [from Table 2]	0.38 / 0.32	**0.46 / 0.38**	**0.46 / 0.44**	0.20 / 0.20	**0.40 / 0.38**	**0.30 / 0.32**	**0.36 / 0.36**	**0.26 / 0.16**
Summarisation	0.32 / 0.28	0.34 / 0.26	0.30 / 0.25	**0.28** / 0.18	0.22 / 0.16	0.24 / 0.16	0.34 / 0.26	0.20 / 0.12
Chunking (NV-Embbed-V2)	0.24 / 0.20	0.08 / 0.06	0.16 / 0.08	0.22 / 0.24	0.02 / 0.02	0.04 / 0	0.06 / 0.02	0.12 / 0.08
Chunking (BM25)	0.26 / 0.24	0.20 / 0.16	0.20 / 0.18	0.24 / **0.26**	0.04 / 0	0 / 0	0.10 / 0.08	0.12 / 0.10
Agentic Ontology-based	**0.40 / 0.36**	0.38 / 0.34	0.38 / 0.30	0.24 / 0.22	0.20 / 0.14	0.12 / 0.14	0.20 / 0.24	0.18 / 0.14

Table 3: FEM (↑) / LLM-as-Judge (↑) accuracy for various representations using Llama 3.1 8B. Best in **bold**.

mation properly in the first place. Another feature of our evaluation is its ability to test perspective-taking. Each question is framed as the last utterance from one dialogue participant, with the model required to act as the other participant to provide an answer in the next utterance. This design forces the model to correctly attribute knowledge and experiences to the interlocutor. For instance, correctly answering "*What was the color of the second house that I visited?*" requires the model to recall the information shared by the questioner about their sequence of actions. However, this changes fundamentally if the question replaces '*I*' with '*you*', which would require the model to access shared information pertaining to its own character's experiences, thereby testing its ability to maintain distinct perspectives within the common ground. This multi-dimensional aspect of `IndiRef` serves as a metric that can potentially help models to create better retrievable representations of common ground. Appendix D contains test examples, further information on benchmark creation and human evaluation where we achieved 87.5% accuracy.

5 Baselines

To evaluate common ground representation capabilities, we adopt a framework inspired by (Zhang et al., 2025) utilizing three operators: Writer (W), Reader (R), and Generator (G). As a baseline, we assess the LLM's ability to comprehend context by providing the full dialogue history (h_t) (containing the speaker information, utterances, and the time stamps) and tasking them with generating a response (u_{t+1}) to a question in the preceding utterance. We test a range of open-weight **instruction-tuned models**, including Gemma2 (2B, 9B, 27B)(Gemma, 2024), Llama 3.1 8B(Llama, 2024), and the reasoning-focused Qwen-QWQ-32B (Qwen, 2025). In this setup, W appends each new utterance to the history, R provides the entire history to the model, and G produces the subsequent utterance.

$$W(h_{t-1}, u_t) = h_t;\ R(u_{t+1}) = h_t;\ G(h_t, u_{t+1}) = u_{t+1}$$

6 Common Ground Representation

The impracticality of using a complete, ever-growing dialogue history in real-world systems necessitates better methods to represent common ground. Thus, we evaluate three established techniques for managing long-form dialogues: summarization, utterance chunking, and ontology creation.

We focus first on **summarization**, a method shown by (Padmakumar et al., 2023) to be a highly effective representation of open-dialogue context for transformer-based models. Its utility is further underscored by its adoption in modern LLM chatbots for personalization (Zhang et al., 2024) by summarizing each session separately. In this scheme, W creates a summary (s_t) of the whole dialogue, which is stored, retrieved by R, and used by G to form the next utterance.

$$W(s_{t-1}, u_t) = s_t;\ R(u_{t+1}) = s_t;\ G(s_t, u_{t+1}) = u_{t+1}$$

Our next approach, **utterance chunking**, is our first to explicitly group information to enhance retrieval. Here, W segments the dialogue into overlapping chunks of utterances (c_i). For our experiments, we define a chunk as seven utterances with a stride of three to ensure informational continuity. R then retrieves the top three chunks most relevant to the user's utterance by matching its embedding against the embeddings of each chunk. To isolate the performance of this structural approach, our Reader only uses the embeddings of the concatenated utterances within a chunk, and does not use additional embeddings of chunk summaries as done by (Wu et al., 2025). To better understand how embeddings affect the retrieval of important context, we tested two types of embedders: a sparse model (BM25 (Robertson et al., 1994)) and a dense model (NV-Embed-V2 (Lee et al., 2024)).

$$\begin{aligned} &W(h_{t-1}, u_t) = [c_1, \ldots, c_i] \\ &R(u_{t+1}) = \text{top}_k(\text{sim}(e(u_{t+1}), [e(c_1), \ldots, e(c_i)])) \\ &G(u_t, [c_1, \ldots, c_k]) = u_{t+1} \end{aligned}$$

Finally we use an **agentic** approach which investigates the use of an **ontology** (O) to create a

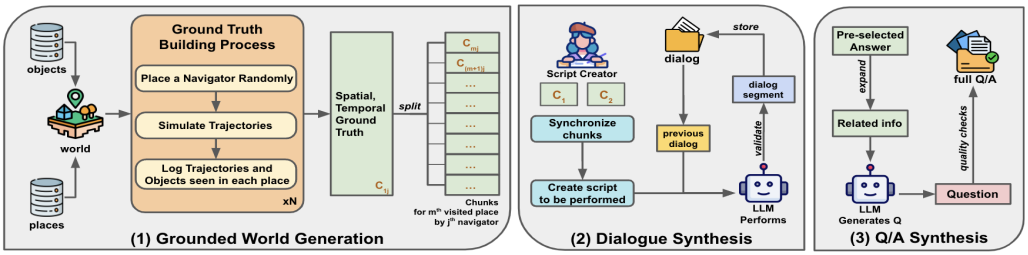

Figure 2: Pipeline for the synthetic data generation involving three main phases: (1) Grounded World Generation, creates the factual basis of the scenario; (2) Dialogue Synthesis, where LLM generates conversation constrained by the script; and (3) Q/A Synthesis, which extracts and formats question-answer pairs for training.

structured common ground. The Writer (W) processes dialogue chunks to extract key entities, assigning each a unique id and mapping its properties, inter-entity relations and source speaker as a proxy for estimating each speaker's belief. This rich structure enables complex relational reasoning (e.g., if *'place A has object b'* and *'b is kept on top of c'*, then *'A also has c'*). To preserve temporal context, W also maintains a chronological log of events, which includes the involved entity IDs and speakers, serving as a primary index for querying. Given this highly structured knowledge, R uses a multi-step query process to retrieve and process information before sending it to G. It first decomposes the user's request into a series of sub-queries (q), each prefixed with a command: $RAG[n]$, $Process$, or $Final$. A $RAG[n]$ query retrieves the top-n relevant contexts (c) from the extracted common ground. Here n is also generated by the LLM. A $Process$ query then performs operations on this retrieved data, such as filtering or selecting specific items. The final retrieved context and the $Final$ query are passed to G to synthesize the final answer. All operations are implemented using langchain (Chase, 2022) for the vectorstore and NV-Embed-V2 for embedding generation.

$$W(O_{t-1}, u_t) = O_t$$
$$R(u_{t+1}) = [q_1....q_n]$$
$$R(q_i, O_t, c_{i-1}) = c_i$$
$$G(q_n, c_{n-1}) = u_{t+1}$$

7 Evaluation

To comprehensively evaluate performance, we employ a Fuzzy Exact Matching (FEM) metric. For each question, an answer is considered correct if it contains a predefined keyword or phrase. Recognizing that keyword presence alone may not guarantee accuracy, we also incorporate an LLM-as-a-judge approach (Zheng et al., 2023). A Qwen2.5-32B-Instruct (Yang et al., 2024) model, provided with the question and both the correct and generated answers, determines correctness based on semantic equivalence. Crucially, dialog context was intentionally excluded from the LLM's input to ensure its judgment focuses solely on the answers' semantic similarities. Since the relational references point to only one correct referent, the overall performance for both fuzzy matching and the LLM-as-a-judge approach was quantified using accuracy.

