MINDMASTER: BENCHMARKING AND ANALYZING MENTAL STATE REASONING VIA AN IMMERSIVE SOCIAL INTERACTION GAME

Anonymous authors
Paper under double-blind review

ABSTRACT

Social explanation benchmarks typically evaluate what an observer believes happened, rather than how people explain decisions while participating in the interaction itself. In this work, we introduce MINDMASTER ROLEPLAY, a participant-grounded resource that links largely nonverbal social interaction to the judgments, intentions, actions, and reasons reported by the people shaping that interaction. Across 223 players, 388 trajectories, and 6,089 steps, these records preserve each explanation together with the local interaction history in which it arose. We use this resource to ask whether third-party observers recover participants' reported judgments and whether model explanations organize social evidence in the same way as Human accounts. Observer labels show limited correspondence with participants' action and intention reports. Explanation graphs differ in absence evidence and reasoning-type composition; differences in intermediate reasoning support persist under answer- and graph-size-matched analyses. A blinded human audit finds over 95% of sampled graph representations faithful to their source text. Together, these results show that "human-like explanation" is not a single target: it depends on whose interpretation is taken as the reference and whether similarity is measured in stated content or in the support structure connecting evidence, reasoning, and answers.

1. INTRODUCTION

Evaluating social reasoning in large language models (LLMs) involves more than checking a selected action or an inferred mental state. It also concerns how an explanation connects observations, goals, and prior actions to that judgment. A human reference for this comparison must specify whose interpretation it records. The reasons given by a person taking part in an interaction concern a situation that person experiences and helps shape; an observer interprets the encounter from outside it. Accounts of second-person social cognition emphasize this distinction between participation and observation (Schilbach et al., 2013). Preserving participants' own reports therefore makes their situated interpretations available for explanation-level evaluation.

Existing work establishes complementary ways to study social reasoning. Mental-state and interactive benchmarks assess perspective taking and social goal pursuit (Kim et al., 2023; Zhou et al., 2024). Process-oriented studies examine evidence grounding, mental–physical chains, and reasoning dependencies (Mathur et al., 2025; Niu et al., 2026; Kargupta et al., 2025; Xiong et al., 2025). These approaches provide behavioral outcomes and structured accounts of reasoning, but do not make reports from the people acting in an unfolding encounter their common reference. Connecting interactive behavior with participants' own reasons requires records that retain both the local history and the judgment being explained. Comparing those records with model explanations also requires separating the information an explanation invokes from the reasoning relations through which it supports an answer.

We address these two requirements with Mindmaster Roleplay, a two-player game centered on nonverbal social interaction. The original study's 223 players contributed 388 trajectories, reporting partner-intention estimates, their own intentions, and action choices with natural-language reasons throughout participation. Each report remains linked to the participant and local interaction history

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

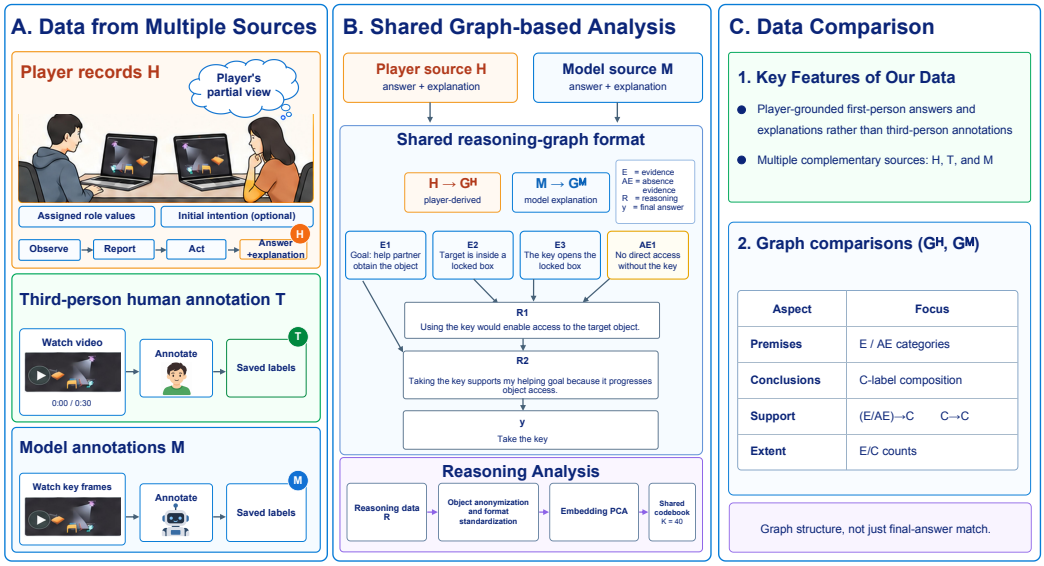


Figure 1: **Participant-grounded analysis of social explanations.** **(A)** H contains players' first-person reports, T third-party video judgments, and M saved model answers and explanations. **(B)** Participant-derived (G^H) and model (G^M) graphs represent evidence (E), absence evidence (AE), reasoning (R), and answers. R statements are compared in a frozen codebook; participant glosses are projected into it and models retain saved direct labels. **(C)** Comparisons examine premises, reasoning composition, support, and extent. C in the comparison panel denotes R. The graph is schematic; shared vocabulary does not imply identical elicitation or label assignment.

(Section 3). We use these records in two complementary analyses: correspondence with separately collected third-party video judgments, and comparison with model explanations for source-linked visual-history questions. Evidence–reasoning graphs represent premises, intermediate reasoning, and stated support while retaining the source of each account (Section 4; Figure 1).

The label analysis finds limited correspondence for recorded action classes and own intentions, including NA and decision-weighting sensitivity checks. The graph analysis combines source-level profiles with comparisons on common decision questions. Model graphs contain more explicit absence evidence and show distinct reasoning-type profiles in the expanded corpus. Astra retains fewer reasoning-to-reasoning links, including under unrounded repeat-mean R equality, with intervals clustered by source game. These are comparisons of generated representations rather than a ranking of reasoning quality (Section 5). A separate blinded human audit provides evidence for fidelity to the analysis text in its reviewed sample (Section 4.3). Together, these analyses motivate evaluating both the choice of human reference and the properties used to compare explanations against it.

Our contributions are a **participant-grounded resource** linking nonverbal interaction to the actor's reports; a **source-linked semantic and relational analysis** of expressed reasons; and an **empirical characterization** of label correspondence, premise selection, and reasoning support. Together, they make participants' interpretations accessible without substituting an external label for the original report or reducing an explanation to its final answer.

2. Related Work

2.1. Social judgments and interactive evaluation

Social-reasoning benchmarks evaluate mental-state inference and the decisions informed by it. FANToM tests perspective tracking in information-asymmetric conversations (Kim et al., 2023); SimpleToM separates mental-state inference from its application to behavior (Gu et al., 2026). SOTOPIA studies goal pursuit in open-ended human and model interactions (Zhou et al., 2024),

while SocialEval assesses outcomes and interpersonal abilities along scripted decision paths (Zhou et al., 2025). These approaches establish ways to evaluate what agents infer, choose, and achieve. An interaction score or an evaluator's justification, however, does not record the actor's own reason for a particular decision. Mindmaster adds this participant-authored layer: judgments and reasons are collected throughout interaction and linked to the behavior of the people providing them. The resulting records complement behavioral evaluation with a reference for the actor's reported interpretation.

2.2. Human explanations as a reference

Explanation-oriented evaluations examine more than answer agreement. Social Genome collects stepwise interpretations of social videos and evaluates semantic similarity, grounding, and reasoning structure (Mathur et al., 2025). Read the Room uses mental–physical causal chains to assess social understanding and reasoning consistency (Niu et al., 2026). Beyond social tasks, Chain-of-Thought as a Lens compares intermediate-step relations with human-preferred reference chains (Wang et al., 2026); graph-oriented analyses characterize reasoning operations and dependencies (Kargupta et al., 2025; Xiong et al., 2025). Such approaches provide representations of interpretation and support, but observing behavior, endorsing an explanation, and reporting one's own reasons are different sources of evidence. Mindmaster combines a participatory source with a semantic and relational comparison: the Human record is tied to the reporter's interaction history, while the graph keeps cited evidence distinct from intermediate reasoning and their support links. This permits analysis of participant-grounded accounts without treating their generated graph representations as direct measurements of internal cognition.

3. Human Study and Dataset

Mindmaster Roleplay is a two-player, turn-based game in which participants pursue individual or joint goals through pointing, gestures, movement, and object manipulation. Speech is restricted to predefined utterances. Coordination is therefore carried largely by observable actions and responses: a participant encounters the partner through the same unfolding scene in which the participant acts (Figure 2A). The dataset links this nonverbal interaction to how the players interpret their partner and justify their own decisions.

Reports are collected during sustained participation rather than as a separate interpretation of a completed episode. Players remain in their assigned roles while reporting a partner-intention estimate, an own-intention update, and an action choice, each with a natural-language reason; partner values are reported periodically. These texts are dataset annotations, not messages to the partner. Keeping the structured choice and its reason together distinguishes what the participant reported from the explanation of that report. Keeping both attached to the local observation and action history preserves the context in which it was given (Figure 2B–C).

The original study recruited **232 participants**, of whom **223 entered gameplay**. The source-file census contains **388 trajectory files and 6,089 step rows**: 5,485 rows have nonempty action and both intention fields; 1,464 contain partner-value entries. These are label-coverage counts, not counts of verified explanations or independent analysis participants (Appendix A.3). The resource's distinctive unit remains a participant-linked report in an ongoing encounter, retaining its local history and turn order. The analyses use source-linked decisions; they do not estimate longitudinal change or a causal effect of participation. Section 4 specifies the analysis samples.

4. Method

The dataset supports two comparisons. The first examines whether third-party video judgments recover participants' recorded labels. The second examines the evidence, intermediate reasoning, and support relations represented in Human and model explanation graphs. Label correspondence and graph profiles have separate sources and analysis units.

