Distributed Partial Information Puzzles: Examining Common Ground Construction Under Epistemic Asymmetry

Yifan Zhu[1], **Mariah Bradford**[2], **Kenneth Lai**[1], **Timothy Obiso**[1]
Videep Venkatesha[2], **James Pustejovsky**[1], **and Nikhil Krishnaswamy**[2]
[1]Brandeis University, [2]Colorado State University
415 South St, Waltham, MA 02453, Fort Collins, CO 80523
{zhuyifan, jamesp}@brandeis.edu,
mbrad@rams.colostate.edu, nkrishna@colostate.edu

Abstract

Establishing *common ground*, a shared set of beliefs and mutually recognized facts, is fundamental to collaboration, yet remains a challenge for current AI systems, especially in multimodal, multiparty settings, where the collaborators bring different information to the table. We introduce the **Distributed Partial Information Puzzle (DPIP)**, a collaborative construction task that elicits rich multimodal communication under epistemic asymmetry. We present a multimodal dataset of these interactions, annotated and temporally aligned across speech, gesture, and action modalities to support reasoning over propositional content and belief dynamics. We then evaluate two paradigms for modeling common ground (CG): (1) state-of-the-art large language models (LLMs), prompted to infer shared beliefs from multimodal updates, and (2) an axiomatic pipeline grounded in Dynamic Epistemic Logic (DEL) that incrementally performs the same task. Results on the annotated DPIP data indicate that it poses a challenge to modern LLMs' abilities to track both task progression and belief state.

Keywords: common ground, collaborative tasks, multimodality

1. Introduction

Effective collaboration between humans depends on establishing *common ground*—a set of shared beliefs and agreed-upon facts in a task context. Doing so enables them to align perspectives, coordinate actions, and achieve joint goals (Clark and Brennan, 1991; Traum, 1994; Asher, 1998; Dillenbourg and Traum, 2006). However, humans rarely begin a task with the exact same background, perspective, or information, meaning that construction of common ground requires them to resolve *epistemic asymmetry* (Zhou et al., 2024a) and effectively communicate and make inferences about partial or privately-held information that their interlocutors may have. These complexities are amplified in *co-situated interaction*, which involves significant multimodal components, as humans communicate through both verbal cues (speech) and non-verbal cues such as gestures and actions in context (Cassell, 2000; Wahlster, 2006; Foster, 2007; Kopp and Wachsmuth, 2010; Marshall and Hornecker, 2013; Schaffer and Reithinger, 2019). The challenges are multiplied further when collaborations involve more than just two parties, as in this case each individual has to integrate and coordinate with multiple information sources. Common ground, dialogue state tracking, collaboration, theory of mind, and related issues have been the objects of intense study and interest within dialogue research, NLP, and AI more generally. However, data that enables the robust study of these topics in multiparty, co-situated, epistemic-asymmetric collaboration, and illuminates the challenges it poses to state-of-the-art (SOTA) AI systems remains sparse.

In this paper, we introduce a task we call a "Distributed Partial Information Puzzle" (DPIP), which is realized as a type of collaborative construction task. In the task, three "directors" are each given individual partial information about a goal structure and must collaboratively instruct one "builder" to construct a single contiguous structure that is that consistent with all three pieces of individual information (Fig. 1). We present a dataset of DPIP interactions annotated with speech transcriptions, gestures, and actions in context, as well as the propositional information communicated through each of these modalities. Further, we investigate the ability of SOTA LLMs to track and reason about information exchange in this context, to establish the challenges this type of task and domain poses to SOTA AI systems. Our annotated data and evaluation code, including prompts, are available through `https://doi.org/10.5281/zenodo.18626419`, and will be a continually updated community resource.

2. Related Work

Common ground has long been studied in linguistics, philosophy, and computational dialogue research as the process by which interlocutors establish shared understanding through communicative acts (Clark and Brennan, 1991; Traum, 1994). Later work formalizes its role within discourse structure and collaborative learning (Asher, 1998; Dillenbourg and Traum, 2006), viewing common ground

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 4974–4987
11–16 May 2026.

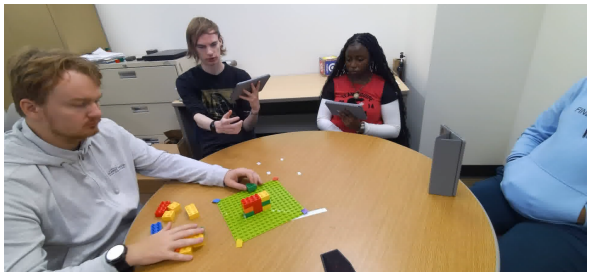

Figure 1: A *builder* and 3 *directors* participating in the DPIP task with a partially-completed structure on the table. Director 1 (second from left) is indicating the position of a block using a gesture as well as a speech act.

as an evolving set of shared beliefs that enable coordination and joint action.

Because human communication is inherently multimodal, grounding extends beyond speech. Prosody conveys intention beyond propositional content (Shriberg and Stolcke, 2004; Chiu, 2008), gaze signals attentional focus and shared reference (Fathi et al., 2012; Huang et al., 2015), and gesture complements or substitutes for verbal acts to enrich meaning (Komorowska, 2017; Brutti et al., 2022). Together these modalities form the substrate through which common ground dynamically emerges.

Situated language understanding has been explored previously in physical and virtual environments (Zarrieß et al., 2016; Krishnaswamy et al., 2017; Pustejovsky et al., 2017; Krishnaswamy and Pustejovsky, 2018; Bisk et al., 2018; Chen et al., 2019; Suhr et al., 2019; Krishnaswamy et al., 2020, 2022; Thomason et al., 2020). Related work models how interlocutors infer intentions and beliefs (Chai et al., 2014; Imai et al., 2025), track dialogue state (Williams et al., 2016), and display theory-of-mind–like reasoning (Sap et al., 2022; Ullman, 2023; Nath et al., 2025; Sicilia and Alikhani, 2025). However, few datasets capture all three challenges—multiparty interaction, co-situatedness, and epistemic asymmetry—simultaneously.

Among existing corpora, the HCRC Map Task (Anderson et al., 1991) and EGGNOG (Wang et al., 2017) model two-party co-situated collaboration; *MindCraft* (Bara et al., 2021) introduces partial observability in a virtual world; and *DeliData* (Karadzhov et al., 2023) provides multiparty deliberation but no embodied action. The Weights Task (Khebour et al., 2024a,b, 2025; VanderHoeven et al., 2025) combines gesture and action in a co-situated setting, yet its full observability constrains the epistemic richness of belief negotiation. Our *DPIP* task complements these efforts by uniting all three aspects—multimodality, multi-party interaction, and asymmetric knowledge—within a collaborative construction domain.