Table 2 presents the performance of various LLMs when provided with the full dialog context. For Meetup, results indicate that model performance improves with increasing size, with Gemma2-2B consistently exhibiting the lowest accuracies across all categories. The relative performance of Fuzzy Matching and LLM-as-a-judge remains consistent across models. Slight variations occur when Fuzzy Matching scores differ due to longer answers with multiple entities or alternate representations (e.g., digits vs. words). Gemma2-9B has similar performance to Llama 3.1-8B as both marginally surpass each other across various categories. Surprisingly, Gemma2-27B emerges as the top performer across all evaluated models, outperforming even Qwen-QWQ which is a slightly larger model specifically trained for enhanced reasoning. Further analysis of Qwen-QWQ's responses revealed difficulties in logical reasoning and frequent hallucinations(Appendix; Figure 12), potentially attributable to a scarcity of spoken dialogs in its training data. Despite this, Qwen-QWQ excelled in inferred grounding questions, a category where other models performed poorly, underscoring the critical role of reasoning for grounding implicit information. It also had better fuzzy exact match scores due to it's verbose nature. However, in STD, Gemma models underperformed compared to Llama. We hypothesis that the longer dialog context in STD as compared to Meetup helped Llama which uses RoPE over Gemma which uses sliding window attention. Furthermore, the models demonstrated significant

LLMs	Meetup				Spot the difference			
	Temporal	Spatial	Attributive	Inferred	Temporal	Spatial	Attributive	Inferred
Original Setting [Table 2]	0.38 / 0.32	0.46 / 0.38	0.46 / **0.44**	0.20 / 0.20	0.40 / 0.38	0.30 / 0.32	0.36 / 0.36	0.26 / 0.16
In-Context Learning	**0.60** / **0.56**	0.58 / **0.54**	**0.62** / 0.58	0.42 / 0.34	0.48 / 0.52	**0.48** / 0.40	0.52 / 0.54	0.36 / **0.44**
GRPO Trained	0.58 / 0.52	**0.66** / **0.54**	**0.62** / **0.60**	**0.46** / **0.42**	**0.52** / **0.54**	**0.48** / **0.42**	**0.56** / **0.58**	**0.38** / 0.42

Table 4: FEM (↑) / LLM-as-Judge (↑) accuracy on entire dialog history for enhanced Llama 3.1-8B. Best in **bold**.

limitations in differentiating between speaker perspectives (e.g., 'your' vs 'mine') and in isolating information conveyed by individual speakers, often resulting in the erroneous merging of conversational content. It is critical to note that **none of the baselines achieved an accuracy exceeding 50% across all the categories**.

The performance of various common ground representation techniques is detailed in Table 3. Due to space constraints, we only present results of the Llama3.1-8B model. For Gemma2 2B and 9B models, refer to Appendix B. We did not test larger models due to resource limitations. The general trend across all models remained consistent. Summarization generally yielded poorer performance compared to the baseline, a likely consequence of information loss during the compression process. Chunking utterances into discrete groups for retrieval proved suboptimal for handling relational references, likely stemming from the disruption of inter-utterance relationships across chunks. Moreover, sparse embedding (BM25) demonstrated a slight advantage over dense embedding, which may be attributed to its enhanced capability in named-entity retrieval. The agentic ontology-based system outperformed other representation techniques but still lagged behind LLMs with full dialog context. While it performed relatively well on the Meetup dataset, it struggled with the STD dataset, often merging the representations of images of both participants due to their similarity. However, its performance compared to other representation techniques suggests that a multi-step retrieval process, particularly one that explicitly extracts speaker-specific information regarding entities and events, helps the LLM understand the context better.

8 Performance Improvement

From our baseline analysis, it becomes evident that existing LLMs even with full dialog history may not be optimally suited for robust information grounding within a situated spoken dialog context. A significant challenge lies in their difficulty in understanding overlapping information from multiple speakers, preventing the construction of a reliable grounding representation. Thus, our initial research focused on enhancing the performance of LLMs while leveraging the entire dialog history. We therefore implemented In-Context Learning (Dong et al., 2022) by augmenting the input prompts with examples for each reference type to instruct the LLMs on how to accurately ground the context within the dialogues. The prompts (see Appendix E) were further enriched with explicit guidance on decoupling the information contributed by each speaker, unless a common entity was being discussed. As shown in Table 4, this resulted in a significant improvement in LLM performance across all categories in both datasets. These findings indicate that enhanced prompt design can substantially improve the models' comprehension of spoken conversations, a critical prerequisite for effective grounding.

To enable the model to develop inherent grounding capabilities applicable to diverse settings, we transitioned to a reinforcement learning approach. We used Group Relative Policy Optimization (GRPO) (Shao et al., 2024) to train the model to identify the important context by providing positive rewards for each accurate response. A primary challenge was the absence of suitable training data for situational spoken dialogues, and so we opted to generate synthetic data to resolve data scarcity with unlimited data scaling, provide granular domain control, and test edge cases. However, since LLMs are not natively trained on spoken dialogues, we first needed to develop a framework to generate high-quality, situational dialogue data.

Synthetic Data: We designed a systematic, multi-phase synthetic data generation framework with an "environment-first, dialogue-later" approach. Inspired by procedural content generation (Smith, 2015) and recent advances in synthetic dialogue generation (Suresh et al., 2025), our framework is engineered to produce a grounded situated dialogue corpus. The process is structured into three phases as shown in Figure 2:

Grounded World Generation: First, a rich and verifiable simulated 'world' is programmatically constructed, starting from a list of places and related objects previously generated (see Appendix

LLMs	Meetup				Spot the difference			
	Temporal	Spatial	Attributive	Inferred	Temporal	Spatial	Attributive	Inferred
Without GRPO training [Table 3]	0.40 / 0.36	0.38 / 0.34	0.38 / 0.30	0.24 / 0.22	0.20 / 0.14	0.12 / 0.14	0.20 / 0.24	0.18 / 0.14
With GRPO training	0.48 / 0.46	0.44 / 0.42	0.52 / 0.44	0.36 / 0.38	0.20 / 0.24	0.22 / 0.28	0.28 / 0.32	0.16 / 0.20

Table 5: FEM (↑) / LLM-as-Judge (↑) accuracy for agentic ontolgy-based representation after GRPO training.