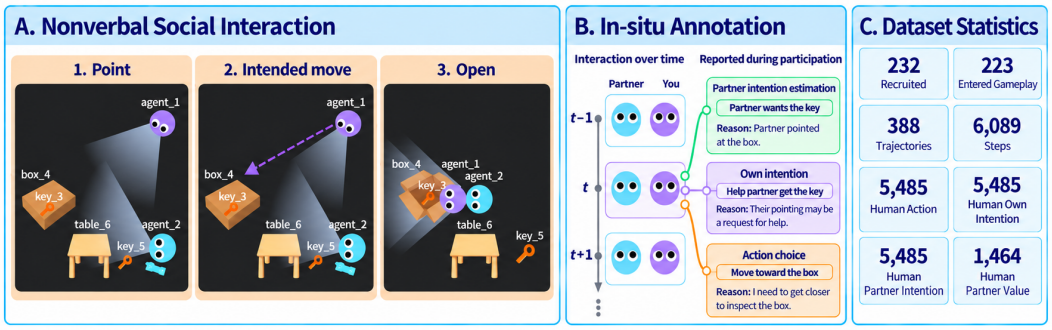


Figure 2: **Participant-grounded data collection. (A)** Pointing, movement, and object manipulation support largely nonverbal interaction. **(B)** Players report partner-intention estimates, own intentions, and action choices with reasons; partner values are elicited periodically. Reports are annotations, not partner-directed messages. Ordered turns preserve the participant and local history; displayed text is illustrative. **(C)** Of 232 recruits, 223 entered gameplay, yielding 388 trajectories and 6,089 steps. Each primary field has 5,485 entries; partner values have 1,464. These collection counts differ from analysis-subset sizes.

4.1. Comparison Design and Explanation Sources

A task $i = (g, u, t, c)$ identifies an interaction g, focal actor u, decision point t, and channel c. We denote the participant-derived Human source by H. A Human record contains the selected or inferred label y_i^H and its written reason. Third-party annotators T view recorded-belief video ending before the target action, with mental-state labels disabled. Human labels are taken from row t, with page-context state from row $t-1$. The primary sample contains 383 decision cells, including 114 repeatedly annotated cells; 52 additional cells contain only final-value reports (Appendices A.4 and F.2).

Action and own-intention questions concern the focal actor's next action and current intention. The partner template asks for the partner's current intention, whereas the paired Human field records the focal actor's estimate of that intention. The latter is therefore retained as a perspective-sensitive field comparison, not a confirmed same-target recovery test. Section 5.1 focuses on action class and own intention; Appendix E.1 specifies all channel definitions and Appendix F.2 retains every reported field and pooled count.

For graph analysis, G^H denotes graphs generated from Human explanation text retained in the source trace, while G^M denotes graphs generated from evaluated models' saved answers and explanations. All source explanations are processed with the same latest graph-construction pipeline, yielding G^H and G^M in the common representation used throughout the paper. The study uses 1,752 visual-history questions from 44 trajectories, with three response rounds each for Astra, DeepSeek, Gemini, and Claude. Models receive focal-actor storyboards and available state text, under target-answer, target-reason, hidden-world-state, and future-action exclusions, and respond to recorded contexts rather than playing the interaction online. Appendix C records response-generation conditions separately from the shared graph-construction stage.

4.2. Source-Linked Explanation Representation

For source $s \in \{H, M\}$, the directed graph represents evidence, absence evidence, intermediate reasoning, and their stated support (Dalvi et al., 2021):

$$G_i^s = \left(V_{E,i}^s \cup V_{AE,i}^s \cup V_{R,i}^s \cup \{v_{y,i}^s\},\ F_i^s\right). \tag{1}$$

E nodes identify cited evidence; AE nodes identify *absence evidence*, namely explicitly stated absence or information-gap premises; R nodes identify intermediate *reasoning* statements; and v_y records the answer. Support is directed from evidence or an intermediate reasoning statement to the statement it supports. R–R edges distinguish dependencies between reasoning steps from the mere presence of the same reasoning categories. Source, question, and response-round identifiers remain attached to the graph.

Source anchoring. The same latest graph-construction pipeline is applied to every analyzed explanation. It links composition statements to quotations, evidence, and dependencies in the source text; the evidence catalog locates expressed premises rather than supplying extra reasoning or repairing the answer. An R node is a saved reasoning/composition unit, not an assumed one-to-one sentence segment. AE requires an expressed absence: a missing catalog entry does not itself establish that something did not occur. Actor, time, negation, and uncertainty are part of the claim. Appendix D specifies the common representation and its validation.

Evidence and reasoning semantics. Ordinary E categories distinguish object information, agent information, self/other mental states, self/other values, and social interaction relative to the focal actor. AE families distinguish behavioral absence, social-support absence, and information-or-goal gaps. R labels use a frozen 40-type reasoning codebook (C types in the exports). Human-derived R statements are summarized as structural glosses, embedded with Jina embeddings v3, reduced by 64-dimensional PCA, and assigned in the frozen 40-cluster base vocabulary. Model comparisons use their saved direct labels in that vocabulary. The analysis retains direct base assignments, not refinement outputs; it does not refit clusters for each source. Glosses summarize complete R statements rather than newly selected evidence neighborhoods (Appendix E.3).

4.3. Measurements and Validation

Label correspondence (Section 5.1). We compare Human and third-party labels at two levels. Root matching evaluates action classes or intention predicates, while the stricter intention comparison uses normalized full strings. Actions remain class-level because separately stored arguments are not scored. Explicit `NA` values are retained, repeated annotations remain separate, and observer–observer agreement is computed independently from repeated cells (Appendix E.1).

Evidence profiles (Section 5.2). We measure both the ordinary evidence cited in an explanation and whether it explicitly represents missing behavior, support, information, or goals. Let $\mathcal{G}_s^E$ denote the eligible graphs for source s, n_g^{AE} the number of absence-evidence nodes in graph g, and n_{gf}^{AE} the number assigned to family f. We report graph-level AE prevalence and the conditional composition of AE nodes:

$$\pi_s^{AE} = \frac{\sum_{g \in \mathcal{G}_s^E} \mathbf{1}[n_g^{AE} > 0]}{|\mathcal{G}_s^E|}, \qquad q_{sf}^{AE} = \frac{\sum_{g \in \mathcal{G}_s^E} n_{gf}^{AE}}{\sum_{g \in \mathcal{G}_s^E} n_g^{AE}}. \tag{2}$$

These measures characterize expressed absence evidence rather than whether an absence claim is correct in the environment.

Reasoning content and support (Sections 5.3–5.4). Mean R/E counts measure extent; R–R links measure support. For graph-level label counts n_{gk} and totals n_g, pooled-node shares are $p_{sk} = \sum_g n_{gk} / \sum_g n_g$; equal-graph shares are $\tilde{p}_{sk} = |\mathcal{G}_s^R|^{-1} \sum_g n_{gk}/n_g$, over graphs with valid labels. We compare both distributions using base-2 Jensen–Shannon (JS) divergence, separately from support topology (Appendices E.3–E.4).

Paired comparisons and uncertainty. For shared-question graph comparisons, available model repeats are averaged within each question before pairing with the Human graph for the same game, focal actor, decision, and channel. We report model-minus-H differences with 95% percentile intervals from 5,000 source-game bootstrap resamples, keeping records from the same game together. Human-label intervals use the same source-game resampling principle. Exact-R sensitivity analyses further restrict H–model comparisons to matched reasoning-node counts (Appendix E.5; results in Appendices F.5–F.6).

Graph-fidelity validation. Two independent human reviewers evaluated 120 blinded analysis-text–graph pairs produced by the graph-construction pipeline. They judged 115/120 (95.83%) and 117/120 (97.50%) pairs faithful, with exact agreement on 112/120 pairs (93.33%; Gwet's AC1 = 0.9287). This audit evaluates fidelity of the graph representation to its source analysis text (Appendix D.3).

5. Results

We examine observer–participant correspondence, expressed evidence, reasoning-type distributions, and reasoning linkage. Human (H) denotes the participant-derived source defined in Section 4.1.

The tables retain each measure's eligible population rather than combining label pairs, graph records, and reasoning nodes into one sample count.

5.1. Observer–participant correspondence

Third-party judgments only partially recover participants' reports of the same decision. Action-class correspondence is 191/579 (32.99%; 95% game-clustered CI [28.13, 39.01]). Own-intention correspondence is 128/579 at the predicate level (22.11%; [17.18, 27.85]) and 107/579 for normalized full strings (18.48%; [13.86, 23.84]). These comparisons preserve the distinction between an actor's situated report and an external interpretation.

Table 1: Action-class and own-intention correspondence between human sources.

Field	Pairs	Root same	Root (%)	Strict (%)
Participant–observer: non-empty label pairs				
Action class	579	191	32.99	32.99
Own intention	579	128	22.11	18.48
Observer–observer: 114 repeated primary cells				
Action class	278	90	32.37	32.37
Own intention	278	71	25.54	16.19

Disagreement also occurs between observers: on the 114 repeatedly annotated cells, action-class agreement is 32.37%, own-intention predicate agreement 25.54%, and strict own-intention agreement 16.19% (Table 1). These estimates use different records and weights from participant–observer correspondence, but show that the same pre-action video does not induce a single external interpretation.

The pattern persists under NA and weighting sensitivities. Excluding pairs with `NA` on either side leaves action correspondence unchanged and raises own-intention root correspondence to 25.00% ([19.61, 30.82]; 492 pairs). Equal decision weighting gives 36.03% for action and 24.37% for own intention; adding NA exclusion gives 27.49% for own intention (Table S6).

Label differences are structured: observers assign `observe` 130 times versus 59 participant reports, and `help` 20 times versus 55 (579 own-intention pairs; Appendix F.2). These results support keeping the two references distinct; they do not show that participant reports are objectively correct or that participation causally improves inference. Full matching rules and the perspective-sensitive partner-field results appear in Appendices E.1 and F.2.