Work on competitive or deceptive multi-agent settings also examines reasoning under partial information. For example, Zhang et al. (2025b) develop *Werewolf* agents that infer hidden roles and intentions, extending theory-of-mind frameworks (Baron-Cohen, 1997; Goldman, 2012; Pustejovsky and Krishnaswamy, 2021). In contrast, the DPIP task is cooperative: participants integrate truthful but partial evidence to build a shared representation of a physical goal. While Zhang et al. (2025b) emphasize emotional and deceptive cues, our focus is on *task-oriented multimodal grounding*, where gesture and action encode propositional information central to problem-solving.

Complementary work by Sileo and Lernould (2023) employs Dynamic Epistemic Logic (DEL) to evaluate theory-of-mind reasoning in text-based scenarios. Our approach extends this logic-driven perspective to a co-situated multimodal setting, where updates to belief states must be inferred jointly from speech, gesture, and action. Large language models (LLMs) show strong results in discourse reasoning and implicit inference (Fei et al., 2024; Niu et al., 2024; Zhang et al., 2025a), yet to our knowledge, no prior study has systematically evaluated whether LLMs can infer and track *common ground* in collaborative, multimodal, and partially-observable environments such as DPIP.

3. Task Description and Dataset

3.1. Distributed Partial Information Puzzle

The *Distributed Partial Information Puzzle* (DPIP), is a collaborative problem-solving paradigm in which essential information is *distributed* among participants, making communication a requirement for task success. Unlike traditional shared-information tasks, DPIPs deliberately prevent any single participant from solving the puzzle independently, thereby enforcing communication as the primary means of coordination.

Our DPIP task is a collaborative construction task, performed in groups of four, wherein each of three "directors" is given different partial information about a structure and have to collaboratively instruct one "builder" to build a single contiguous structure out of large Lego blocks, that is consistent with all pieces of information. See Fig. 1. The task is complete when all three directors agree that the structure that the builder has built is consistent with all three of their views. The group is *successful* if that structure actually matches the originally-generated side views.

In our data collection implementation, the goal structure has dimensions of 3 (width) $\times$ 3 (depth) $\times$

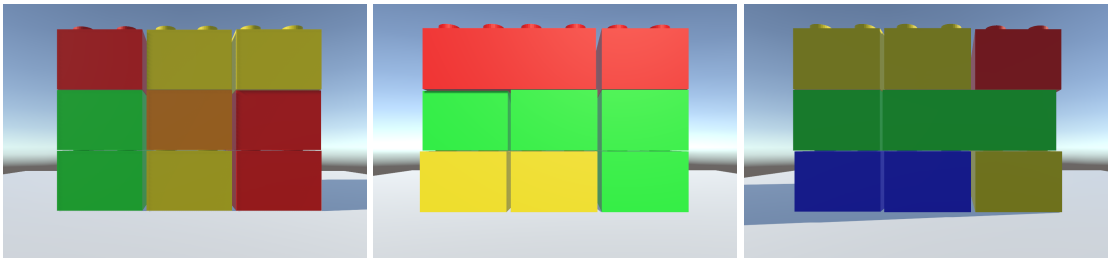

Figure 2: 3 individual side views of the same complete structure, each assigned to a director.

3 (height), where the unit of the height dimension is a layer of blocks and the unit in the width and depth dimensions is *2* of the pegs atop a Lego block (thus a single "square" Lego is considered to be $1 \times 1 \times 1$ — see Fig. 2). The structure is composed entirely of square or rectangular blocks, and there are no gaps in between blocks in any of the walls. In a variant of the task, the structure footprint is extended to $4 \times 4 \times 3$, and blocks with curved sides and gaps in the structure are allowed—however, our annotations and evaluation extend only to the first variant described above.

Epistemic asymmetry is established by distributing distinct 2D side views to the three directors, while only the builder is allowed to manipulate the blocks. The ground truth goal structure is procedurally computed and rendered in the Unity game engine, and screenshots of the walls (e.g., Fig. 2) are distributed to the directors. The directors must collectively guide the builder to reconstruct the goal structure, and in the process must necessarily reconcile discrepancies across their partial perspectives, such as inferring when another participant is expressing or doing something that is consistent or inconsistent with their own private information. Each director holds unique spatial information, and as only the builder can touch the blocks, no group can complete the task without contributions from all members. This setup simulates teams with members with different backgrounds and expertise.

This setup enables the study of epistemic group states and common ground formation, as participants must externalize private visual knowledge through language and gestures (and in the case of the builder, actions). The task further demands precise spatial reasoning and description, and even mental simulation (Goldman, 2006), since the builder must form a mental model of the structure without visual reference.

3.2. Data

The annotated DPIP Lego dataset consists of 10 audiovisual recordings of groups of four performing the task described above (Fig. 1). Each director was provided with a tablet showing a different side view of the target structure (e.g. front, left, or right side views—Fig. 2) without knowing what side they are seeing. Participants were seated around a table and their interactions were captured using 3 Microsoft Kinect Azure cameras to log different angles of the task. Audio was recorded on a single conference-style tabletop microphone. The study was approved by university Institutional Review Boards (IRBs) and participants received USD 15.00 each.

Table 1 provides descriptive statistics over the 10 groups. One of the groups, Group 7, was an extreme outlier in terms of task completion time (Z score of 14.7), and did not successfully complete the task, and so we provide statistics with and without this group, which was held out from the main evaluation for a secondary experiment (Sec. 5.2). Of those subjects who completely filled out a demographic survey at study time, the average age was 23 years (SD = 3.58). Ten subjects were female and 21 were male. Eleven were Caucasian, 18 were Asian, and 2 reported other ethnicity. Fourteen were native English speakers, 11 were native Telugu speakers, 6 spoke other native languages.[1] The task was conducted entirely in English.

4. Annotation Pipeline

As task-relevant propositional content, as well as epistemic positioning, may be communicated through any of three primary modalities—*action*, *speech*, and *gesture*—these form the empirical basis for modeling multimodal, multiparty communication in the DPIP task. Following VanderHoeven et al. (2025), we treat gaze as an implicit perceptual cue, while gesture is considered a distinct communicative modality complementary to speech. Each constituent modality is dually-annotated by independent human annotators, and inter-annotator agreement (IAA) scores are computed to assess reliability. When performing the annotations, all human annotators received access to the video recordings of the task, specifically to the central angle of the 3 camera angles recorded. Our annotation extends the process initially described in Zhu et al. (2025).

[1] Numbers are not reported for subjects who did not respond to questions on the demographic survey. Other native languages in the subject pool included Indonesian, Kannada, Malayalam, Spanish, Tamil, and Yoruba.