A). Two algorithmic navigators explore this world, and their experiences are logged to establish the spatio-temporal ground truth (i.e., what they saw, where they went, etc.) that we can then split into smaller chunks ($\mathcal{C}_{m,j}$) which is the chunk for the mth visited place by the jth navigator

Dialogue Synthesis: We then synthesize dialogs where the two navigators describe their locations to find each other. It is constrained by the rules and facts of the created world. A programmatic Script Creator synchronizes the chunks of the two Navigators and generates the script for the LLM to follow, thus generating dialogue segments based on the log events. In this way, we rely on the Script Creator for the 'reasoning' part of the task (i.e., noticing similarities, addressing ambiguities, resolving references, etc), while the LLM focuses on constrained Language Generation. This generation process is iterative and considers the previously generated utterances for cohesiveness.

Q/A Synthesis: The final stage automatically identifies scenarios to generate accurate Question-Answer (Q/A) pairs requiring complex reference resolution. The process is deterministic: a fact is preselected from the Ground Truth as the 'expected answer', and the LLM is prompted to generate a question, constrained to the context, that elicits this specific answer. These are designed to probe various aspects of common ground, including spatial, temporal, and relational facts, and undergo a strict validation process. We generated ~ 600 Q/A pairs from various synthetic dialog scenarios.

Table 4 demonstrates that reinforcement learning on the synthetic data improved model accuracy by 15-20% not only on the Meetup dataset (which closely resembles the training data), but also on the Spot-the-Difference dataset. This indicates that the models successfully acquired the underlying concepts needed to resolve such relational references to the common ground. We note that for the incorrect responses, the models would often reason in the right direction but then hallucinate while providing the final answer. We expect this to improve with larger models but, due to lack of computational resources, we did not train on them.

To assess whether the enhanced contextual understanding of GRPO-trained models improves their representational capabilities, we applied the agentic ontology-based representation using these same GRPO-trained models trained on synthetic Q/A pairs, given their promising results in earlier experiments. As shown in Table 5, the GRPO-trained model outperformed its non-GRPO counterpart on the Meetup dataset but continued to struggle with the Spot-the-Difference dataset. The main issue concerned representing the two images as distinct within the common ground, as the models frequently merged them due to their high degree of similarity. While more carefully designed task-specific prompts could improve performance, these would limit generalizability and thus our findings fall short of our broader goal: developing a universal model for common ground representation that relies only on spoken utterances and minimal contextual descriptions.

9 Conclusion and Future Work

In this work, we investigated the critical ability of dialog systems to construct and utilize common ground in dialogue. To this end, we developed `IndiRef` a benchmark derived from human conversations, to evaluate how models represent, store, and exploit this shared knowledge. Our experiments show that both in-context learning and GRPO training with synthetic data improve an LLM's ability to resolve relational references from the full dialogue history. We also found that GRPO training moderately enhances common ground extraction for long-term use in agentic setups. However, inconsistent performance across datasets highlights the need for more robust representation methods. For future work, while our study focused on interpretable textual representations, exploring more efficient parametric methods is a promising direction. Furthermore, a crucial next step is to integrate the creation of common ground with grounding acts, such as clarifying ambiguities or updating shared knowledge when new information creates contradictions. We believe our benchmark and proposed methods provide a solid foundation for advancing research into more sophisticated and robust common ground representation in dialogue agents.

Limitations

Certain limitations offer avenues for future research. First, the dataset used for our test cases, while selected to approximate spontaneous spoken dialogue, has inherent constraints. The ideal dataset would comprise spontaneous, multi-session spoken dialogs. The logistical challenges in acquiring such data necessitated the use of two specialized dialog corpora, Meetup and Spot the Difference that, while effective in providing a range of grounding phenomena through a game-based framework, may not entirely capture the nuances of spontaneous multi-topic human interaction.

Second, due to computational resource limitations, our empirical evaluation was restricted to the 8-billion parameter Llama 3.1 instruction-tuned model especially for the GRPO-based training. A central hypothesis in current NLP research is that increasing model scale can reduce issues like hallucination. While we conjecture that our findings would benefit from this effect, validating this hypothesis was beyond the scope of our resources. A valuable direction for future work would be to replicate and extend our experiments using larger models.

Lastly, while our test cases provide a solid foundation for testing the model's ability to use established common ground, they do not comprehensively test it's capacity to handle 'assumptions' and 'beliefs'. These deeper cognitive elements, while tilting towards theory of mind research, are still essential for robust conversational grounding and present a clear direction for future investigation.

References

Philippe Blache and Matthis Houlès. 2021. Common ground, frames and slots for comprehension in dialogue systems. pages 33–43.

Chenyang Bu, Guojie Chang, Zihao Chen, CunYuan Dang, Zhize Wu, Yi He, and Xindong Wu. 2025. Query-driven multimodal GraphRAG: Dynamic local knowledge graph construction for online reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21360–21380, Vienna, Austria. Association for Computational Linguistics.

J Cassell, T Bickmore, L Campbell, H Vilhjálmsson, and H Yan. 2001. More than just a pretty face: conversational protocols and the affordances of embodiment. *Knowledge-Based Systems*, 14(1):55–64.

Harrison Chase. 2022. Langchain. https://github.com/langchain-ai/langchain.

Huili Chen, Yubin Kim, Kejia Patterson, Cynthia Breazeal, and Hae Won Park. 2025. Social robots as conversational catalysts: Enhancing long-term human-human interaction at home. *Science Robotics*, 10(100):eadk3307.

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.

Herbert H. Clark and Ed Schaefer. 1989. Contributing to discourse. *Cogn. Sci.*, 13:259–294.

Alexandre Denis. 2010. Generating referring expressions with reference domain theory. In *Proceedings of the 6th International Natural Language Generation Conference*. Association for Computational Linguistics.

Maria Di Maro, Antonio Origlia, and Francesco Cutugno. 2021. Cutting melted butter? common ground inconsistencies management in dialogue systems using graph databases. In *IJCoL, 7-1, 2*, pages 157–190.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2022. A survey on in-context learning. In *Conference on Empirical Methods in Natural Language Processing*.

Team Gemma. 2024. Gemma 2: Improving open language models at a practical size. *ArXiv*, abs/2408.00118.

Mikael Henaff, Jason Weston, Arthur Szlam, Antoine Bordes, and Yann LeCun. 2016. Tracking the world state with recurrent entity networks. *ArXiv*, abs/1612.03969.

Mohammad Javad Hosseini, Filip Radlinski, Silvia Pareti, and Annie Louis. 2023. Resolving indirect referring expressions for entity selection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12313–12335, Toronto, Canada. Association for Computational Linguistics.

Nikolai Ilinykh, Sina Zarrieß, 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.

Léo Jacqmin, Lina M. Rojas Barahona, and Benoit Favre. 2022. "do you follow me?": A survey of recent approaches in dialogue state tracking. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 336–350, Edinburgh, UK. Association for Computational Linguistics.

John D. Kelleher and Geert-Jan M. Kruijff. 2006. Incremental generation of spatial referring expressions in situated dialog. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 1041–1048, Sydney, Australia. Association for Computational Linguistics.