5.2. Absence evidence distinguishes the represented premise profiles

The evidence comparison concerns cited information and what an explanation explicitly treats as absent or missing. At least one AE node appears in 74 of 1,325 Human (H) graphs (5.58%), versus 34.42–64.69% of model graphs with E/AE annotations (Table 2). These global prevalences use the full sidecar populations; the paired analysis below uses R-bearing shared questions.

Table 2: Absence evidence: graph-level presence and conditional node composition.

Source	Graphs with AE / eligible	With AE (%)	Behavioral absence (within AE, %)	Social-support absence (within AE, %)	Information/ goal gap (within AE, %)
Human (H)	74 / 1,325	5.58	11.54	21.79	66.67
Astra	1,791 / 5,204	34.42	26.04	33.49	40.48
DeepSeek	2,160 / 5,144	41.99	35.93	25.46	38.61
Gemini	2,541 / 5,212	48.75	41.82	30.87	27.30
Claude	3,279 / 5,069	64.69	31.68	25.09	43.23

Family shares condition on AE nodes: 78 / 2,777 / 3,888 / 5,966 / 7,509 in table order. Global prevalence is unadjusted (Appendix F.3).

On common decision questions, model AE presence is higher by 23.21 percentage points for Astra, 30.14 for DeepSeek, 38.41 for Gemini, and 55.08 for Claude. All four pointwise 95% game-clustered intervals lie above zero (Table S13). These comparisons use 1,167 Astra–H or 1,168 other model–H questions across 41 games, averaging available model repeats before pairing. They cover action, own-intention, and partner-estimation graph channels, not value questions. Recovered model analysis texts are substantially longer than Human participant reports, so we treat AE as a source-specific representation difference rather than a length-independent reasoning property (Appendix F.3).

The AE-family composition asks a separate question: what is represented as absent? Information-or-goal gaps constitute 66.67% of the Human reference's 78 AE nodes, versus 27.30–43.23% in models. The low Human AE-node count limits this conditional profile. Neither AE presence nor AE family establishes that an absence claim is warranted by the scene.

Ordinary evidence also shows source-specific emphasis (Table S7). Every model has a smaller self-mental-state share than H. Agent information is more prominent in Astra, Gemini, and Claude; DeepSeek instead combines a smaller agent-information share with a larger self-value share. These are coded-evidence compositions, not the agents' personal values or the reasoning conclusions supported by each evidence node.

5.3. Reasoning-type distributions differ across sources

We examine the final 40-type codebook in a broader materialized corpus containing 73,604 valid labeled R nodes: 12,959 from $\text{Human}_{\text{full}}$, 10,515 from Astra, 22,608 from DeepSeek, and 27,522 from Gemini. These span 6,619 / 5,202 / 5,113 / 5,212 graphs, respectively. $\text{Human}_{\text{full}}$ is the broader participant-derived reference for this analysis, distinct from the 1,298-graph H subset used for structure comparisons; Claude is absent from this export. Empty or unknown labels are excluded.

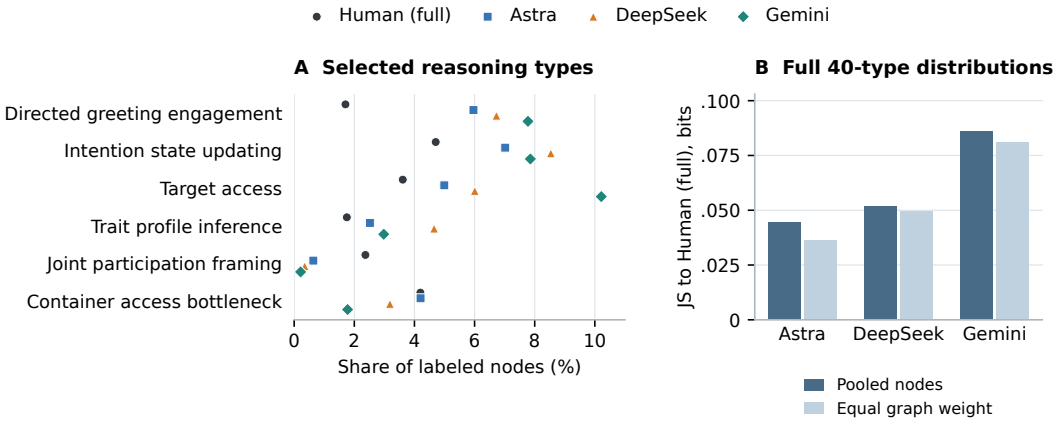


Figure 3: **Reasoning-type composition in the expanded corpus. (A)** Pooled-node shares for six illustrative codebook types, selected to show common and source-specific contrasts; Appendix F.8 reports all 40. **(B)** JS divergence from $\text{Human}_{\text{full}}$ over all 40 types. Equal graph weighting normalizes within each graph before averaging, reducing the influence of graphs with more labeled nodes. Both panels describe source-level composition, not matched-question effects or reasoning quality.

All three models allocate more labeled nodes to *Directed Greeting Engagement* (5.96–7.78%, versus 1.71% in $\text{Human}_{\text{full}}$) and *Intention State Updating* (7.02–8.54%, versus 4.71%; Figure 3A). Conversely, *Joint Participation Framing* accounts for only 0.21–0.64% in the models, versus 2.37% in $\text{Human}_{\text{full}}$. Additional contrasts are source-specific: Gemini assigns 10.21% to *Target Access*, versus 3.61% in $\text{Human}_{\text{full}}$, while DeepSeek assigns 4.65% to *Trait Profile Inference*, versus 1.75%.

Pooled-node JS divergences from $\text{Human}_{\text{full}}$ are 0.0444, 0.0520, and 0.0859 bits for Astra, DeepSeek, and Gemini. Equal graph weighting yields 0.0361, 0.0496, and 0.0809 bits, preserving their descriptive ordering (Figure 3B). The greeting differences also persist with equal graph weights: +4.63, +5.22, and +6.22 percentage points. Thus the composition contrasts are not solely due to larger graphs receiving more node weight.

These profiles remain conditional on the corpus and label-assignment pipeline; equal weighting does not match questions, task mix, or information conditions. Appendix F.8 retains the separate 1,000-resample, independent-graph bootstrap as an exploratory sensitivity; it does not preserve game-level dependence. The original H-subset JS values remain separate (Table 3) and are not overwritten by this broader-reference analysis.

5.4. Reasoning linkage differs on shared decision questions

The global H–Astra comparison combines similar mean R counts (1.98 and 2.02) with different mean R–R counts (0.97 and 0.43; Table 3). Pairing 1,167 common decision questions across 41 games retains this difference: Astra averages 0.481 R–R links and H 0.962, giving a model-minus-H difference of -0.481 links per question (95% CI $[-0.555, -0.402]$). The channel-specific contrasts are provided in Table S14; pooled means are not treated as exact size matching.

Table 3: Explanation extent, linkage, and composition against the original H subset.

Source	Mean R	Mean E	Mean R–R links	R-label JS to H (bits)
Human (H)	1.98	4.20	0.97	—
Astra	2.02	4.60	0.43	0.1360
DeepSeek	4.42	9.41	2.45	0.1731
Gemini	5.28	10.22	2.84	0.1444
Claude	6.22	14.87	4.09	0.1204

R/E and R–R means use 1,298 / 5,202 / 5,113 / 5,212 / 5,069 R-bearing records in row order. JS retains its global labeled-node populations (Table S9). The juxtaposition is descriptive, not a within-question association or a size-adjusted effect; node and support-count summaries remain in Appendix F.4.

Exact-R sensitivities preserve the H–Astra contrast: unrounded repeat-mean R equality retains 326 questions across 40 games and gives -0.383 R–R links (95% CI $[-0.436, -0.339]$), while graph-by-graph equality gives -0.395 $[-0.434, -0.355]$ (Table S15). In an additional offline sensitivity requiring both normalized-answer agreement and exact R count within the joined model-round/channel strata, the Astra–H difference remains negative (-0.373 links), but mean coverage is only 8.0%; we therefore treat it as a descriptive robustness check rather than a primary estimate. Full overlap definitions and coverage are in Appendix F.6.

On the original H subset, Claude has the smallest direct-label JS divergence (0.1204 bits) but the largest mean R/E counts and most R–R links. Separate structure-label recomputation preserves that ordering (Table S16). Category composition does not specify support topology, and these source-level summaries do not establish reasoning quality.

Link counts do not by themselves imply a fixed chain. In the saved topology audit, 98.6% of H graphs satisfy $|F_{RR}| = |V_R| - 1$, but only 38.8% meet the stricter non-branching single-chain rule, compared with 38.5% for Astra. These descriptive checks distinguish edge-count identities from actual topology (Appendix F.6).

6. Discussion

6.1. The actor's report as an evaluation reference

A Human participant report records the judgment the actor gave during the interaction; its value does not require that judgment to be universally correct. Observer consensus and recovery of a particular actor's report are different evaluation goals. The action and own-intention results remain limited under game-clustered uncertainty and the stated NA/decision-weighting sensitivities. This supports retaining both sources rather than replacing the actor's report without changing the evaluation target.

Mindmaster also preserves the connection between a report and the participant's local history. This is the resource contribution of sustained participation: reasons remain associated with a particular actor, decision, and evolving encounter. The current analyses compare decision-level records and

their representations; they do not demonstrate a temporal law of social reasoning or an advantage caused by immersion. The ordered records preserve the basis for investigating such questions without pretending that those additional analyses have already been performed.

6.2. What content-only comparison leaves unspecified

The graph analysis makes two properties of explanations inspectable beyond a final answer or a global content distance. First, an explanation may use explicitly stated absence evidence as a premise. The AE profiles identify this behavior in the generated representations, but its interpretation depends on what is claimed to be absent and whether the available information supports that claim. More AE is neither intrinsically better nor worse: justified statements of missing information and unwarranted assertions of non-occurrence would require different judgments. Checking that distinction is an implication of the result, not a completed correctness evaluation.