	Ave	SD	Ave*	SD*	Min	Max	Max*	Total
Task Time (s.)	703.53	903.48	423.52	190.43	159.82	3223.60	800.40	7035.32
# Utterances	311.20	397.90	190.33	117.34	38	1399	350	3112
# Actions	47.90	48.34	33.00	11.48	19	182	53	479
# Gestures	27.80	19.74	22.56	11.35	7	75	37	278
# Props	30.70	15.54	26.44	8.23	18	69	40	307
# Tokens	1302.70	1442.70	860.44	375.71	350	5283	1517	13027

Table 1: Summary statistics of 10 annotated DPIP Lego task groups. *Indicates statistics excluding outlier group (Sec. 3.2).

4.1. Speech Annotations

For the speech modality, we employ the Whisper ASR model (Radford et al., 2023) for automatic transcription. The automatic transcriptions are verified and corrected by a human who reviewed the transcripts by watching the associated videos. Annotators then annotated the resulting post-corrected utterances with the *propositions* they express, in the form of relational information among blocks referenced in dialogue.

Each resulting proposition encodes four elements: the *timestamp*, *speaker identity*, *relative position* between two blocks, and the *side* information corresponding to the director's assigned perspective. For example, if Director 2 says, "a small blue block is next to a green block from my side," the extracted proposition is represented as: `(nextto(BlueShortBlock, GreenShortBlock), D2's side)`.

4.2. Action Annotations

The action space of the DPIP Lego task is limited to block manipulations by the builder: *put* (a block appears at a location on the board), *remove* (a block disappears from a location), and *move* (a sequential *remove-put* operation). Therefore to derive gold-standard action sequences, we first had annotators log the evolution of blocks on the board in a 3D Structure Annotation Tool (SAT). While watching the video, annotators reconstruct the evolving board state by placing color- and shape-specified blocks onto a grid. The SAT interface records these placements as JSON logs containing block identifiers, spatial coordinates, and timestamps, while a 3D visualization panel provides a real-time way to verify annotation against the video recording.

From these logs, we then deterministically extracted discrete *put*/*remove*/*move* actions according to the above definitions. Locations are first encoded as absolute coordinates and subsequently transformed into task-relevant spatial relation predicates (e.g., on, $leftof$, etc.).

Each action annotation integrates four key elements: the *timestamp*, *action type*, a *unique object identifier with its block attributes (color and shape)* automatically generated by the system, and *side* information added by human annotators. For example, the annotation $put(rs1,\ on(base)),\ D1's\ side$ indicates that a `red short block` with the identifier `1` (unique among blocks of the same properties) is introduced into the structure from `Director 1`'s perspective, placed on the base layer.

4.3. Gesture Annotations

Participant gestures are annotated using Gesture Abstract Meaning Representation (GAMR; Brutti et al. (2022)), a framework designed to capture gesture semantics in ELAN (Wittenburg et al., 2006). Gestures may be deictic, iconic, or emblematic, indicating structural descriptions (e.g., *side by side*), block attributes (e.g., *square*, *rectangular*, *curved*), or actions (e.g., *bring forward/backward*), showing epistemic position such as agreement (e.g., *nodding*).

Annotators with previous expertise in GAMR watched the video while annotating, to ensure annotations were time-stamped for alignment with utterances and actions, and are produced with access to object IDs, allowing gestures that reference specific objects to be concretely recorded. Annotators recorded when they saw gestures occurring and recorded their meaning. For example, the following GAMR annotation:

```
(d / deixis-GA
    :ARG0 (d1 / director-1)
    :ARG1 (bs1 / blue-square-1)
    :ARG2 (g / group))
```

indicates that Director 1 (`ARG0`) is performing a pointing gesture toward a specific block, `blue-square-1` (`ARG1`), with the intended referent of the gesture being the group (`ARG2`).

4.4. Inter-Annotator Agreement

From the two annotation sets produced for each modality, inter-annotator agreement (IAA) was computed to assess reliability. For the speech and structure modalities, we used Cohen's Kappa coefficient (Cohen, 1960) as the evaluation metric. In the speech modality, agreement was assessed

Group	3	4	5	7	9	10	11	12	13	14	Ave
Speech	1.000	0.979	1.000	0.990	1.000	1.000	0.832	1.000	0.908	1.000	0.971
Structure	0.826	0.943	0.901	0.939	1.000	0.587	0.924	0.890	0.916	1.000	0.893
Gesture (union)	0.578	0.540	0.638	0.507	0.568	0.542	0.526	0.378	0.452	0.222	0.495
Gesture (intersection)	0.944	0.882	0.839	0.854	0.853	0.795	0.793	0.806	0.886	0.800	0.845

Table 2: Inter-annotator agreement (IAA) scores by modality across 10 groups.

on both the number and the propositional content of relational expressions extracted from speech acts. For the structure modality, agreement was calculated by comparing the human-annotated 3D structures, represented in xyz-coordinates along with the corresponding layer indices derived from the task video. Since action annotations were deterministically extracted from structure annotations, IAA over the structure annotations serves as validation of the action annotations as well. For the gesture modality, we report two scores. Annotators needed to record both when they saw a gesture occurring and what they interpreted it to mean. Since the two annotators may be more or less conservative with their identification of gesture *occurrence*, independent of interpretation, the two annotators may report different numbers of gestures in a video. This is similar to the problem of *mention* over- or under-annotation reported in coreference corpora (Bugert et al., 2021). Therefore we report IAA scores over the union of the sets of gestures annotated by both annotators, and over the intersection only. For each annotator, following Lai et al. (2024), we concatenate all or some of their GAMRs into a single AMR: to calculate union scores we include all GAMRs, while for intersection scores we only include those gestures annotated by both annotators. We then calculate SMATCH F1 scores (Cai and Knight, 2013) between the two annotators' resulting AMRs.

Table 2 presents the inter-annotator agreement (IAA) scores across three modalities for ten groups. Speech annotations exhibit consistently high levels of agreement, with an average Cohen's Kappa score of 0.971. Structure annotations generally demonstrate high reliability, with many groups achieving strong consistency, though some variability is observed. This variability arises when the same block is annotated with different xy-coordinates by different annotators; for example, in Group 10 such discrepancies led to divergent representations of the same structure and consequently a lower agreement score. For the gesture modality, we see that IAA over the union is significantly lower than IAA over the intersection, due to the occurrence identification differences, although the numbers still signal moderate to substantial agreement (avg. of 0.495). IAA over the intersection is uniformly high (avg. 0.845), signaling reliable gesture annotation. Therefore we use the union of the reported annotations since we are concerned with gesture interpretation rather than mere occurrence identification, under the assumption that, when given the same set of identified gestures, the annotators display high agreement. We likewise assume that over-identified gesture occurrences that do not contribute to multimodal interpretation of meaning do not align with meaningful features in other modalities, and would therefore be filtered out in the next step (Sec. 4.5).