Ibrahim Khalil Khebour, Kenneth Lai, Mariah Bradford, Yifan Zhu, Richard A. Brutti, Christopher Tam, Jingxuan Tu, Benjamin A. Ibarra, Nathaniel Blanchard, Nikhil Krishnaswamy, and James Pustejovsky. 2024. Common ground tracking in multimodal dialogue. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3587–3602, Torino, Italia. ELRA and ICCL.

Seokhwan Kim, Spandana Gella, Chao Zhao, Di Jin, Alexandros Papangelis, Behnam Hedayatnia, Yang Liu, and Dilek Z Hakkani-Tur. 2023. Task-oriented conversational modeling with subjective knowledge track in DSTC11. In *Proceedings of the Eleventh Dialog System Technology Challenge*, pages 274–281, Prague, Czech Republic. Association for Computational Linguistics.

Emiel Krahmer and Kees van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Jaap Kruijt and Piek Vossen. 2022. The role of common ground for referential expressions in social dialogues. In *Proceedings of the Fifth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 99–110, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Yuxuan Li, Vijay Veerabadran, Michael L. Iuzzolino, Brett D. Roads, Asli Celikyilmaz, and Karl Ridgeway. 2025. Egotom: Benchmarking theory of mind reasoning from egocentric videos. *Preprint*, arXiv:2503.22152.

Team Llama. 2024. The llama 3 herd of models. *ArXiv*, abs/2407.21783.

José Lopes, Nils Hemmingsson, and Oliver Åstrand. 2018. The spot the difference corpus: a multi-modal corpus of spontaneous task oriented spoken interactions. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Chris Madge, Matthew Purver, and Massimo Poesio. 2025. Referential ambiguity and clarification requests: comparing human and llm behaviour. *ArXiv*, abs/2507.10445.

Biswesh Mohapatra, Seemab Hassan, Laurent Romary, and Justine Cassell. 2024a. Conversational grounding: Annotation and analysis of grounding acts and grounding units. In *Proceedings of LREC-COLING*. Association for Computational Linguistics.

Biswesh Mohapatra, Manav Nitin Kapadnis, Laurent Romary, and Justine Cassell. 2024b. Evaluating the effectiveness of large language models in establishing conversational grounding. In *Conference on Empirical Methods in Natural Language Processing*.

Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. 2022. Teach: Task-driven embodied agents that chat. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36:2017–2025.

Vishakh Padmakumar, Behnam Hedayatnia, Di Jin, Patrick Lange, Seokhwan Kim, Nanyun Peng, Yang Liu, and Dilek Hakkani-Tur. 2023. Investigating the representation of open domain dialogue context for transformer models. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 538–547, Prague, Czechia. Association for Computational Linguistics.

Shuwen Qiu, Mingdian Liu, Hengli Li, Song-Chun Zhu, and Zilong Zheng. 2024. MindDial: Enhancing conversational agents with theory-of-mind for common ground alignment and negotiation. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 746–759, Kyoto, Japan. Association for Computational Linguistics.

Team Qwen. 2025. Qwq-32b: Embracing the power of reinforcement learning.

Janarthanan Rajendran, Jatin Ganhotra, and Lazaros C. Polymenakos. 2019. Learning end-to-end goal-oriented dialog with maximal user task success and minimal human agent use. *Transactions of the Association for Computational Linguistics*, 7:375–386.

Sanjana Ramprasad, Elisa Ferracane, and Zachary Chase Lipton. 2024. Analyzing llm behavior in dialogue summarization: Unveiling circumstantial hallucination trends. In *Annual Meeting of the Association for Computational Linguistics*.

Stephen Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at trec-3. pages 0–.

David Schlangen and Gabriel Skantze. 2009. A general, abstract model of incremental dialogue processing. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 710–718, Athens, Greece. Association for Computational Linguistics.

Melanie Sclar, Jane Yu, Maryam Fazel-Zarandi, Yulia Tsvetkov, Yonatan Bisk, Yejin Choi, and Asli Celikyilmaz. 2024. Explore theory of mind: Program-guided adversarial data generation for theory of mind reasoning. *Preprint*, arXiv:2412.12175.

Olga Seminck. 2023. Conversational AI: Dialogue systems, conversational agents, and chatbots by Michael McTear. *Computational Linguistics*, 49(1):257–259.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junmei Song, Mingchuan Zhang, Y. K. Li, Yu Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *ArXiv*, abs/2402.03300.

T. Shopen. 1985. *Language Typology and Syntactic Description: Volume 3*. Language Typology and Syntactic Description. Cambridge University Press.

Mohit Shridhar, Dixant Mittal, and David Hsu. 2020. Ingress: Interactive visual grounding of referring expressions. *Int. J. Rob. Res.*, 39(2–3):217–232.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Gillian Smith. 2015. An analog history of procedural content generation. In *FDG*, pages 1–6. Boston, MA.

Sathya Krishnan Suresh, Wu Mengjun, Tushar Pranav, and EngSiong Chng. 2025. DiaSynth: Synthetic dialogue generation framework for low resource dialogue applications. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 673–690, Albuquerque, New Mexico. Association for Computational Linguistics.

David Traum and James Allen. 1994. A "speech acts" approach to grounding in conversation.

David R. Traum and Staffan Larsson. 2003. *The Information State Approach to Dialogue Management*, pages 325–353. Springer Netherlands, Dordrecht.

Lina Varonina and Stefan Kopp. 2021. Knowledge modelling for establishment of common ground in dialogue systems. *Italian Journal of Computational Linguistics*, 7:9–26.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. Trl: Transformer reinforcement learning. `https://github.com/huggingface/trl`.

Nicholas Thomas Walker, Stefan Ultes, and Pierre Lison. 2022. Graphwoz: Dialogue management with conversational knowledge graphs. *ArXiv*, abs/2211.12852.

Tom Williams, Fereshta Yazdani, Prasanth Suresh, Matthias Scheutz, and Michael Beetz. 2019. Dempster-shafer theoretic resolution of referential ambiguity. *Auton. Robots*, 43(2):389–414.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. 2025. Longmemeval: Benchmarking chat assistants on long-term interactive memory. *ICLR*, abs/2410.10813.

Qingyang Wu and Zhou Yu. 2024. Stateful memory-augmented transformers for efficient dialogue modeling. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 853–867, St. Julian's, Malta. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Zeyu Zhang, Quanyu Dai, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. 2025. A survey on the memory mechanism of large language model-based agents. *ACM Trans. Inf. Syst.*, 43(6).

Zhehao Zhang, Ryan A. Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, Ruiyi Zhang, Jiuxiang Gu, Tyler Derr, Hongjie Chen, Ju-Ying Wu, Xiang Chen, Zichao Wang, Subrata Mitra, Nedim Lipka, and 2 others. 2024. Personalization of large language models: A survey. *ArXiv*, abs/2411.00027.

Zheng Zhang, Minlie Huang, Zhongzhou Zhao, Feng Ji, Haiqing Chen, and Xiaoyan Zhu. 2019. Memory-augmented dialogue management for task-oriented dialogue systems. *ACM Trans. Inf. Syst.*, 37(3).