Second, the expanded-corpus analysis shows different reasoning-type compositions even under equal graph weighting. Such compositions still do not specify which intermediate reasoning step supports another. The common-question comparison shows fewer represented R–R links for Astra, and the contrast remains negative in Exact-R subsets and in the smaller answer-matched plus Exact-R sensitivity. These checks narrow simple answer- and size-based explanations without constituting a causal control. The representation makes reasoning-support links inspectable alongside content; it does not demonstrate superiority over all text-based metrics.

6.3. Scope and limitations

The human results remain conditional on the implemented video task, explicit-NA rule, and class-only action comparison. Actual historical visibility, account-to-person identity, and cross-game participant/session membership are not fully verified. The source-game bootstrap retains within-game dependence but does not separately estimate dependence from contributors shared across games. The partner-field perspective caveat remains as defined in Section 4.1. The broader $\text{Human}_{\text{full}}$ distribution is not question-matched to the models, and its graph bootstrap does not account for game-level clustering.

Common-question pairing and finite intervals strengthen the source comparisons. All reported graphs use the same latest graph-construction pipeline, and blinded human review supports fidelity of its outputs to the analyzed text in the audited sample. This validation concerns representation fidelity, not whether an explanation is factually warranted by the environment or whether it reveals an agent's internal computation. Exact-R comparisons remain restricted to their reported overlap subsets (Jacovi & Goldberg, 2020; Turpin et al., 2023).

7. Conclusion

Mindmaster Roleplay preserves judgments and reasons from people participating in nonverbal social interaction. Third-party labels show limited action and own-intention correspondence under NA and weighting sensitivities. Shared-question graph comparisons retain more explicit absence evidence across models and fewer R–R reasoning links for Astra, including in Exact-R overlap subsets, with game-clustered intervals. The expanded corpus also shows source-specific reasoning-type distributions under both node pooling and equal graph weighting. Blinded human review supports fidelity to the analysis text in the audited sample. The resource supports evaluating whose judgment a reference records, which evidence an explanation expresses, and how intermediate reasoning supports its final answer.

ETHICS STATEMENT

The original human-study documentation reports approval by the Institutional Review Board of the School of Psychological and Cognitive Sciences, Peking University, under protocol #2023-03-16, together with informed consent, compensation, withdrawal rights, debriefing opportunities, and anonymized, secure storage. Assigned role values and in-game decisions do not describe participants' personal traits. Appendix A.5 records the reported safeguards and payment schedule. The dataset and derived annotations used in this work will be released after de-identification; personally identifying information will not be released.

REPRODUCIBILITY STATEMENT

The study artifacts provide label pairings, graph metrics, shared-question and repeat-level counts, saved labels, validation summaries, and analysis scripts. Human and paired-AE intervals are reproducible from delivered rows, and the repeat-level table permits replay of the paired R–R and Exact-R analyses without new model calls or reasoning-codebook fitting. All reported graphs use the same latest graph-construction pipeline; the blinded human audit in Appendix D.3 provides semantic-fidelity evidence for its analysis-text-to-graph outputs. Appendices E.5 and F.5–F.7 define the statistical populations and overlap restrictions. We will release the de-identified Mindmaster dataset, derived annotations, and the data-processing, graph-construction, and statistical-analysis code used for the reported results. Personally identifying information will not be released. Raw participant text and identifying source paths are not included in the anonymous manuscript source ZIP.

AI USE STATEMENT

Generative AI was used in the historical analyses and current visual-response and graph-conversion runs to generate explanations and graph candidates, and to assist structural glossing, semantic labeling, and validation. It also assisted code writing, literature synthesis, and manuscript revision. The authors are responsible for reviewing the final text, claims, and artifacts.

REFERENCES

Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. Explaining answers with entailment trees. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7358–7370, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.585. URL https://aclanthology.org/2021.emnlp-main.585/.

Yuling Gu, Oyvind Tafjord, Hyunwoo Kim, Jared Moore, Ronan Le Bras, Peter Clark, and Yejin Choi. SimpleToM: Exposing the gap between explicit ToM inference and implicit ToM application in LLMs. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/524ff06d5375d76e93d8490e654a81bf-Abstract-Conference.html.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.386. URL https://aclanthology.org/2020.acl-main.386/.

Priyanka Kargupta, Shuyue Stella Li, Haocheng Wang, Jinu Lee, Shan Chen, Orevaoghene Ahia, Dean Light, Thomas L. Griffiths, Max Kleiman-Weiner, Jiawei Han, Asli Celikyilmaz, and Yulia Tsvetkov. Cognitive foundations for reasoning and their manifestation in LLMs, 2025. URL https://arxiv.org/abs/2511.16660.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Le Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. FANToM: A benchmark for stress-testing machine theory of mind in interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp.

14397–14413, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.890. URL https://aclanthology.org/2023.emnlp-main.890/.

Leena Mathur, Marian Qian, Paul Pu Liang, and Louis-Philippe Morency. Social genome: Grounded social reasoning abilities of multimodal models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 24868–24891, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1264. URL https://aclanthology.org/2025.emnlp-main.1264/.

Lixing Niu, Jiapeng Li, Xingping Yu, Xinyi Dong, Shu Wang, Ruining Feng, Bo Wu, Ping Wei, Yisen Wang, and Lifeng Fan. Read the room: Video social reasoning with mental-physical causal chains. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/0ba10e63a87590a9364f33b9040e4301-Abstract-Conference.html.

Leonhard Schilbach, Bert Timmermans, Vasudevi Reddy, Alan Costall, Gary Bente, Tobias Schlicht, and Kai Vogeley. Toward a second-person neuroscience. *Behavioral and Brain Sciences*, 36(4): 393–414, 2013. doi: 10.1017/S0140525X12000660.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, and Han Xiao. jina-embeddings-v3: Multilingual embeddings with task LoRA, 2024. URL https://arxiv.org/abs/2409.10173.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, pp. 74952–74965. Curran Associates, Inc., 2023. doi: 10.52202/075280-3275. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/ed3fea9033a80fea1376299fa7863f4a-Abstract.html.

Boxuan Wang, Zhuoyun Li, Xinmiao Huang, Xiaowei Huang, and Yi Dong. Chain-of-thought as a lens: Evaluating structured reasoning alignment between human preferences and large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 39514–39530, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1834. URL https://aclanthology.org/2026.acl-long.1834/.

Zhen Xiong, Yujun Cai, Zhecheng Li, and Yiwei Wang. Mapping the minds of LLMs: A graph-based analysis of reasoning LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 17751–17763, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.896. URL https://aclanthology.org/2025.emnlp-main.896/.

Jinfeng Zhou, Yuxuan Chen, Yihan Shi, Xuanming Zhang, Leqi Lei, Yi Feng, Zexuan Xiong, Miao Yan, Xunzhi Wang, Yaru Cao, Jianing Yin, Shuai Wang, Quanyu Dai, Zhenhua Dong, Hongning Wang, and Minlie Huang. SocialEval: Evaluating social intelligence of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30958–31012, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1496. URL https://aclanthology.org/2025.acl-long.1496/.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/b3075b88e583a0e98d8b24338a613060-Abstract-Conference.html.

SUPPLEMENTARY MATERIAL

This supplement contains the collection and information conditions needed to interpret the participant reference (A–C), graph-construction and measurement definitions (D–E), and the complete supplementary results supporting Sections 5.1–5.4 (F). The source-game intervals, NA/weighting sensitivities, and channel-specific paired contrasts are retained with their eligibility rules. Appendix D.3 reports the blinded human analysis-text–graph fidelity audit; source and measurement limitations remain explicit.

Notation. The archived analyses below retain C for the intermediate reasoning units denoted R in the main text, and SE for absence evidence denoted AE. C–C and Exact-C therefore correspond to R–R and Exact-R; they do not define additional node types. Appendix F.8 reports the expanded $\text{Human}_{\text{full}}$ C-type distribution separately from the original H-subset statistics.

APPENDIX A. HUMAN STUDY AND PARTICIPANT RECORDS

A.1. COLLECTION AND QUALIFICATION

The original study documentation reports controlled, in-person sessions with 232 recruits, of whom 223 entered gameplay after qualification. The reported collection comprises 388 trajectories and 6,089 steps. These quantities describe the original collection, not the number of independent contributors to each analysis subset. Participant-to-session membership and the link to the current 217-interaction pool are not fully established.

The documented onboarding included an approximately 25-minute tutorial and 15-minute quiz, examples, warm-up play, and help resources. Training covered agent identity, observation versus memory, object affordances, available actions, assigned role values, and justification reporting. Qualification preceded gameplay; failed attempts required restudy. The precise deployed quiz, feedback, pass threshold, and example explanations are unavailable. Qualification is therefore not evidence that explanations are accurate or that training did not influence their content.

A.2. INTERACTION AND REPORT ELICITATION

Participants were paired in a virtual room and assigned role-value profiles; an initial intention could be absent. The Active, Social, and Helpful dimensions describe assigned roles, not participants' personal traits. Intentions include object-directed goals and social goals such as informing, helping, requesting help, or hindering, and may be updated during interaction. Scenarios provide compatible or conflicting goals and variations in object access, location, and visibility. A final scenario roster and generation distribution for the analyzed corpus are not established by the supplied descriptions.

The interface displays the focal actor's system-maintained observation/memory representation. Currently perceived objects are vivid and remembered objects outside the view are faded; absence from the display does not establish nonexistence. The backend logs world state separately. Actions include navigation, orientation, object manipulation, pointing/following, greetings, nodding or shaking the head, and predefined utterances such as "hello" and "thank you." Actual out-of-interface speech, screen sharing, and access to the partner's report fields during the sessions remain unconfirmed.

Participants report a partner-intention estimate, own intention, and action choice with separate reasons. Partner values are reported periodically; process-value entries and dedicated terminal surveys remain distinct. The recorded sequence associates reports with decision stages, but does not establish that text was committed before outcomes became visible. Editing access, original language, and preprocessing/translation between the original report and the graph's logged input also require confirmation. We describe in-situ reports, not verified pre-outcome commitments. The Figure 2 report text is illustrative rather than verbatim, and its selected turns need not be adjacent.