4.5. Cross-Modality Alignment

In isolation, the individual modalities may not fully capture the instructions communicated by the directors, their expressed beliefs or personal grounding, or the builder's interpretation thereof. Therefore, we standardized and aligned the propositional content expressed through each primary modality to achieve a complete representation.

Each builder action (e.g., Builder: $(rs1,\ on(gs2),\ layer\ 2, D1's\ side)$) is parsed into a structured event specifying the acting block (top), the supporting block (base), the spatial relation (on or below), the layer number, and the speaker's perspective. These structured action events establish mappings between actions and linguistic or gestural communications. For every speech or gesture interval, the pipeline retrieves temporally adjacent actions within configurable windows, enabling utterances such as "put the green on the red" and corresponding pointing gestures in the correct physical context.

Verbal and gestural propositions are normalized and symbolically grounded. Color-shape terms in speech (e.g., $RedShort, YellowLong$) are automatically replaced with concrete block identifiers (e.g., $rs2$, $yl1$) inferred from future action pairs. In gesture annotations, gesturer and gesture content are replaced with the participant ID and the nearest object, including its construction layer.

All layer-aware propositions are then serialized into temporally aligned outputs for each modality and merged with emblematic gesture annotations that convey disagreement or confirmation.

5. Evaluation

Our experimental setup evaluates the abilities of both (1) LLMs and (2) an axiomatic processing pipeline, at inferring what the collaboratively-built structure should look like, given the observable interaction between the participants, and at inferring their apparent shared beliefs. The state of the structure, as derived from the structure and action annotations, represents an objective observable ground truth that can be taken to be ultimately reflective of what the accepted common ground of the group is, after they have negotiated and disambiguated the placement of each block.

5.1. Common Ground (CG) Calculation

The temporally aligned multimodal propositions form the basis of our belief-tracking framework. For each speaker, *belief states* were annotated by two annotators over each dialogue, and they were encoded as to whether each relation is currently accepted, doubted, or negated. The dual annotations yielded an average inter-annotator agreement score of $\kappa = 0.809$. The annotated propositions, together with acceptance and doubt labels, are temporally aligned (Sec. 4.5).

We then implemented a common-ground inference module that axiomatically predicts how shared knowledge and mutual belief evolve within and across turns. We specify the following axioms, inspired by Bolander (2014); Pustejovsky and Krishnaswamy (2021) and Grice (1989) and grounded in a simplified model of evidence-based dynamic epistemic logic (EB-DEL) following van Benthem et al. (2014) and Pacuit (2017), and used by Khebour et al. (2024b) and Pustejovsky and Zhu (2024):

1. **Seeing is Believing**: perceptual context directly updates belief;
2. **Acting is Believing**: embodied action reveals intention;
3. **Saying is Believing**: language communicates epistemic state.

These axioms leverage the primary available modalities of language, gesture, and action, as available in the DPIP task. As belief states evolve based on these axioms, we incrementally construct and update sets of participants who share the same positive stance toward a given proposition. Whenever at least two individuals concurrently accept a relation, a CG(Common Ground) label (e.g., $CG_{\{D1,D2,Builder\}} : on(gs3,\ rs2,\ layer\ 2)$) is generated. Doubts temporarily remove the doubting participant from that CG(Common Ground) set, while negations delete the shared belief entirely—analogous to removing a physical block from the construction. Multi-clause acceptances and subset relations are also handled, allowing higher-level CG(Common Ground) formations to emerge from overlapping atomic agreements. This process yields a structured, dynamically updated representation of the group's common ground under the above axiomatic assumptions.

Those CG calculations are then used to define the turn boundaries, where each turn represents a complete episode of belief formation. A turn is bounded by two successive updates in the shared common ground—the set of task-relevant beliefs mutually recognized by the group. Rows containing explicit common-ground (CG) annotations mark these boundaries, so that each turn spans the interval between consecutive CG labels, encompassing all intermediate multimodal events (verbal, gestural, and physical) that collectively constitute the evidential context leading to a belief update. The same turn boundaries were used to segment the dialogue sequences provided to the LLMs for evaluation, enabling direct comparison between human-annotated and model-predicted belief dynamics.

5.2. Experiments

The following experiments were performed:

1. *Structure prediction from actions* (**Action** $\rightarrow$ **Structure**): We input the annotated actions taken during each turn, and prompt an LLM to output the status of the structure after that turn, as the group proceeds toward task completion. This provides a strict measure of how actions can indicate *task progress*, and serves as a baseline to demonstrate how well LLMs can track actions taken on the board over time.
2. *Structure prediction from aligned annotations* (**Aligned** $\rightarrow$ **Structure**): We input annotations from *all* modalities without fusion for each turn, and prompt an LLM to output the status of the structure after that turn. This measures how much the additional modalities that inform the block placements help or hurt LLMs' structure prediction ability.
3. *Structure prediction from axiomatic CG* (**CGC** $\rightarrow$ **Structure**): We run the CG inference (Sec. 5.1) over the aligned and merged multimodal annotations to generate a predicted set of beliefs about block positions. These beliefs are then used to specify what the structure would be in the case that all the predicted beliefs were true. This demonstrates how well an axiomatic prediction of common ground matches the reality of what the group actually builds.

Because each turn represents an episode of belief formation, sharing, and negotiation, and because

the structure annotations demonstrate a high reliability, the physical board state after each turn represents an objective "observable ground truth" of the common ground between the group members, and so the outputs of these experiments are compared to the actual structure state.

4. *Common ground prediction from aligned annotations* (**Aligned** $\rightarrow$ **CGC**): We input aligned annotations for each turn to an LLM and, rather than prompt it to predict the structure state, we prompt it to predict the *common ground* (shared belief set) of the group, and compare this output to the CG calculated axiomatically over the same annotations. This provides an indication of how much LLMs and an axiomatic approach overlap or differ in their predictions, and serves as an intrinsic evaluation of the CG calculation (CGC) axioms.

We conduct experiments using Qwen3-4B-Instruct-2507 (Yang et al., 2025) and Llama-3.2-3B-Instruct (Dubey et al., 2024), to assess multiple models on a limited compute budget. For a larger/stronger model comparison, we also evaluated GPT-5-mini (OpenAI, 2025), and, when necessary, GPT-5 (OpenAI, 2025). The latter was employed when GPT-5-mini omitted outputs when performing structured reasoning or JSON generation.