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

A Synthetic Data Generation

We expand upon the Synthetic Data generation process to offer a comprehensive understanding of the

framework's design and operation. The core objective of this generation process extends beyond mere factual accuracy to the creation of dialogues that are inherently spoken-like and situational. We specifically design the framework to model the complex dynamics of conversational grounding, where participants actively establish and maintain mutual understanding of entities and facts. This is achieved by programmatically incentivizing and modeling real-world conversational phenomena, including high volumes of back-and-forth exchanges, frequent backchanneling (e.g., "uh-huh," "yeah"), and explicit affirmations or clarifications of shared knowledge. By ensuring the agents are constantly listening to and reliably building upon each other's words, the resulting corpus aims to a deeper, more reliable form of collaborative communication.

A.1 Grounded World Generation

The environment is not a simple random collection of items; it's built from a structured taxonomy (see Figure 3). The process starts with a curated list of abstract places, like `festival grounds` or `historic museum`. For each place, we have generated inventories of objects and edges (paths). Objects are coherent with the place and highly detailed with properties and relations to other objects. This granular data provides the raw materials for a rich, fact-based world.

The simulation is a time-series event logger. Once a place is selected and populated with objects, two non-sentient virtual navigators are placed within it. The system then simulates their movement, creating a `full_scenario` file that logs their 'trajectories' over a set time span. This file is the definitive ground truth, recording what each navigator "sees," their paths, and their arrival/departure times. This detailed record ensures that every piece of information later referenced in the dialogue is verifiable. The continuous Ground Truth is then partitioned into smaller chunks ($\mathcal{C}_{m,j}$), where j denotes the navigator and m denotes the index of the visited location $\mathcal{P}$ (or time segment). The entire corpus of chunks is represented as $\mathcal{C}_{\text{all}} = \bigcup_{j=1}^{|\mathcal{N}|} \bigcup_{m=1}^{|\mathcal{P}_j|} \mathcal{C}_{m,j}$.

A.2 Dialogue Synthesis

A programmatic "Script Writer" acts as the crucial link between the simulated world and the LLM. It takes the structured information from a time segment, algorithmically 'reasons' over it, and converts this into a "script" prompt (in dictionary format) for the LLM to "perform" like an actor. This pre-processing is key to offloading the reasoning burden from the LLM. For example, the model will directly receive something like Listing 1.

Listing 1: Example of a $\mathcal{C}_{m,j}$ chunk for the LLM

```
"place_type": "large bakery production facility",

"time_span_s": [175, 212],
"navigator_A": {
  "name": "Michael",
  "zone_type": "packaging and storage zone",
  "status": "arrived & departing",
  "arrived_from": "office area connection",
  "objects_to_mention": [
    {"type": "packaging machine"},
    {
        "type": "label applicator",
        "property": "wipe-on",
        "relation": "holding",
        "reference": "label roll holder"
    },
    {"type": "delivery cart"},
    {"type": "refrigeration unit"}
  ]
}
... //continues
```

The Script Creator systematically introduces complexity to the dialogue (i.e., noticing similarities, addressing ambiguities, etc.) This includes generating sentences that reference objects seen by the other navigator (indirect reference) or that require a deeper understanding of the shared context.

Listing 2: 'Reasoned' part by the Script Writer

```
... //continues
"current_observer": {
    "name": "Michael",
    "full_details_of_what_Michael_is_seeing": {
        "type": "delivery cart",
        "property": "fixed axle",
        "relation": "secured by",
        "reference": "securing strap"
    },
    "ask_question_about": "Michael asks about a
     delivery cart seen before by Kenneth",
    "to_avoid_in_question": "**Michael does NOT
     ask about specific properties that only
     Kenneth has seen.**"
},
"to_discuss_with": {
    "partner_name": "Kenneth",
    "full_details_of_what_Kenneth_saw": {
        "type": "delivery cart",
        "property": "solid rubber",
        "relation": "near",
        "reference": "loading ramp"
    },
    "seen_after_taking": "pallet movement lane",
    "that_place_also_had": "waste compactor"
}
```

The LLM then outputs the dialogue:

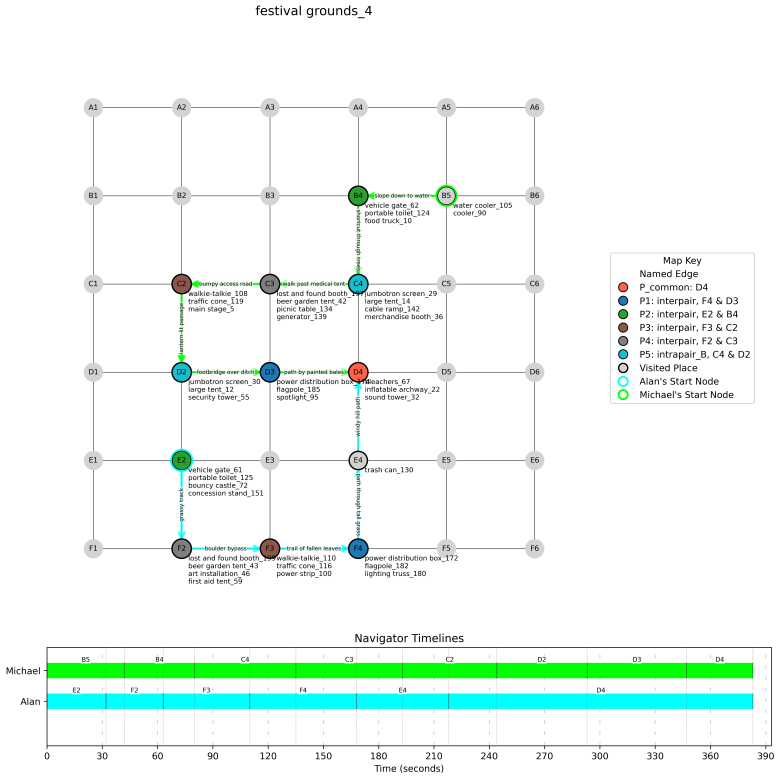

Figure 3: A visualization of a `full_scenario`. Unique objects, with their id, are crafted to have similar characteristics to induce ambiguity that has to be resolved. The ambiguity can arise from a navigator's experience alone (`intrapair`) or for their experience combined (`interpair`). Each chunks ($\mathcal{C}_{m,j}$) is delimited by the densely dotted lines.

Listing 3: Example of LLM-generated dialogue segment

```
Michael:    I just arrived in the packaging and
    storage zone via the office area connection.
Kenneth:    Uh-huh, I'm still in the mixing and
    preparation zone.
Michael:    There's a packaging machine here
    with an adhesive seal protected by a safety
    interlock.
Kenneth:    Got it. My packaging machine here
    uses an ultrasonic seal and receives from a
    product feed conveyor.
Michael:    Noted. There's also a label
    applicator with a wipe-on method holding a
    label roll holder.
Kenneth:    Mine's a tamp-apply applicator
    triggered by a product sensor. Any delivery
    carts around?
Michael:    Yes, there's a delivery cart with a
    fixed axle secured by a securing strap.
Kenneth:    Mine's near a loading ramp with
    solid rubber wheels.
... //continues
```

A.3 Q/A Synthesis

First, we automatically identify "questionable moments" - where there are multiple entities of similar nature and a relational reference to them would only be resolved with a good grounding mechanism. We do this using both the ground truth ($\mathcal{C}_{\text{all}}$) and the dialogue transcript ($\mathcal{D}$), algorithmically analyzing the conversation and flagging instances in a structured list.