A.3. COLLECTION-TO-ANALYSIS SCOPE

The source-row inventory enumerates 6,089 unique file/step pairs from 388 files. Its label coverage is shown below; nonempty includes explicit NA and does not imply a usable accompanying reason. Full-source coverage is distinct from the third-party pairing sample and from the graph-analysis

subset. The preprocessing description does not provide an attrition mapping sufficient to connect all collection and analysis populations in Table S1. The 44 visual trajectories were selected without replacement from the 217-interaction pool. The 1,752-question inventory includes action, own intention, partner-intention estimation and process/terminal values; its per-channel question totals remain unspecified.

Source field	Nonempty, including NA	Missing
Action	5,485	604
Own intention	5,485	604
Partner-intention estimate	5,485	604
Partner values	1,464	4,625

The same 5,485 rows contain the three primary fields. An auxiliary `completed_trajectory_rows` view contains 372 files and 5,896 rows, using the last nonempty status of SUCCESS or REACH_MAX_ITER. It has 5,292 nonempty entries per primary field and 1,440 value entries. This export-defined view does not replace the full census or reselect the main analyses. Label-status NA counts and roots parsed as NA are not assumed interchangeable; only their combined nonempty coverage is used here. Neither view supplies participant/session linkage, free-text quality, or complete reason coverage.

Table S1: Collection and analysis scopes; the rows do not constitute a verified attrition chain.

Source scope	**Reported quantity**	**Link still required**
Original recruitment and gameplay	232 recruited; 223 entered gameplay	Participant-to-session membership and current-corpus inclusion
Original collected corpus	388 trajectory files; 6,089 source rows	Source game/trajectory IDs and mapping to the current pool
Current interaction pool	217 interaction samples	Mapping to the original collection and preprocessing version
Current visual selection	44 trajectories	Frozen selection IDs and distinct contributors
Current task inventory	1,752 source-aligned questions	Original-reason availability, generated responses, and accepted pairs

A.4. Third-party judgments and source identity

Participant records P contain a selected or inferred label and its reason. Third-party labels T are independent annotations of pre-action video linked to the same source decision. Model explanations M are saved model responses; participant-derived graphs G^H and model-explanation graphs G^M are distinct derived records. No third-party graph comparison is performed.

The observer implementation takes target labels from source row t, restores page context from $t-1$, and sets the video cutoff at t with recorded-belief rendering enabled and mental-state labels disabled. Templates ask what the focal actor will do next, the focal actor's current intention, and the partner's current intention. The participant partner field instead records the focal actor's estimate of the partner's intention. That channel remains field correspondence, not a confirmed same-perspective recovery test. Matching the row does not resolve this target distinction.

The interface selects an action class with separate arguments and uses intention selectors. Partner-value sliders range from 0 to 1 at 0.05 increments. The exported configuration does not settle actual tutorials, replay access, private information visibility, or whether accounts identify independent people. Source fields that exist in a stored record are not assumed to have been displayed. Complete label counts and weighting sensitivities appear in Appendix F.2.

A.5. Ethics, consent, and access

The collection documentation reports approval by the Institutional Review Board of the School of Psychological and Cognitive Sciences, Peking University, protocol #2023-03-16, informed consent, compensation, withdrawal rights, debriefing, anonymization, and secure storage. The dataset and

derived annotations used in this work will be released after de-identification; personally identifying information will not be released.

The documented payment schedule specifies 25 for tutorial/qualification, 95 for gameplay and a 1–30 quality-dependent bonus; its prospective wording does not verify actual payments. The operational bonus rubric and whether explanation content affected rewards are not established. Data release will use de-identified records, and the accompanying data-processing, graph-construction, and analysis code will be open-sourced with the research release.

APPENDIX B. VISUAL HISTORIES AND INFORMATION BOUNDARIES

B.1. RECORDED-BELIEF HISTORIES

The documented visual materializer replays the focal actor's recorded-belief history, not omniscient world state. Its canonical asset is a 720×360, 2:1 storyboard per completed action: the left panel is the selected in-action frame, the right panel the recorded terminal frame. Step 1 lacks a pre-action snapshot, so its left panel places a non-text action glyph over the recorded terminal view. Sidecars retain the image, dimensions and render identity; the exact salience-selection implementation is not fully documented. Source/model aliases alone do not establish that every output group uses identical rendered assets.

Question construction retains source identity, actor, target row, role information and decision-time context. Reference answers remain offline for scoring; model graph terminals must represent the model's saved answer rather than substitute a source answer. Histories are chronological, and the query is separated from the history items. History-dependent claims are instructed to cite `[Step N]` or `[Steps N, M]`; `[Background]` identifies task definitions. Citation syntax is not evidence that the claim is grounded.

B.2. MODEL-FACING INFORMATION

Table S2 records the declared input boundary. State text may contain mental-state information already available to the focal actor at the corresponding point; it is not an additional visual observation. Exact task-specific field placement and time cutoffs require the frozen requests. Current target reasons are excluded from the evaluated-model input. Some model state fields are absent from the third-party video condition, so their information conditions are not assumed equivalent.

Table S2: Declared information boundary of response generation; this is not a completed leakage audit.

Information	**Model-facing use**
Scene header, shared vocabulary, task definitions	Included as context
Focal actor's recorded-belief visual history up to the query	Included as chronological storyboards
Mental state already available at the corresponding decision point	Included as state text under the task's time boundary
Current target answer and answer labels	Excluded; retained only offline where required
Hidden world state and future actions	Excluded
Graph annotations and extractor evidence catalog	Not included in the task-answering prompt

APPENDIX C. RESPONSE GENERATION AND RUN IDENTITY

C.1. REQUESTS AND RECORDED CONDITIONS

Each trajectory's prefix contains scene information and shared vocabulary, followed by chronological images/text and a question. The response contract requests a concise reasoning summary and an action, intention, or value answer. Models answer recorded contexts rather than playing online. The backend manifest supplies the following *recorded response identifiers*, not independent verification of provider snapshots:

Reported group	Saved response-model identifier
Astra	`gpt-6-astra`
DeepSeek	`deepseek-flash`
Gemini	`gemini-3.8-flash-high`
Claude	`claude-opus-5-5`

The earlier response records describe medium reasoning for Astra, high reasoning/temperature 1 for Gemini, and medium effort for Claude. The new manifest's `reasoning_effort`, temperature and token-limit columns instead come from *graph-extraction* requests: the sampled successful trace for each M group names `gpt-5.6-sol`, medium effort, temperature 0, and a configured 65,536-token maximum. They must not be attributed to the models producing the original explanations. Response-generation budgets, seeds, prompt-template hashes, and final DeepSeek decoding settings remain unavailable.

Each sampled model graph trace records conversion version `multimodal_reason_graph_prompt_v13`, schema `multimodal_compact_reason_graph_v9`, parser `node_only_references_v2_agent_alias`, and 720×360 images. One successful trace and one example conversion hash do not establish uniform settings or full membership of an archive. The DeepSeek manifest still points to 4,223 successful graphs, not the 5,151-payload analysis inventory (Appendix D.2). Failure/retry event counts are not unique question failures. The participant-side graph signature remains separately specified in C.2.

C.2. Responses, repeats, and source-to-graph lineage

The four evaluated groups retain three requested response rounds on the same 1,752-question inventory. Saved response IDs and run metadata distinguish repeats from resume attempts. An unresolved transport attempt is not a semantic task failure. Payload, evidence-bearing graph, C-bearing record and labeled-node coverage are reported separately in Appendix F.1; three requested rounds do not guarantee three available outputs for every question.

For the participant side, 1,325 current graph rows link explanation text retained in a prompt trace directly to a single-pass generated graph. The registered run signature identifies `gpt-5.6-terra`, `xhigh`, temperature 0.2 and `agent_view`. These are source-side graph-generation identifiers, not settings for the four evaluated model answers. The effective serving token limit is not established. The current route contains no independent reconstructed prose stage. Provenance identifies the transformation but does not by itself certify semantic preservation. The separate audit in D.3 assesses analysis-text-to-graph fidelity, not original-language-to-trace preservation.

C.3. Reproduction and data availability

Research inputs contain the source-label census, paired human labels, common-question joins, graph metrics, codebook counts and automatic checks. The backend increment adds 13,797 question–repeat rows with H/model C and C–C counts and actual available-round identifiers. Its eight exact-C summaries can be reproduced from these rows. Joining their C–C values to the prior common-question order also reproduces the original pooled and channel-specific H–Astra intervals. Human and paired-SE intervals retain their previous row-based checks. These replays verify saved counts and statistics, not graph generation or upstream response/question identity.

Partial analysis-text length counters, registered audit IDs and selected case files are also available; they are not completed length-adjusted estimates, per-item human review scores, or publication clearance for examples. Full regeneration still depends on original trajectory and graph records, complete requests, the code environment, and embedding/fitted-codebook artifacts. The public research release will include the de-identified dataset, derived annotations, and the data-processing, graph-construction, and statistical-analysis code used for the reported results. Sensitive identifiers will be removed before release, and identifying source paths remain excluded from the manuscript archive.

APPENDIX D. GRAPH CONSTRUCTION AND SEMANTIC VALIDITY

D.1. SOURCE-CONSTRAINED REPRESENTATION

Equation equation 1 represents ordinary evidence E, explicit absence premises SE, intermediate conclusions C and the saved answer. On the model side, conversion requires a saved successful response with nonempty reasoning and answer. The question-specific evidence catalog identifies recorded observations in the focal actor's view. An evidence ID locates a premise and its time; it is not permission to add information absent from the explanation.

The documented extraction principles link composition statements to source quotations, premise references and earlier conclusions. Actor identity, time, negation, uncertainty and epistemic status must be preserved, including source statements that conflict with the catalog. The selected C statement can contain antecedent–consequence content; it is not defined as one sentence or one atomic proposition. A support edge follows an expressed dependency, not textual adjacency or the analyst's judgment that an inference would be reasonable. C–C counts refer to conclusion support rather than terminal-answer links.