Our primary metric is the Dice Similarity Coefficient (DSC; Sørensen (1948); Dice (1945)), which measures the overlap between predicted and ground-truth sets with sensitivity to the size of the set. The Action $\rightarrow$ Structure task evaluates an LLM on its ability to produce each director's view as a 3x3 list of blocks, which is then translated to a set of configurational relations. For the Aligned $\rightarrow$ Structure prediction experiments, the LLM produces a set of spatial relations between blocks, whereas for the axiomatic CG prediction experiment, the LLM output is the set of predicted shared beliefs of the group. We report the *Average* of DSC after each turn (reflecting local agreement on per-turn state changes) and DSC at the *Global* dialogue level (agreement between the final predicted state and final actual state).

As mentioned in Sec. 3.2, in the outlier group, Group 7, the participants did not successfully build a structure consistent with all three directors' side views. Thus, it provided an opportunity to study LLM performance and the contribution of multimodal annotations in a setting decoupled from task success as a completion state. In this case, since intermediate actions did *not* lead to successful completion, we evaluate only on the aligned and axiomatic inputs (i.e., Experiments 2–4 as described above), to examine multimodal alignment and common ground on an example of task failure.

6. Results and Analysis

Table 3 shows results in all four experiments for all methods and metrics over 9 of the 10 annotated DPIP groups.

In the first experiment, *Actions* $\rightarrow$ *Structure*, GPT-5 is the best performer, with a statistically significant advantage over the next-best model, Qwen, according to group-wise Average DSCs ($p \approx 0.004$ by a Wilcoxon signed-rank test). However, Global DSC scores are worse for all models than Average DSCs, suggesting that when given only information about actions, LLMs are better at inferring structure at the turn level than over entire dialogues. This aligns with previous results showing that LLMs and common methods with which they are aligned for generation struggle with longer multi-turn interactions and reasoning (Zhou et al., 2024b).

In the *Actions* $\rightarrow$ *Structure* experiment, LLMs were provided with only one modality, and so it can serve as an approximate baseline to assess how adding other modalities helps or hurts inference ability. When provided with aligned annotations across all available modalities (*Aligned* $\rightarrow$ *Structure*), Qwen is the top performer (statistically significant advantage over nearest competitor Llama, $p \approx 0.004$). Here, in all cases between the two experiments, Average step-wise DSC goes down but Global DSC goes up. This may reflect added context via other modalities. If so, the additional context seems more useful when looking at entire dialogues together rather than segmented turns, suggesting that with sufficient context, additional modalities help, but without it, they simply add noise. Interestingly in this case, GPT-5-mini/GPT-5 goes from the best-performing LLM to the worst, predicting exactly none of the correct structure for Groups 4 and 5. It seems surprisingly bad at processing the DPIP task data despite its formidable reputation, and also has a very high standard deviation compared to raw per-group scores.

CGC $\rightarrow$ *Structure*, the deterministic axiomatic approach to inferring common ground, is sometimes surprisingly effective at predicting the structure state. CG axioms achieve higher mean Average and Global DSC scores at structure prediction than GPT-5-mini does when given access to all modalities (non-significant medium difference, $p \approx 0.2$).

Group 9 is consistently among the highest-scoring groups, and seems to be relatively easy for both LLMs and CGC to process.

Aligned $\rightarrow$ *CGC* computes the overlap between LLMs' predictions of common ground and common ground calculated axiomatically. We find that overall overlap is quite low, indicating the LLMs frequently infer different belief states from multimodal dialogues than are calculated axiomatically.

Actions → Structure	**Group**	3	4	5	9	10	11	12	13	14	μ	**SD**
Llama 3.2-3B	Average	0.173	0.202	0.170	0.275	0.157	0.402	0.259	0.134	0.200	0.219	0.078
	Global	0.000	0.037	0.056	0.056	0.000	0.148	0.074	0.037	0.037	0.049	0.042
Qwen3-4B	Average	0.185	0.193	0.232	0.457	0.235	0.392	0.276	0.251	0.264	0.276	0.085
	Global	0.037	0.000	0.148	0.185	0.074	0.148	0.148	0.037	0.185	0.107	0.066
GPT-5-mini / GPT-5	Average	0.343	0.270	0.308	0.633	0.254	0.561	0.333	0.362	0.378	0.382	0.122
	Global	0.133	0.148	0.000	0.482	0.000	0.222	0.000	0.000	0.444	0.159	0.180
Aligned → Structure	**Group**	3	4	5	9	10	11	12	13	14	μ	**SD**
Llama 3.2-3B	Average	0.110	0.097	0.155	0.198	0.105	0.099	0.080	0.106	0.115	0.118	0.036
	Global	0.439	0.453	0.419	0.718	0.556	0.348	0.375	0.457	0.474	0.471	0.110
Qwen3-4B	Average	0.148	0.153	0.157	0.146	0.131	0.191	0.098	0.098	0.132	0.139	0.029
	Global	0.811	0.704	0.545	0.903	0.667	0.703	0.541	0.516	0.621	0.668	0.130
GPT-5-mini / GPT-5	Average	0.030	0.000	0.000	0.075	0.017	0.025	0.010	0.065	0.046	0.029	0.027
	Global	0.261	0.000	0.000	0.696	0.091	0.174	0.125	0.556	0.348	0.250	0.233
CGC → Structure	**Group**	3	4	5	9	10	11	12	13	14	μ	**SD**
	Average	0.121	0.073	0.095	0.057	0.036	0.012	0.045	0.081	0.037	0.062	0.034
	Global	0.519	0.364	0.455	0.889	0.357	0.239	0.197	0.167	0.271	0.369	0.215
Aligned → CGC	**Group**	3	4	5	9	10	11	12	13	14	μ	**SD**
Llama 3.2-3B	Average	0.167	0.242	0.400	0.185	0.000	0.016	0.000	0.098	0.153	0.140	0.131
	Global	0.190	0.353	0.800	0.400	0.000	0.062	0.000	0.073	0.108	0.221	0.261
Qwen3-4B	Average	0.167	0.000	0.000	0.074	0.083	0.000	0.104	0.275	0.334	0.115	0.121
	Global	0.286	0.000	0.000	0.182	0.250	0.000	0.073	0.296	0.158	0.138	0.124
GPT-5-mini / GPT-5	Average	0.375	0.515	0.200	0.333	0.278	0.000	0.292	0.407	0.195	0.288	0.148
	Global	0.435	0.571	0.500	0.429	0.727	0.000	0.214	0.281	0.108	0.363	0.232

Table 3: Results in four experimental conditions for all models, methods, and metrics (DSC).

Aligned → Structure	**Llama 3.2-3B**	**Qwen3 -4B**	**GPT-5-mini / GPT-5**
Average	0.040	0.066	0.007
Global	0.182	0.298	0.055
CGC → Structure			
Average		0.250	
Global		0.226	
Aligned → CGC	**Llama 3.2-3B**	**Qwen3 -4B**	**GPT-5-mini / GPT-5**
Average	0.334	1.000	1.000
Global	0.500	1.000	1.000

Table 4: DSC Results in three experimental conditions for outlier group, Group 7.