A fact is preselected from the Ground Truth as the "expected answer", and the LLM is prompted to generate a question that elicits this specific answer. The complex context for a moment is captured in a JSON fact:

Listing 4: Example of a fact for Q/A generation

```
"fact_id": 17,
"chunk_id": "chunk_007",
"question_type": "inferred_comparison_result",
"questioner": "Kenneth",
"answerer": "Michael",
"context": {
    "property_of_obj_seen_by_Michael": "fixed
     axle",
    "property_of_obj_seen_by_Kenneth": "solid
     rubber",
    "ambiguous_object_type": "delivery cart",
    "disambiguating_fact": "the one that was
     where there was also a waste compactor"
},
"expected_answer": "solid rubber",
```

Given the whole context, the LLM generates the following questions:

Listing 5: Example of the generated and validated Question Pair

```
"question_full": "Of the two delivery carts we
    discussed-yours with a fixed axle and mine
    with solid rubber wheels-which one was
    located where there was also a waste
    compactor?"
"question_short": "Between your delivery cart
    with a fixed axle and mine with solid rubber
     wheels, which was near the waste compactor
    ?",
```

A.4 Validation and Quality Control

To ensure the highest quality of our generated data, the entire pipeline is subject to algorithmic validation, enforcing structural integrity and semantic accuracy. On the Dialogue, we prevent excessive monologues by both parts, we ensure all mandatory objects seen in $\mathcal{C}_{m,j}$ are mentioned, and we verify that movement intentions are stated. On the Questions, instead, we verify that the questions contain mandatory keywords from the scenario, making them directly tied to the ground truth; we ensure the question text does not contain the expected answer; and finally we also ensure that the questions use the correct pronouns based on the point-of-view of the questioner and answerer.

Of all the questionable facts identified, only ~ 6 thousand ($\sim 36\%$) of them were successfully converted into validated Q/A pairs. This high rejection rate directly deals with logical inconsistencies, or factual errors, enhancing the final quality of the corpus.

B Gemma Results

Tables 6 and 7 contain all the baseline results for Gemma2 2B and 9B models. As seen from the Tables, the overall pattern remains the same as compared to Llama 3.1-8B. The 2 Billion model performs relatively poor compared to the 9 Billion model across all the categories. Chunking in general provides the worst results among all the representation techniques suggesting the importance of good abstraction of information that preserves the links between the information across different segments of the conversation. Surprisingly, for the gemma2 models that struggled with longer context of STD, the agentic system performs better than when provided with the full dialog for Meetup test cases. This suggests that providing a structure might help it reason better for longer contexts.

C GRPO Training

For our GRPO training, we used 600 Question/Answer pairs from 60 artificially generated dialogs. While we can generate much more synthetic data, we were contrained by the computing resources and decided to use only 600 instances which showed to be enough for the model as the reward stabilized towards the end of the training as seen in Figure 4. The huggingface TRL (von Werra et al., 2020) library was used for GRPO training.

C.1 Hyperparameters

We used a batch size of 16 for the RL training. The training used 8 A100 GPUs in total out of which 4 A100s were used by VLLM for the inference. Here are our other hyperparameters : learning rate - 1e-6, GRPO beta - 0.025, weight_decay - 0.1, adam_beta1 - 0.9, adam_beta2 - 0.99, warm-up ratio - 0.1, and a cosine learning scheduler.

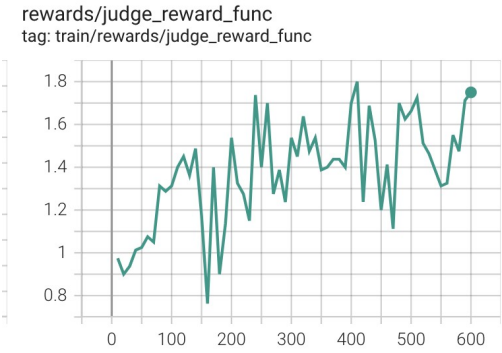

Figure 4: Training reward over time while using the synthetic data. Here the y-axis is the reward while x-axis is the step.

D Benchmark Creation

The creation of our test cases involved an initial selection of 50 dialogues per category from both datasets. The authors of this paper then undertook the task of creating question/answer pairs that contained relational references to entities within the established common ground of each dialogue. A key methodological consideration was the exclusive use of conversational data to formulate these questions. This approach ensures that the common ground is derived solely from the information discussed, rather than any external knowledge individual participants may have from the images they viewed. For validation, the benchmark was evaluated by 10 volunteers with expertise in Natural Language Processing. These volunteers were provided with the dialogues and asked to answer the corresponding questions. We randomly selected 10% of test cases from each category for this validation process and each test case was provided

LLMs	Meetup				Spot the difference			
	Temporal	Spatial	Attributive	Inferred	Temporal	Spatial	Attributive	Inferred
Full Dialog [from Table 2]	0.20 / 0.18	0.18 / 0.16	0.24 / 0.26	0.26 / 0.16	0.18 / 0.12	0.18 / 0.10	0.20 / 0.12	0.16 / 0.10
Summarisation	0.16 / 0.14	0.12 / 0.12	0.20 / 0.14	0.18 / 0.14	0.12 / 0.10	0.14 / 0.12	0.12 / 0.10	0.08 / 0.02
Chunking (NV-Embbed-V2)	0.16 / 0.14	0.04 / 0	0.12 / 0.10	0.22 / 0.16	0 / 0	0.02 / 0	0.02 / 0	0.02 / 0
Chunking (BM25)	0.24 / 0.20	0.24 / 0.20	0.20 / 0.18	0.16 / 0.12	0 / 0	0 / 0	0.06 / 0.02	0.04 / 0
Agentic Ontology-based	0.40 / 0.36	0.32 / 0.26	0.26 / 0.22	0.20 / 0.16	0.12 / 0.06	0.10 / 0.04	0.14 / 0.08	0.06 / 0.10

Table 6: FEM (↑) / LLM-as-Judge (↑) accuracy for various representations using Gemma2-2B.

LLMs	Meetup				Spot the difference			
	Temporal	Spatial	Attributive	Inferred	Temporal	Spatial	Attributive	Inferred
Full Dialog [from Table 2]	0.48 / 0.38	0.50 / 0.46	0.38 / 0.40	0.26 / 0.18	0.34 / 0.20	0.24 / 0.16	0.28 / 0.16	0.20 / 0.10
Summarisation	0.28 / 0.22	0.24 / 0.18	0.28 / 0.24	0.16 / 0.14	0.12 / 0.10	0.10 / 0.10	0.12 / 0.12	0.16 / 0.14
Chunking (NV-Embbed-V2)	0.22 / 0.16	0.10 / 0.10	0.18 / 0.16	0.22 / 0.22	0 / 0	0 / 0.02	0 / 0	0.04 / 0
Chunking (BM25)	0.34 / 0.26	0.22 / 0.24	0.24 / 0.26	0.26 / 0.24	0.02 / 0	0.02 / 0	0.02 / 0	0.06 / 0.02
Agentic Ontology-based	0.44 / 0.46	0.40 / 0.38	0.42 / 0.36	0.26 / 0.24	0.32 / 0.32	0.36 / 0.32	0.30 / 0.26	0.28 / 0.16

Table 7: FEM (↑) / LLM-as-Judge (↑) accuracy for various representations using Gemma2-9B.

to two volunteers. The accuracy rate of 87.5% affirmed the correctness of our test suite. A subsequent error analysis indicated that the majority of mistakes were the result of human error.