D.2. ABSENCE, COMPILATION, AND ELIGIBILITY

SE is reserved for an explicitly expressed missing behavior, support, information or goal. A missing catalog item does not license SE. "Not observed" and "did not occur" have different epistemic scope and must not be silently exchanged. Ordinary E and special-evidence counters also cannot be assumed identical to every structural evidence-role counter; the implemented graph representation governs the reported counts.

All reported Human and model graphs are produced by the same latest graph-construction pipeline and share the E/SE/C/answer schema in Equation equation 1. Structural consistency checks validate node references, dependency ordering and terminal membership, while semantic fidelity is evaluated separately by the blinded human audit in Appendix D.3.

D.3. HUMAN VALIDATION OF SOURCE-TO-GRAPH FIDELITY

Audit unit and judgments. Two independent human reviewers evaluated the same 120 blinded items, each pairing an `analysis_text` with its graph. Each reviewer made one binary judgment per item of whether the graph was faithful to that analysis text. The evaluation therefore contains 120 reviewed pairs and two judgments per pair, not 240 independent pairs. The reported aggregate outcomes are:

Measure	Count	Value
Reviewer 1: faithful	115/120	95.83%
Reviewer 2: faithful	117/120	97.50%
Exact inter-reviewer agreement	112/120	93.33%
Gwet's AC1	—	0.9287

The first two rows report the fraction of pairs each reviewer judged faithful; exact agreement and AC1 describe agreement between reviewers, not an external-gold-standard extraction accuracy. Eight items received different judgments. No adjudicated score or separate node/edge precision or recall is reported.

Scope. The audit supports fidelity of the common graph-construction pipeline to the analyzed text in the reviewed sample. It does not evaluate whether the explanation itself is factually correct in the environment or whether it reflects an agent's internal computation.

APPENDIX E. MEASUREMENT DEFINITIONS AND STATISTICAL PROTOCOL

E.1. HUMAN-LABEL COMPARISON

A participant–observer pair is eligible when both exported strings are nonempty. The main rule retains literal `NA` and prefixed NA values and gives each eligible annotation pair equal weight. Repeated responses remain separate; no majority vote is used. Graph availability does not determine eligibility for this human-label analysis.

Action root is the `Action...` class token. Own-intention root is the first hyphen-delimited predicate. Partner root removes the leading numeric estimator prefix before predicate extraction. Roots are stripped and case-folded. For intentions, the strict function applies Unicode NFKC, whitespace collapse and case folding to the complete string, retaining arguments and the partner prefix. For actions, strict intentionally returns the class root: observer arguments are stored separately and are not scored. Thus equal action columns do not demonstrate agreement on targets or parameters.

Let $\mathcal{I}_c$ contain the eligible pairs for channel c and ρ_c be its root parser. The reported rate is

$$A_c = |\mathcal{I}_c|^{-1} \sum_{(i,a)\in\mathcal{I}_c} \mathbf{1}\{\rho_c(y_{ia}^T) = \rho_c(y_i^P)\}.$$

Strict rates substitute the implemented strict function. Full partner-field results retain the target distinction in A.4. Pooling fields mixes targets and matching granularities and is not a single social-accuracy score.

Observer–observer agreement uses unordered, nonempty response pairs from different accounts at the same decision and channel. Its repeated-cell population and pair weights differ from the participant–observer sample. Different accounts are not verified independent people. No observer–observer interval is included in the executed statistical output.

E.2. ORDINARY AND EXPLICIT-ABSENCE PREMISES

The seven ordinary E classes are object information, agent information, self/other mental states, self/other values and social interaction relative to the focal actor. E shares divide class counts by the ordinary-E count of their source. The denominators are 5,399 Human, 21,158 Astra, 44,553 DeepSeek, 47,322 Gemini and 67,842 Claude E nodes. The classifier's complete ambiguous-case priorities remain unverified.

SE1/SE2 denote behavioral absence, SE3/SE4 social-support absence, and SE5/SE6 information-or-goal gaps. Equation equation 2 distinguishes a graph with at least one SE from the composition among SE nodes. An eligible zero-SE graph contributes zero; a missing graph record does not. The sidecar graph populations and node counts are retained in F.3. Expressed absence is not an environmental correctness or psychological-uncertainty score.

E.3. FROZEN CONCLUSION VOCABULARY

The coding unit is a complete C statement summarized as a structural gloss, not a newly selected evidence neighborhood. The registered participant-derived projection uses `jinaai/jina-embeddings-v3` (Sturua et al., 2024), encoding task `separation`, maximum length 256 and batch size 16, followed by PCA with 64 components, randomized solver and seed 0. Assignment uses the frozen `kmeans_k040_pca064` base vocabulary. Its recorded base settings are seed 0, batch size 2,048, 10 initializations and reassignment ratio 0.01. Applying this vocabulary is not a new cluster fit for each source, and E classes are not C clusters.

Global summaries use all saved direct base assignments, not a separate refinement output. Human has 2,567 labeled C nodes and 33 observed base labels; a configured size of 40 does not require every label to occur. Model direct-label populations are 10,515 Astra, 22,608 DeepSeek, 27,522 Gemini and 31,547 Claude. These differ from all-C structure rows by three Astra and five DeepSeek blank-label rows. Shared vocabulary IDs do not establish identical elicitation or assignment implementations. The separate structure-label calculation is retained in F.7 and does not replace the main direct-label results.

E.4. Aggregation, content extent, and support

Global C-label distributions pool labeled nodes within each declared source, rather than averaging normalized graph distributions. Jensen–Shannon divergence is measured in bits:

$$\mathrm{JS}(p,q) = \tfrac{1}{2}\,\mathrm{KL}_2(p\|m) + \tfrac{1}{2}\,\mathrm{KL}_2(q\|m), \qquad m = \tfrac{1}{2}(p+q).$$

Entropy $H(p) = -\sum_k p_k \log_2 p_k$ and effective labels $2^{H(p)}$ describe the same distributions. The saved summaries do not fully specify every smoothing choice. Human–model global distances and the 4,865-key model-only common set have different populations and are not a shared five-source ranking test.

Mean C/E and C–C counts are per C-bearing structure record. Node counts, support-edge counts, semantic-node proportions and graph-level prevalence weight different units; they are not combined into one sample count. Two exported Human metric views with 1,298 records each are not independent cohorts. Likewise, a model repeat is not a new participant. Similar mean C counts do not match graph-size distributions or task composition. C-node/E-class conditional Human comparisons remain unavailable because their key mapping is unresolved; the reported marginal profiles do not establish such an effect.

E.5. Shared-question estimates and uncertainty

The finite statistical analysis was author-approved after the original results, not preregistered. It uses 95% percentile intervals from 5,000 bootstrap resamples of source game files, seed 20260925. Focal views and all decisions from the same file stay together. This preserves within-game dependence while conditioning on the observed contributor panel; cross-game participant/session and observer dependence are not separately estimated.

Available model repeats are first averaged within each question. Sources are joined by game, focal actor, decision and channel; H contributes once per paired question. For metric z, the point estimate is the question-weighted mean

$$\widehat{\Delta} = |Q|^{-1} \sum_{i \in Q} (\bar{z}_i^{M} - z_i^{H}).$$

The resampling unit does not convert this into an equal-game point estimate. The model SE quantity is the within-question average of repeat-level presence, not whether any repeat contains an SE. Pair-specific intersections contain 1,167 H–Astra or 1,168 H–other-model questions over 41 game files; all cover action, own intention and partner estimation, not values. These populations are subsets of the 1,298 C-bearing H records, not all 1,325 sidecar graphs.

Human-label intervals use 36 game files. Pair weighting gives each eligible annotation pair equal weight; decision weighting first averages within a decision and then weights decisions equally. NA exclusion removes a pair if either side has exported NA status. All combinations are retained in F.2. Intervals are pointwise, not multiplicity-adjusted joint statements or causal effects. Actual available model repeats, rather than a hardcoded nominal count of three, define each mean. The corrected size analysis specifies **A**, equality of H's C count with the unrounded mean over available model repeats, and separately **B**, exact equality for each model-repeat graph. A averages differences over retained questions; B averages over retained question–repeat pairs, so questions with more matching repeats receive greater weight. Games, not pairs, remain bootstrap units. Equality of means in A does not require every model-repeat graph to have H's size. F.6 reports both rules and overlap coverage.

Appendix F. Supplementary results for the current analyses

F.1. Graph coverage and distinct populations

Each model has 5,256 requested slots (1,752 questions in three rounds). Table S3 separates successful payloads from missing slots. C-bearing structure counts are 5,202 Astra, 5,113 DeepSeek, 5,212 Gemini and 5,069 Claude. H has 1,325 sidecar graph keys, 1,298 C-bearing records and 2,567 C nodes. The 27-record difference reflects representation eligibility, not 27 missing trajectories. Successful payloads, sidecar graphs, C-bearing records and direct-label nodes retain their distinct

denominators; DeepSeek, for example, has 5,151 payloads, 5,144 sidecar keys and 5,113 structure records.

The H C-bearing inventory separates 468 action graphs (966 C nodes), 421 own-intention graphs (798), 341 partner-estimation graphs (632), and 68 value graphs (171). These sum to 1,298 graphs and 2,567 C nodes; they are not the per-channel counts of the original 1,752-question inventory or the 120-item human audit.

Table S3: Successful graph payloads by model and response round.

Model	r1	r2	r3	Total	Missing
Astra	1,735	1,734	1,735	5,204	52
DeepSeek	1,667	1,743	1,741	5,151	105
Gemini	1,729	1,741	1,742	5,212	44
Claude	1,688	1,692	1,689	5,069	187

The common-question graph intersections cover 89.91% or 89.98% of C-bearing H records and span 41 source games. Their available model-repeat rows number 3,491/3,432/3,488/3,386 for Astra/DeepSeek/Gemini/Claude. The corresponding numbers of questions with one, two and three available repeats are (2, 6, 1,159), (0, 72, 1,096), (0, 16, 1,152), and (3, 112, 1,053). These are observed repeat counts within the paired C-bearing subsets, not the original 5,256 requested slots per model. They differ from the 4,865 model-only round/question/channel keys used for model–model C-label distances. Exact person/session overlap across the human-label and graph samples remains unverified; these denominators are not counts of independent people.