In many cases the DSC is 0, indicating completely disjunct sets predicted by the two methods. Interestingly, on Group 9, on which both LLMs and CGC achieve high global DSC on structure prediction, the LLM/CGC overlap never exceeds 0.5, indicating that even on an apparently "easy" dialogue in this task, LLMs and CG axioms are retrieving at least very different aspects of the interaction.

Overall there is a high variance in performance across groups, similar to what is observed by Khebour et al. (2024b) in their common ground tracking task in a multimodal dialogue. This convergent result further highlights the diverse ways in which collaborative groups express belief convergence and epistemic positioning.

Table 4 shows performance on Group 7, the outlier group that failed to complete the task correctly. In the *Aligned* → *Structure* and *CGC* → *Structure* experiments, we see results that fall within the distribution of results for the other groups. However, a very interesting result emerged when evaluating *Aligned* → *CGC*: both Qwen and GPT perfectly infer the axiomatically-calculated common ground. This result seems sharply divergent from expectations of what should be a harder inference problem (i.e., common ground in a group that fails at the task). However, a closer examination of the group dynamics explains this outcome: widespread confusion about the task, goals, and available information leaves little shared common ground that can be axiomatically extracted. The LLMs, when presented with the aligned annotations, infer small or even empty belief sets, which matches the lack of common ground displayed by the group. They appear able to correctly detect when there is a *lack* of common ground displayed in the group dialogues and annotations. Conversely, as indicated by *Aligned*

$\rightarrow$ *CGC* over other groups, when there exists substantive common ground in a group, as calculated over the annotations, SOTA LLMs remained challenged by the task of inferring what its contents are.

7. Conclusion

In this paper, we introduced a challenging new task, the Distributed Partial Information Puzzle (DPIP), realized as a collaborative construction task under partially-observable conditions. This task and data uniquely combines a multiparty, co-situated interaction with rich belief dynamics induced by the partial information setting, which simulates bringing together teams composed of members with different backgrounds and expertise. We performed rich multimodal annotations over the data, and evaluated 3 modern LLMs as well as an axiomatic belief extraction pipeline on the task of predicting the structure being built based on the participants' utterances, gestures, and/or actions in context. Our results show how challenging this task setting and data is for the modeling of belief state, group dynamics, and task progressing by state of the art systems, establishing a challenging new task and benchmark for multimodal dialogue research, including common ground tracking (Khebour et al., 2024b; Tu et al., 2024; VanderHoeven et al., 2025), but also for tasks such as modeling theory of mind (Bara et al., 2021; Sileo and Lernould, 2023) in small groups, or spatiotemporal reasoning.

Limitations

Although 3 camera angles were recorded during collection, annotations were conducted from a single camera angle for each session (the central of the three). As a result, it is possible that at times part of the structure could be occluded from view, e.g., when participants placed blocks behind existing ones. This could potentially introduce uncertainty while annotating the spatial relationships of the blocks.

The reliability of identifying gesture occurrences remains a significant methodological challenge within the field. While we currently evaluate inter-annotator agreement (IAA) based on the union of annotated gesture sets, this approach may be refined through the application of tools such as Staccato (Lücking et al., 2011).

The fully annotated data which we evaluated on constitutes only a subset of the total data collected. We annotated and reported statistics and results on 10 groups out of a full 33. The total length of all 33 videos is 19 hours, 46 minutes, and 27 seconds in length. As mentioned in Sec. 3.1, sessions contain a variant of the task that is less structured (larger footprint, gaps allowed in the structure, greater variety of block types). Our results on the 10 groups in the simpler of the two tasks already showcase the challenge of modeling and tracking common ground in the DPIP Lego task, and the less restrictive task variant is likely to be more challenging still.

Ethical Considerations

Collaborative construction tasks like Lego building exemplify creative construction as a developmental and epistemic practice extending beyond childhood. From a developmental ethics perspective, such tasks can foster autonomous imagination and collaborative sense-making, as participants negotiate meaning, share limited resources, and construct shared models of understanding through tangible forms. In our study, participants were provided with visual stimuli, pictures of target structures but no verbal or procedural instructions, compelling them to infer, coordinate, and reconstruct spatial relations through embodied reasoning and dialogue. Whereas this openness exposes the limits and frictions of creative collaboration, interpretive freedom may lead to conflict, misalignment, or the dominance of certain voices, while material constraints and the standardized logic of Lego pieces channel imagination into predefined geometries. What appears as freedom of construction thus may carry subtle forms of restriction and hierarchy, reminding us that even creative play may reproduce systems of order, control, and negotiation that mirror the social and epistemic structures from which it emerges.

Our data collection took place under an instantiated human subjects research protocol that was reviewed and approved by Colorado State University and Brandeis University Institutional Review Boards (IRB), as well as human subjects research protection offices at the agencies that funded the data collection. Participants consented to be recorded and were fairly compensated for their time (Sec. 3.2). Training and evaluating AI models over recorded data containing human likenesses risks exposing participants to invasion of privacy, however in our experiments, no video, raw audio, or human likeness data was sent to an AI model (e.g., GPT-5), only transcripts and annotations.

Acknowledgements

We are grateful to Tarun Varma Buddaraju, Jack Fitzgerald, Sai Kiran Ganesh Kumar, Sai Shruthi Garlapati, Carine Graff, Shamitha Gowra, Huma Jamil, Changsoo Jung, Ibrahim Khebour, Maniteja Vallala, Yangyang Chen, and Marc Verhagen for their valuable assistance in the data collec-

tion and annotation process, and to Bruce Draper and Nathaniel Blanchard for additional work on the task development. We would also like to thank the anonymous reviewers whose valuable feedback helped improved the quality of the final copy of this manuscript. This material is based in part upon work supported by Other Transaction award HR00112490377 from the U.S. Defense Advanced Research Projects Agency (DARPA) Friction for Accountability in Conversational Transactions (FACT) program, the U.S. National Science Foundation (NSF) under award DRL 2454151 (Institute for Student-AI Teaming), and by award W911NF-25-1-0096 from the U.S. Army Research Office (ARO). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

Bibliographical References

Anne H Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. 1991. The hcrc map task corpus. *Language and speech*, 34(4):351–366.

Nicholas Asher. 1998. Common ground, corrections and coordination. *Journal of Semantics*.

Cristian-Paul Bara, Sky CH-Wang, and Joyce Chai. 2021. MindCraft: Theory of mind modeling for situated dialogue in collaborative tasks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1112–1125, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Simon Baron-Cohen. 1997. *Mindblindness: An essay on autism and theory of mind*. MIT press.