Illustrative examples for each test case category are provided in Figures 7, 6, 8, and 5. Specifically, we present two examples (Spatial and Temporal) from the Spot the Difference dialogues and two examples (Attributive and Inferred) from the Meetup dialogues. In these figures, we have highlighted the essential segment of the dialogue containing the answer (in purple) to the question (in violet). We also provide the keywords needed for the exact match analysis.

As can be seen in the examples, our questions are always regarding an object that has multiple instances with different properties. For example, Figure 5 provides an example of inferred grounding, where, while A asks B if his child's room has a teddy bear, B says no. This should ground two facts - A's room has a teddy bear while B's room doesn't. However, the question is asked by A about his child's room. This should have made B answer about A's room but since in the original dialog, the answer to the question is 'nope', our models generally provide a negative answer. '...' in our figures indicate a time gap between the chunk of utterances as can be seen in Figure 8 where we have a gap from the time [02:21] till [10:45] and again till [12:04]. Figure 8 references a white cloud mentioned in preceding images, yet the question is specifically directed at the final image. Figure 9 provides the system prompt used for judging the correctness of the answers using LLM as a judge.

E Prompts

During testing, we utilized the system prompts shown in Figures 9, 10 and 11. Our approach leverages chain-of-thought prompting (Figure 10), which we found to be more effective than direct questioning. This prompt outlines the game's rules to help the LLM understand the context. However, it differs from the in-context prompt (Figure 11) by not specifying how to formulate the answer. We have omitted the prompts for our Agentic method from this paper due to the numerous large prompts used in its different phases, but they will be included in the open-source code release. Furthermore, we deliberately excluded textual descriptions of the images from the prompts. This ensures that the questions can only be answered using the conversational context, thereby testing the model's ability to form an accurate common ground representation based on have has been discussed.

F License

All tools and models used in this research were employed under licenses that permit academic use. Specifically, we utilized the Huggingface TRL library under the Apache-2.0 license. The Llama 3.1, Gemma 2, and Qwen models were used in accordance with the Llama 3.1 Community License Agreement, Google's usage license, and the Apache-2.0 license respectively.

[00:15] **B**: hello

[00:29] **B**: I've found a child's room.

[00:31] **A**: i am in a child

[00:34] **A**: 's room too

[00:42] **A**: blue comforter and teddy bear?

[00:43] **B**: nope.

[00:48] **B**: There are two beds. with a purplish mattress, and a biege comforter across each

[00:57] **A**: oh i can look for that one

[01:03] **B**: okay.

[01:14] **B**: there's also a set of pink and white towels hanging by the foot of the bed

[02:01] **B**: there's a small white nightstand between them, with a little lamp on it.

[02:01] **B**: I'm heading to a new location.

[02:01] **A**: i am in the bathroom

[02:12] **A**: where are you at?

[02:16] **B**: I'm outside. Let's move around a bit.

[02:21] **A**: k

...

[08:23] **A**: Was there a teddy bear in that child's room that I visited?

B: Yes there was

Exact Match: Yes

Figure 5: Test example for inferred grounding from Meetup dataset

...

[02:37] **A**: I am in the basement

[02:37] **B**: I'm in a basement.

[02:49] **B**: Mine has a white staircase

[02:54] **A**: No

[03:03] **A**: mine has wooden stair case

[03:21] **B**: Mine has a leather chair. Should I try to move towards you?

[03:35] **A**: Sure

[03:37] **B**: Wooden? What else?

[03:53] **A**: water heater and washer and dryer

[04:08] **A**: a plastic chair and a screen door

[04:08] **A**: I'm heading to a new location.

[04:08] **A**: I have found a staircase

[04:34] **A**: The staircase is on my right and there is a hallway in front of me

[04:58] **A**: Find it?

[05:01] **B**: found the staircase

[05:15] **A**: is it the one with the hallway in front of you?

[05:27] **B**: it is to my east

[05:30] **A**: and the staircase on the right?

[05:38] **B**: its on my left

[05:53] **A**: We're not at the same staircase

...

[07:12] **B**: What was the material of the chair present in the **white staircase basement**?

A: It was a leather chair.

Exact Match: leather

Figure 6: Test example for attributive grounding from Meetup dataset.

...

[09:45] **A**: Okay I see.

[09:46] **A**: Two sharks on the right top

[09:49] **A**: of the image.

[09:50] **A**: They are blue

[09:51] **A**: and gray.

[09:52] **A**: Or

[09:52] **A**: another kind of blue.

[09:53] **B**: The one below.

[09:55] **A**: Yes.

[09:56] **B**: In my

[09:57] **B**: picture it's green, the one on the top is blue. Which color did you say it was yours?

[10:01] **A**: They are both blue and

[10:03] **A**: gray.

[10:03] **B**: Okay.

[10:04] **B**: The down one in mine is green.

[10:07] **A**: Okay.

[10:08] **A**: They are facing the right.

[10:10] **B**: Correct.

[10:10] **A**: Okay and they have

[10:12] **A**: two fins.

[10:15] **B**: Two.

[10:16] **A**: Fins.

[10:16] **B**: Yes.

[10:17] **A**: aa so below the sharks I have three yellow

[10:21] **A**: octopuses.

[10:23] **B**: I have two.

...

[19:02] **A**: What were the colors of the animals right above the octopuses in my image?

B: There were blue and gray sharks above the octopuses in your image.

Exact Match: blue and gray

Figure 7: Test example for spatial grounding from Spot the difference dataset

...

[02:11] **B**: Ahan.

[02:11] **A**: <ahm>

[02:12] **A**: To the left there is a cloud a gray cloud with raindrops underneath.

[02:18] **B**: Oh so it's a white cloud.

[02:21] **A**: Yeah.

[02:21] **B**: And there's no rain drops.

[02:21] **A**: Okay.

...

[10:45] **A**: Let's switch to the next image now.

...

[12:04] **A**: Okay and up in the sky to the right there's <aa> three clouds.

[12:09] **B**: I have four.

[12:10] **B**: <ahm>

[12:10] **A**: Okay.

[12:11] **B**: Two big ones on the right and two little ones <on the> on the left and upper side.

[12:17] **A**: Okay I have three ones and

[12:21] **A**: I have <aa> the the small one is <aa> above the two big ones.

[12:27] **B**: <ahm>

[12:28] **B**: Okay <ahm>.

...

[14:07] **B**: Do you remember the number of clouds in my last image?

A: You had four clouds.

Exact Match: four

Figure 8: Test example for temporal grounding from Spot the difference dataset

You are a strict evaluator. Given the Question, LLM response and the correct response, judge whether the LLM response and the correct response both have the same meaning provided the question.