F.2. Human correspondence and all executed sensitivities

The frozen annotation export has 663 completed rows, 94 assignments and 15 observer accounts across 60 trajectory–focal-agent keys. It contains 383 primary decision cells and 52 final-value-only cells. Primary multiplicities are 269 single, 32 double and 82 triple annotations, giving 579 rows; final-value-only multiplicities are 33 single, 6 double and 13 triple annotations, giving 84 rows. Final-value-only records are not additional primary-channel decisions.

Tables S4 and S5 retain all three primary channels and the pooled descriptive counts. Their partner and pooled rows are not restored as a common-target recovery claim. The implemented different-account, nonempty observer rule yields 278 pairs per primary channel; action strict remains class-only. Counts preserve the source rule, not majority voting.

Table S4: Complete participant–observer field correspondence under the implemented matching rules.

Field	Pairs	Root same	Root different	Root (%)	Strict same	Strict (%)
Action	579	191	388	32.99	191	32.99
Own intention	579	128	451	22.11	107	18.48
Partner field	579	185	394	31.95	88	15.20
Pooled	1,737	504	1,233	29.02	386	22.22

Table S5: Complete observer–observer agreement; accounts are not independently verified people.

Field	Pairs	Root same	Root (%)	Strict same	Strict (%)
Action	278	90	32.37	90	32.37
Own intention	278	71	25.54	45	16.19
Partner intention	278	102	36.69	89	32.01

Table S6 contains all 32 estimate/interval outputs, with root and strict side by side. “Retain” keeps literal NA; “exclude” removes a pair with NA on either side. Pair/cell counts refer to that row’s eligible population. Pooled cells count channel-specific decision cells rather than independent interactions.

No observer–observer interval was computed. The current statistical sample has 36 source game files; interpretation is conditional on the actual viewing and matching rules, not an immersion effect.

Table S6: Human-label estimates and game-clustered 95% intervals (percent).

Channel / NA	Pairs / cells	Weight	Root % [CI]	Strict % [CI]
Action / retain	579 / 383	Pair	32.99 [28.13, 39.01]	32.99 [28.13, 39.01]
Action / exclude	579 / 383	Pair	32.99 [28.13, 39.01]	32.99 [28.13, 39.01]
Action / retain	579 / 383	Decision	36.03 [29.74, 42.92]	36.03 [29.74, 42.92]
Action / exclude	579 / 383	Decision	36.03 [29.74, 42.92]	36.03 [29.74, 42.92]
Own intention / retain	579 / 383	Pair	22.11 [17.18, 27.85]	18.48 [13.86, 23.84]
Own intention / exclude	492 / 331	Pair	25.00 [19.61, 30.82]	20.73 [15.59, 26.44]
Own intention / retain	579 / 383	Decision	24.37 [19.01, 30.17]	19.71 [14.71, 24.98]
Own intention / exclude	492 / 331	Decision	27.49 [21.83, 33.38]	22.05 [16.61, 27.69]
Partner field / retain	579 / 383	Pair	31.95 [23.09, 42.61]	15.20 [10.13, 21.96]
Partner field / exclude	503 / 332	Pair	24.65 [17.41, 33.78]	17.50 [11.54, 25.49]
Partner field / retain	579 / 383	Decision	34.86 [26.84, 42.80]	18.49 [11.99, 25.62]
Partner field / exclude	503 / 332	Decision	27.26 [19.35, 35.74]	21.34 [14.01, 29.15]
Pooled fields / retain	1737 / 1149	Pair	29.02 [24.13, 34.84]	22.22 [18.76, 26.83]
Pooled fields / exclude	1574 / 1046	Pair	27.83 [23.36, 33.31]	24.21 [20.25, 29.35]
Pooled fields / retain	1737 / 1149	Decision	31.75 [27.26, 36.39]	24.75 [20.72, 29.30]
Pooled fields / exclude	1574 / 1046	Decision	30.54 [25.91, 35.53]	26.94 [22.36, 31.94]

Label-frequency diagnostics. The delivered confusion matrices reproduce 579 eligible pairs per channel and the 191 action/128 own-intention root matches. In these pair-weighted marginals, `ActionMoveTo` occurs 187 times in participant records and 210 times in observer labels; recorded `ActionWait` versus observer `ActionMoveTo` accounts for 25 mismatches. For own intention, `observe` occurs 59 versus 130 times and `help` 55 versus 20. The largest single off-diagonal cell is participant `NA` versus observer `observe` (35 pairs). There are 62 participant-NA and 30 observer-NA labels, with five jointly NA: excluding the 87-pair union gives the existing 492-pair sensitivity. Counts describe exported label choices, not verified intent errors.

The separate frequency export includes 84 blank final-value-only rows per source/channel, so it is not the 579-pair marginal used above. Its field called `n_decisions`=60 counts trajectory–focal-agent keys, not the 383 primary cells. Six own-intention source roots are serialized numeric strings (one `1`, five `2`); they are retained without reinterpreting them as semantic categories or changing the existing matching results. Complete supplied-label confusion counts are available with the statistical tables.

F.3. Evidence and explicit-absence composition

Table S7 reports all seven ordinary-E shares from the supplied integer counts. The class names refer to the focal actor, and rounded percentages are not renormalized. Table S8 retains the eligible graph counts and complete SE family counts; the corresponding prevalence and conditional percentages appear in main Table 2. Human's conditional profile has only 78 SE nodes. These unadjusted source-level profiles are distinct from the paired estimates in F.5.

Table S7: Ordinary-E semantic shares from the supplied counts (percent).

E category	Human (H)	Astra	DeepSeek	Gemini	Claude
`object_info`	19.43	17.71	15.50	13.60	16.92
`agent_info`	30.62	38.72	24.81	39.22	36.37

E category	Human (H)	Astra	DeepSeek	Gemini	Claude
self_mental	28.95	20.74	20.93	21.02	19.34
other_mental	10.34	10.87	14.51	9.85	10.18
self_value	7.26	6.80	19.61	11.29	12.56
other_value	0.76	0.01	0.78	0.27	0.57
social_interaction	2.65	5.15	3.86	4.75	4.05

Table S8: SE coverage and family counts. Percentages are reported in main Table 2.

Source	Eligible graphs	Graphs with SE	SE nodes	Behavioral absence	Support absence	Information/ goal gap
Human (H)	1,325	74	78	9	17	52
Astra	5,204	1,791	2,777	723	930	1,124
DeepSeek	5,144	2,160	3,888	1,397	990	1,501
Gemini	5,212	2,541	5,966	2,495	1,842	1,629
Claude	5,069	3,279	7,509	2,379	1,884	3,246

The supplied totals are 186,274 ordinary E nodes and 20,218 SE nodes. Recovered analysis texts are much shorter for Human than for models: mean character lengths are 16.7 for Human action explanations and 16.9 for Human own-intention explanations, versus 201–766 and 233–827 across the four model groups, respectively. These are saved analysis-text lengths, not tokenizer counts or a matched-length design. We therefore do not interpret the SE prevalence difference as length-independent. A represented absence is not automatically a warranted absence, and these counts do not complete an environmental-grounding audit.

F.4. Conclusion composition and graph extent

Table S9 retains the original direct-label population used in main Table 3. The source distributions pool labeled C nodes. Model–model distances in Table S10 use the separate 4,865-key common-model set; they are not substituted for global H–model distances. Higher entropy or more observed categories are not measures of reasoning quality.

Table S9: Global C-codebook summaries.

Source	C nodes	Entropy (bits)	Effective labels	JS to Human
Human (H)	2,567	4.676	25.56	–
Astra	10,515	5.022	32.49	0.1360
DeepSeek	22,608	4.980	31.55	0.1731
Gemini	27,522	4.822	28.29	0.1444
Claude	31,547	5.012	32.26	0.1204

Table S10: All six supplied model–model C-label JS divergences (bits).

Model 1	Model 2	JS (bits)
Gemini	DeepSeek	0.0401
Gemini	Astra	0.0519
DeepSeek	Astra	0.0323
Claude	Gemini	0.0408
Claude	DeepSeek	0.0265
Claude	Astra	0.0218

Table S11 gives source-level C-bearing counts and the associated node/link means. Table S12 retains every recomputed channel mean, including value graphs. The second exported Human view is not counted as a second source. Channel means include different records from the common-question

comparisons below; their aggregation is not treated as a size-adjusted effect. Exact segmentation, evidence-role counters and direct-support conventions remain bounded by Appendix D.

Table S11: Node and support-link counts per C-bearing structure record.

Source	Records	Mean C	Mean E	Mean C–E	Mean C–C
Human (H)	1,298	1.98	4.20	4.20	0.97
Astra	5,202	2.02	4.60	4.75	0.43
DeepSeek	5,113	4.42	9.41	10.31	2.45
Gemini	5,212	5.28	10.22	11.28	2.84
Claude	5,069	6.22	14.87	16.60	4.09

Table S12: Recomputed channel means of saved graph counts.

Source	Channel	Mean C	Mean E	Mean C–C
Human (H)	Action	2.0641	4.4573	1.0641
Human (H)	Partner estimate	1.8534	4.1026	0.8563
Human (H)	Own intention	1.8955	3.7696	0.8931
Human (H)	Value	2.5147	5.5441	1.2794
Astra	Action	1.8078	3.9479	0.3719
Astra	Partner estimate	1.9169	4.6006	0.2974
Astra	Own intention	2.0546	4.3967	0.7370
Astra	Value	2.9665	7.4591	0.0168
DeepSeek	Action	4.2634	9.2398	1.9488
DeepSeek	Partner estimate	4.3424	8.3476	2.7141
DeepSeek	Own intention	4.6633	10.5636	2.9342
DeepSeek	Value	4.4135	9.6629	1.6449
Gemini	Action	5.4594	10.7244	2.6765
Gemini	Partner estimate	5.3844	9.9058	2.8450
Gemini	Own intention	4.7654	9.8405	2.8941
Gemini	Value	6.0588	10.8761	3.1429
Claude	Action	6.1478	14.7063	3.5300
Claude	Partner estimate	6.0334	13.2318	4.3995
Claude	Own intention	6.4549	16.3238	4.4958
Claude	Value	6.3363	15.9535	3.6040

F.5. Paired comparisons and channel-specific intervals

Tables S13 and S14 retain all reported common-question contrasts under E.5. The H and model entries are means on each row's population. Differences are model minus H: C–C is measured in links per paired question, whereas SE differences and intervals are percentage points. Each model SE mean averages repeat-level presence within question. Pair-specific intersections are not one identical population for ranking all four models. The intervals are pointwise; graph-construction fidelity is evaluated separately in Appendix D.3.