Yonatan Bisk, Kevin Shih, Yejin Choi, and Daniel Marcu. 2018. Learning interpretable spatial operations in a rich 3d blocks world. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Thomas Bolander. 2014. Seeing is believing: Formalising false-belief tasks in dynamic epistemic logic. In *European conference on social intelligence (ECSI 2014)*, pages 87–107.

Richard Brutti, Lucia Donatelli, Kenneth Lai, and James Pustejovsky. 2022. Abstract Meaning Representation for gesture. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1576–1583, Marseille, France. European Language Resources Association.

Michael Bugert, Nils Reimers, and Iryna Gurevych. 2021. Generalizing cross-document event coreference resolution across multiple corpora. *Computational Linguistics*, 47(3):575–614.

Shu Cai and Kevin Knight. 2013. Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 748–752.

Justine Cassell. 2000. *Embodied conversational agents*. MIT press.

Joyce Y Chai, Lanbo She, Rui Fang, Spencer Ottarson, Cody Littley, Changsong Liu, and Kenneth Hanson. 2014. Collaborative effort towards common ground in situated human-robot dialogue. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 33–40.

Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. 2019. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12538–12547.

Ming Ming Chiu. 2008. Flowing toward correct contributions during group problem solving: A statistical discourse analysis. *The Journal of the Learning Sciences*, 17(3):415–463.

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.

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

Lee R Dice. 1945. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

Pierre Dillenbourg and David Traum. 2006. Sharing solutions: Persistence and grounding in multimodal collaborative problem solving. *The Journal of the Learning Sciences*, 15(1):121–151.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Alireza Fathi, Yin Li, and James M Rehg. 2012. Learning to recognize daily actions using gaze. In *European Conference on Computer Vision*, pages 314–327. Springer.

Hao Fei, Yuan Yao, Zhuosheng Zhang, Fuxiao Liu, Ao Zhang, and Tat-Seng Chua. 2024. From multimodal llm to human-level ai: Modality, instruction, reasoning, efficiency and beyond. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): Tutorial Summaries*, pages 1–8.

Mary Ellen Foster. 2007. Enhancing human-computer interaction with embodied conversational agents. In *International Conference on Universal Access in Human-Computer Interaction*, pages 828–837. Springer.

Alvin I Goldman. 2006. *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford University Press.

Alvin I Goldman. 2012. Theory of mind. In *The Oxford Handbook of Philosophy of Cognitive Science*. Oxford University Press.

H Paul Grice. 1989. *Studies in the Way of Words*. Harvard University Press.

Nikita Haduong, Irene Wang, Bo-Ru Lu, Prithviraj Ammanabrolu, and Noah A. Smith. 2024. CPS-TaskForge: Generating collaborative problem solving environments for diverse communication tasks. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, pages 86–112, Miami, Florida, USA. Association for Computational Linguistics.

Chien-Ming Huang, Sean Andrist, Allison Sauppé, and Bilge Mutlu. 2015. Using gaze patterns to predict task intent in collaboration. *Frontiers in psychology*, 6:1049.

Saki Imai, Mert İnan, Anthony Sicilia, and Malihe Alikhani. 2025. Measuring how (not just whether) vlms build common ground. *arXiv preprint arXiv:2509.03805*.

Georgi Karadzhov, Tom Stafford, and Andreas Vlachos. 2023. Delidata: A dataset for deliberation in multi-party problem solving. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–25.

Ibrahim Khebour, Richard Brutti, Indrani Dey, Rachel Dickler, Kelsey Sikes, Kenneth Lai, Mariah Bradford, Brittany Cates, Paige Hansen, Changsoo Jung, et al. 2024a. When text and speech are not enough: A multimodal dataset of collaboration in a situated task. *Journal of open humanities data*, 10.

Ibrahim Khebour, Changsoo Jung, Jack Fitzgerald, Huma Jamil, and Nikhil Krishnaswamy. 2025. Feature contributions to multimodal interpretation of common ground. In *International Conference on Human-Computer Interaction*, pages 265–281. Springer.

Ibrahim Khebour, Kenneth Lai, Mariah Bradford, Yifan Zhu, Richard Brutti, Christopher Tam, Jingxuan Tu, Benjamin Ibarra, Nathaniel Blanchard, Nikhil Krishnaswamy, and James Pustejovsky. 2024b. Common ground tracking in multimodal dialogue.

Hanna Komorowska. 2017. Speech and its silent partner: Gesture in communication and language learning. In *Challenges of Second and Foreign Language Education in a Globalized World: Studies in Honor of Krystyna Droździał-Szelest*, pages 73–87. Springer.

Stefan Kopp and Ipke Wachsmuth. 2010. *Gesture in embodied communication and human-computer interaction*, volume 5934. Springer.

Nikhil Krishnaswamy, Pradyumna Narayana, Rahul Bangar, Kyeongmin Rim, Dhruva Patil, David McNeely-White, Jaime Ruiz, Bruce Draper, Ross Beveridge, and James Pustejovsky. 2020. Diana's world: A situated multimodal interactive agent. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13618–13619.

Nikhil Krishnaswamy, Pradyumna Narayana, Isaac Wang, Kyeongmin Rim, Rahul Bangar, Dhruva Patil, Gururaj Mulay, Ross Beveridge, Jaime Ruiz, Bruce Draper, et al. 2017. Communicating and acting: Understanding gesture in simulation semantics. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS)—Short papers*.

Nikhil Krishnaswamy, William Pickard, Brittany Cates, Nathaniel Blanchard, and James Pustejovsky. 2022. The voxworld platform for multimodal embodied agents. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1529–1541.

Nikhil Krishnaswamy and James Pustejovsky. 2018. An evaluation framework for multimodal interaction. In *Proceedings of the eleventh international conference on language resources and evaluation (lrec 2018)*.

Nikhil Krishnaswamy and James Pustejovsky. 2020. A formal analysis of multimodal referring strategies under common ground. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5919–5927.

Kenneth Lai, Richard Brutti, Lucia Donatelli, and James Pustejovsky. 2024. Encoding gesture in multimodal dialogue: Creating a corpus of multimodal AMR. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5806–5818, Torino, Italia. ELRA and ICCL.

Andy Lücking, Sebastian Ptock, and Kirsten Bergmann. 2011. Assessing agreement on segmentations by means of staccato, the segmentation agreement calculator according to thomann. In *International gesture workshop*, pages 129–138. Springer.

Paul Marshall and Eva Hornecker. 2013. Theories of embodiment in hci. *The SAGE handbook of digital technology research*, 1:144–158.

Patrick W Miller. 2005. Body language in the classroom. *Techniques: Connecting education and careers*, 80(8):28–30.