KEEP IN MIND TO ALWAYS FOLLOW THESE RULES -

1) You have to give 'DIFFERENT' if they are not having the same meaning and 'SAME' if they mean the same.
2) If the words are synonyms then they shall be considered the same.
3) If the response is a negation of the other response then it does not have the same meaning and hence should be 'DIFFERENT'.
4) If correct answer is kitchen but the response is dining room, then it should be 'DIFFERENT' because they are different. Do not reason that they have similar purpose. Just because the place might have similar function does not make it the same.
5) Sometimes the answers from LLM might contain reasoning, but you should only provide the score based on the final answer provided which is at the end of the answer.
6) If the two responses are negations but in different styles, they should still be considered as the 'SAME'. eg. 'do not recall presence of a cat' is the same as 'no cats were there' are the same, the first one is more indirect form of saying it.
7) If the main content of the answers are the same then answer 'SAME'. The style might be different. For example, 'yes there was a car there' and 'a car was there' are the 'SAME' in meaning even though one is more affirmative than the other.

You have to give the answer in the following format:
<reasoning>
(your reasoning here in one or two sentences where you concretely mention the reason why they mean the same or not.)
</reasoning>
<answer>
(your final verdict here)
</answer>

Figure 9: The system prompt for LLM as a judge.

You are participant {current_participant} who is navigating a shared environment with another participant {other_participant}. You are in a virtual space that has lots of different areas. The participants are initially located in different areas.

The goal is for the two participants to locate themselves in the same area. In order to achieve this goal, the participants communicate with one another and describe the area they find themselves in. On the basis of those descriptions, they move to different areas and describe their new areas to the other participant.

The two participants need to find themselves in the same area. In the last utterance, participant {other_participant} asks a question. Think about the question, get all the information required from the dialog as participant {current_participant} and then answer that question as your response. First reason and interpret the question. Identify whose experience you have to refer to {other_participant} or {current_participant} or both. Fully break it down and then give the answer. Reason step by step and then answer.

The final answer should be in the format -
<reasoning>
(your reasoning here which requires analysing the dialog and understanding what are the important information to look at before answering. Take all the important information into consideration step by step in your reasoning.)
</reasoning>
<answer>
(your final answer here. DO NOT REASON HERE!!)
</answer>.

Figure 10: The system prompt for the LLM when provided with the full dialog for Meetup test cases. It was also used for the GRPO trained models.

[FIRST PARAGRAPH SAME AS Figure 10]

The two participants need to find themselves in the same area. In the last utterance, participant {other_participant} asks a question. Answer the question. First understand what it means and then answer as the other user. In other words, if A asked the question then answer as if you are B and vice versa. Remember they start at different locations and try and meet and are may be in different rooms and may have parallel observations which are valid from their perspective. Identify whose experience and observations you have to refer to. First reason and interpret the question. Identify whose experince you have to refer to, then A or B or both fully break it down and then give the answer.

Remember most importantly -

1) Reason step by step and then answer. Example:
Question from A: do you remember the park I visited?
Although B has to answer here, the reference is to A's experience, so check what A observed.
Similarly, For, B: Do you remember the garden i visited? -> A should answer based on what B experienced.

2) Most Crucial: If the question is about "my x" (e.g., "my dining room", "my office"), always refer to the one who asked the question. So:
- "My dining room" (asked by A) → Refer to A's observation of their dining room.
- "My chandelier" (asked by B) → Refer to B's observation of their chandelier.

3) Also, prioritize semantic ownership and pronoun resolution carefully. Don't let "described by B" override the ownership stated in the question. Always prioritize the experience of the person who owns the object/location. If A says "My chandelier," then even if B is answering, the correct grounding is in A's description — because A is the owner. Use the other speaker's description if that's whose experience is referenced in the question.

4) If observations conflict strongly, assume two different entities till it has been resolved. Example conflict: A: "I see one white fridge", B: "I have five yellow fridges" → this means they are in two different kitchens.

Example grounding from dialogue:
00:02:01 — A: I'm in a kitchen and see a fridge
00:02:10 — B: yellow?
00:02:15 — A: No not yellow but white.
00:03:10 - A: Found another kitchen. White walls.
Spatial Grounding: A is in Kitchen A, B is in Kitchen B. A then moves to kitchen C.
Object Grounding: Fridge A (white) in Kitchen A — seen by A, Fridge B (yellow) in Kitchen B — seen by B, Kitchen C has white walls - seen by A.
Temporal Grounding: A first goes to kitchen A, B simultaniously goes to kitchen B. A then move and finds another kitchen C which is the second kitchen that A saw.

The final answer should be in the format -
<reasoning> ... [Like in Figure 10] ... </answer>

Figure 11: The system prompt for in-context learning with full dialog history for Meetup test cases.

Question : Can you please remind me which room is positioned east of the dining room that I visited?

Correct Answer : Attic

Generated Answer : Okay, let's see what the user needs here. They're asking for the next utterance in the conversation based on the provided dialog history. The last message from A is a question: C̈an you remind me which room is positioned east of the dining room that I visited?S̈o, I need to figure out how B would respond.Looking through the conversation history, we need to track where each participant has been. Let's start by recalling the key points:First, Participant B started in a dining room with a red chandelier (timestamp [00:00:27]). Then, A moved to the east and mentioned being in an attic ([00:01:04]), but then later described being in a room with a curved ceiling, ceiling fan, and triangular shape, possibly an attic again. At [00:02:43], A mentions finding a jacuzzi with beach chairs, which seems like a bathroom or maybe a spa room. But then moves to a pool area ([00:03:31] mentions brown tiles and pool). Meanwhile, B went to a street ([00:03:48]) and then an attic with a bed and closet, which was different from A's attic without a bed. Eventually, B ended up in the room A was referring to, which had a window and a wooden table on the right, plus a 2007 date ([00:05:04]).The question now is about the room east of the original dining room. From earlier messages, when A first said they were in the dining room with a red chandelier, then moved east to what they thought was an attic. However, after moving, the next image B received was a library/museum building ([00:01:16 Image B]). Wait, actually, the timestamps show that after A says they're heading east, the next Image B is the library building. Hmm, but maybe the movement directions correspond to changing rooms. Since in virtual spaces, moving east could lead them to adjacent rooms. Wait, the initial dining room was B's starting point. When A was in the dining room too? At [00:01:04], A says ïn dining room with red chandelier,s̈o they were both in the same dining room at some point. But since they needed to get to the same room, perhaps moving east from there led A to a different room. The Image B after A's move to east showed the library/museum building. So maybe moving east from the dining room leads to the library? Because the next image after A moving east was the library. So when A asks which room is east of the dining room, B would refer to that library or museum.But looking at the latest interactions, B is currently in a room with a window, wooden table on the right, and a 2007 date. That room's description ([00:05:04]) is an empty room with a dining table, chairs, desk, laptop, etc. Maybe that's a different area. Alternatively, B might remember previous movements. Alternatively, maybe the library was the room east of dining because after A went east, the next image was the library. Therefore, B would say something like ẗhe libraryör m̈useum building.̈ Also, considering that in the final messages, they ended up in a room with the 2007 date, which might not be the library, but another spot. So the response should be B telling A that the room east of the dining room is the library or museum. Let me check the exact description again. The Image B after A moved east was: Ïarge, unique building with a distinctive dome...library or museum.̈ So B should mention that. Hence, the answer is probably pointing out the library/museum as the eastern room.</think>[00:05:42] B: The room east of the dining room looked like a library or museum with a big dome. check the building with cars parked out front and traffic lights

Figure 12: Failure case from Qwen-QWQ where it produces a long answer despite reasoning in it's <reasoning> tags separately