Selene Mota and Rosalind W. Picard. 2003. Automated posture analysis for detecting learner's interest level. In *2003 Conference on Computer Vision and Pattern Recognition Workshop*, volume 5, pages 49–49.

Anjali Narayan-Chen, Prashant Jayannavar, and Julia Hockenmaier. 2019. Collaborative dialogue in minecraft. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5405–5415.

Abhijnan Nath, Carine Graff, Andrei Bachinin, and Nikhil Krishnaswamy. 2025. Frictional agent alignment framework: Slow down and don't break things. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11042–11089, Vienna, Austria. Association for Computational Linguistics.

Cheng Niu, Xingguang Wang, Xuxin Cheng, Juntong Song, and Tong Zhang. 2024. Enhancing dialogue state tracking models through llm-backed user-agents simulation. *arXiv preprint arXiv:2405.13037*.

OpenAI. 2025. Gpt-5 system card.

Eric Pacuit. 2017. *Neighborhood semantics for modal logic*. Springer.

Alexis Plaquet and Hervé Bredin. 2023. Powerset multi-class cross entropy loss for neural speaker diarization. In *Proc. INTERSPEECH 2023*.

James Pustejovsky and Nikhil Krishnaswamy. 2021. Embodied human computer interaction. *KI-Künstliche Intelligenz*, 35(3):307–327.

James Pustejovsky, Nikhil Krishnaswamy, Bruce Draper, Pradyumna Narayana, and Rahul Bangar. 2017. Creating common ground through multimodal simulations. In *Proceedings of the IWCS workshop on Foundations of Situated and Multimodal Communication*.

James Pustejovsky and Yifan Zhu. 2024. Lexical event models for multimodal dialogues. In *International Conference on Human-Computer Interaction*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large lms. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780.

Stefan Schaffer and Norbert Reithinger. 2019. Conversation is multimodal: thus conversational user interfaces should be as well. In *Proceedings of the 1st International Conference on Conversational User Interfaces*, pages 1–3.

Elizabeth Shriberg and Andreas Stolcke. 2004. Prosody modeling for automatic speech recognition and understanding. In *Mathematical Foundations of Speech and Language Processing*, pages 105–114. Springer.

Anthony Sicilia and Malihe Alikhani. 2025. Evaluating theory of (an uncertain) mind: Predicting the uncertain beliefs of others from conversational cues. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8007–8021.

Damien Sileo and Antoine Lernould. 2023. Mindgames: Targeting theory of mind in large language models with dynamic epistemic modal logic. *arXiv preprint arXiv:2305.03353*.

Thorvald Sørensen. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons. volume 5, page 1.

Alane Suhr, Claudia Yan, Jack Schluger, Stanley Yu, Hadi Khader, Marwa Mouallem, Iris Zhang, and Yoav Artzi. 2019. Executing instructions in situated collaborative interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2119–2130.

Jesse Thomason, Aishwarya Padmakumar, Jivko Sinapov, Nick Walker, Yuqian Jiang, Harel Yedidsion, Justin Hart, Peter Stone, and Raymond Mooney. 2020. Jointly improving parsing and perception for natural language commands through human-robot dialog. *Journal of Artificial Intelligence Research*, 67:327–374.

David Traum. 1994. A computational theory of grounding in natural language conversation. *PhD thesis, University of Rochester*.

Jingxuan Tu, Kyeongmin Rim, Bingyang Ye, Kenneth Lai, and James Pustejovsky. 2024. Dense paraphrasing for multimodal dialogue interpretation. *Frontiers in artificial intelligence*, 7:1479905.

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

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

Johan van Benthem, David Fernández-Duque, and Eric Pacuit. 2014. Evidence and plausibility in neighborhood structures. *Annals of Pure and Applied Logic*, 165(1):106–133.

Hannah VanderHoeven, Brady Bhalla, Ibrahim Khebour, Austin C Youngren, Videep Venkatesha, Mariah Bradford, Jack Fitzgerald, Carlos Mabrey, Jingxuan Tu, Yifan Zhu, et al. 2025. Trace: Realtime multimodal common ground tracking in situated collaborative dialogues. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 40–50.

Wolfgang Wahlster. 2006. Dialogue systems go multimodal: The smartkom experience. In *SmartKom: foundations of multimodal dialogue systems*, pages 3–27. Springer.

Isaac Wang, Mohtadi Ben Fraj, Pradyumna Narayana, Dhruva Patil, Gururaj Mulay, Rahul Bangar, J Ross Beveridge, Bruce A Draper, and Jaime Ruiz. 2017. Eggnog: A continuous, multimodal data set of naturally occurring gestures with ground truth labels. In *2017 12th Ieee international conference on automatic face & gesture recognition (fg 2017)*, pages 414–421. IEEE.

Peter C Wason. 1968. Reasoning about a rule. *Quarterly journal of experimental psychology*, 20(3):273–281.

Jason D Williams, Antoine Raux, and Matthew Henderson. 2016. The dialog state tracking challenge series: A review. *Dialogue & Discourse*, 7(3):4–33.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. Elan: A professional framework for multimodality research. In *5th international conference on language resources and evaluation (LREC 2006)*, pages 1556–1559.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Sina Zarrieß, Julian Hough, Casey Kennington, Ramesh Manuvinakurike, David DeVault, Raquel Fernández, and David Schlangen. 2016. Pentoref: A corpus of spoken references in task-oriented dialogues. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 125–131.

Hang Zhang, Zhuoling Li, and Jun Liu. 2025a. Scenellm: Implicit language reasoning in llm for dynamic scene graph generation. *Pattern Recognition*, page 111992.

Zheng Zhang, Nuoqian Xiao, Qi Chai, Deheng Ye, and Hao Wang. 2025b. Multimind: Enhancing werewolf agents with multimodal reasoning and theory of mind. *arXiv preprint arXiv:2504.18039*.

Xuhui Zhou, Zhe Su, Tiwalayo Eisape, Hyunwoo Kim, and Maarten Sap. 2024a. Is this the real

life? is this just fantasy? the misleading success of simulating social interactions with LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21692–21714, Miami, Florida, USA. Association for Computational Linguistics.

Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. 2024b. Archer: Training language model agents via hierarchical multi-turn rl. In *Forty-first International Conference on Machine Learning*.

Yifan Zhu, Changsoo Jung, Kenneth Lai, Videep Venkatesha, Mariah Bradford, Jack Fitzgerald, Huma Jamil, Carine Graff, Sai Kiran Ganesh Kumar, Bruce Draper, et al. 2025. Multimodal common ground annotation for partial information collaborative problem solving. In *Proceedings of the 21st Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-21)*, pages 85–91.