Table S13: Paired graph contrasts on common decision questions.

Model	Metric	Q	Games	H	Model	Difference [95% CI]
Astra	C–C	1,167	41	0.962	0.481	-0.481 [-0.555, -0.402]
Astra	SE (pp)	1,167	41	4.80	28.01	23.21 [18.79, 27.49]
DeepSeek	SE (pp)	1,168	41	4.88	35.02	30.14 [25.94, 34.56]
Gemini	SE (pp)	1,168	41	4.88	43.29	38.41 [33.65, 42.89]
Claude	SE (pp)	1,168	41	4.88	59.96	55.08 [50.07, 59.76]

Table S14: All channel-specific paired contrasts. Differences are model minus H.

Model	Metric	Channel	Q	Games	Difference [95% CI]
Astra	C–C	Action	443	40	-0.697 [-0.827, -0.571]
Astra	C–C	Partner estimate	320	39	-0.579 [-0.744, -0.419]
Astra	C–C	Own intention	404	40	-0.168 [-0.327, -0.010]
Astra	SE (pp)	Action	443	40	17.53 [12.96, 21.89]
Astra	SE (pp)	Partner estimate	320	39	18.28 [12.37, 23.93]
Astra	SE (pp)	Own intention	404	40	33.33 [27.06, 39.57]
DeepSeek	SE (pp)	Action	443	40	22.76 [16.29, 29.80]
DeepSeek	SE (pp)	Partner estimate	321	39	41.17 [35.26, 47.04]
DeepSeek	SE (pp)	Own intention	404	40	29.46 [24.41, 34.54]
Gemini	SE (pp)	Action	443	40	30.36 [24.92, 35.77]
Gemini	SE (pp)	Partner estimate	321	39	36.50 [29.89, 42.70]
Gemini	SE (pp)	Own intention	404	40	48.76 [42.34, 55.20]
Claude	SE (pp)	Action	443	40	46.58 [39.62, 53.41]
Claude	SE (pp)	Partner estimate	321	39	65.78 [60.75, 70.94]
Claude	SE (pp)	Own intention	404	40	55.90 [49.88, 61.76]

F.6. EXACT-C COMPARISONS

The local replay now uses unrounded equality. Rule A retains questions satisfying $C_i^H = \bar{C}_i^M$, where $\bar{C}_i^M$ averages available model repeats. Rule B retains individual question–repeat pairs satisfying $C_i^H = C_{ir}^M$. Both use the source-game bootstrap from E.5, with 5,000 resamples and seed 20260925. The delivered rule names and all eight outputs are preserved in Table S15; the column called `n_questions` in the backend summary counts repeat pairs for B, which are distinguished from unique questions here.

Table S15: Exact-C overlap and model-minus-H C–C differences. A is question-level unrounded mean equality; B is individual-repeat equality. Intervals are pointwise.

Rule	Model	Units / unique Q	Games	Coverage	Difference [95% CI]
A	Astra	326 / 326	40	27.93%	-0.383 [-0.436, -0.339]
A	DeepSeek	8 / 8	6	0.68%	-0.938 [-1.389, -0.583]
A	Gemini	8 / 8	7	0.68%	-0.375 [-0.767, -0.048]
A	Claude	0 / 0	0	0.00%	—
B	Astra	1,275 / 527	41	36.52%	-0.395 [-0.434, -0.355]
B	DeepSeek	139 / 109	32	4.05%	-0.712 [-0.853, -0.568]
B	Gemini	96 / 72	30	2.75%	-0.417 [-0.581, -0.278]
B	Claude	13 / 11	9	0.38%	-0.385 [-1.091, 0.077]

For A, eligible denominators are 1,167 Astra–H questions and 1,168 for each other model. For B they are 3,491/3,432/3,488/3,386 available repeat pairs, respectively. Thus B's 1,275 Astra pairs are 527 distinct questions, not 1,275 independent questions. The A result averages one difference per retained question; B weights every retained repeat pair equally. An H graph can appear in several B pairs, all within its resampled game. Neither rule creates new independent participants.

For Astra, A retains 326 questions (27.93%) over 40 games and yields -0.383 links [-0.436, -0.339]; B yields -0.395 [-0.434, -0.355]. Under A, 282 retained H graphs have two C nodes, 43 have one and one has three, restricting the size range represented. DeepSeek and Gemini each retain only eight A questions; Claude retains none. Their B overlaps are also sparse, and Claude's B interval spans zero. These results support the selected H–Astra contrast, not a full-population ranking or a universal source effect at every graph size.

The prior rounded-mean rule selected 428 Astra questions, of which 102 had unequal unrounded means. Its reported -0.390 diagnostic is superseded here by explicitly distinct A/B definitions, rather than renamed as exact. The primary all-common-question and global tables are unchanged.

Answer-matched robustness. A separate offline sensitivity uses exact normalized-answer equality together with exact C count in the joined model-round/channel strata. For Astra, the mean C–C

difference remains negative as the restrictions tighten: -0.665 over all joined entries, -0.422 for answer-matched entries, -0.697 for exact-C entries, and -0.373 when both answer and C count match. The last subset has only 8.0% mean coverage across the nominal round/channel question slots (512 retained entries) and is therefore descriptive rather than a replacement for the primary paired estimate. Joint answer-plus-exact-C coverage is much smaller for DeepSeek (0.52%), Gemini (0.75%), and Claude (0.09%), so no cross-model ranking is inferred from this sensitivity.

Astra subset	Mean coverage	Mean C–C difference (model–H)
All joined	68.4%	-0.665
Answer matched	20.8%	-0.422
Exact C	25.4%	-0.697
Answer matched + exact C	8.0%	-0.373

C–C topology sanity check. The saved sidecar topology also distinguishes an edge-count identity from a forced chain. Human graphs satisfy $\mathrm{CC} = C - 1$ in 98.6% of C-bearing records, but only 38.8% meet the stricter non-branching single-chain candidate rule; Astra is 51.0% and 38.5%, respectively. Thus equality of the C–C edge count with $C - 1$ is not equivalent to a fixed-chain topology. These are descriptive structure audits, not additional inferential tests.

F.7. Separate structure-label recomputation

The auxiliary calculation in Table S16 uses saved structure-table labels, unsmoothed base-2 entropy and midpoint JS with zero-frequency categories retained; it does not refit embeddings or the codebook. Astra contributes three additional blank-label rows and DeepSeek five; the supplied computation counts the blank key as a separate category. Those blank keys are not a semantic class. This population remains separate from the direct-label distributions in Table S9 and main Table 3; no denominator or source selection is overwritten. The calculation does not provide a semantic-fidelity audit or a matched-distribution test.

Table S16: Separate structure-label recomputation, including explicitly counted blank-label rows.

Source	Structure rows	Blank keys	Entropy (bits)	JS to H (bits)
Human (H)	2,567	0	4.675830	—
Astra	10,518	3	5.024165	0.136150
DeepSeek	22,613	5	4.981640	0.173181
Gemini	27,522	0	4.822420	0.144397
Claude	31,547	0	5.011554	0.120382

F.8. Expanded-corpus reasoning-type distribution

Section 5.3 analyzes the final materialized C-type table, where C denotes the R units of the main text. This export contains 73,604 valid labeled nodes across 22,146 graph records. $\text{Human}_{\text{full}}$ contributes 6,619 graphs and 12,959 nodes; Astra 5,202 and 10,515; DeepSeek 5,113 and 22,608; Gemini 5,212 and 27,522. Claude is not represented in this export. Unknown and empty labels are excluded. These graph counts are not independent participant counts, and $\text{Human}_{\text{full}}$ is not the original 1,298-graph H subset. No new graph extraction, model query, embedding, or codebook fitting is performed.

Weighting and divergence. Pooled-node shares count every labeled node equally. Equal-graph shares first divide each graph's label counts by its total valid labeled-node count and then average these normalized vectors within source. Both use the same 40 base IDs. Base-2 JS divergence is computed as the square of the Jensen–Shannon distance returned by the supplied analysis script. Table S17 retains all six source pairs; the three $\text{Human}_{\text{full}}$ comparisons appear in Figure 3. These expanded-corpus results do not replace Tables S9 or S16.

Table S17: Expanded-corpus JS divergence (bits) over all 40 reasoning types.

Source 1	Source 2	Pooled nodes	Equal graph weight
$\text{Human}_{\text{full}}$	Astra	0.0444	0.0361
$\text{Human}_{\text{full}}$	DeepSeek	0.0520	0.0496
$\text{Human}_{\text{full}}$	Gemini	0.0859	0.0809
Astra	DeepSeek	0.0323	0.0358
Astra	Gemini	0.0519	0.0506
DeepSeek	Gemini	0.0401	0.0365

Complete label composition. Table S18 reports all 40 types and pooled-node shares. Display names are the saved codebook glosses (not independently adjudicated semantic categories). The six types in Figure 3 are illustrative rather than a complete list or a prespecified significance screen. The delivered CSV additionally retains every equal-graph share and exact node count.

Table S18: All 40 reasoning types in the expanded corpus: pooled-node percentages. H-full is $\text{Human}_{\text{full}}$.

ID	Codebook gloss	H-full	Astra	DeepSeek	Gemini
00	Perceptual monitoring inference	4.50	3.95	2.95	3.47
01	Cooperative Goal Adoption	2.83	1.24	2.76	2.21
02	handoff subgoal inference	2.04	3.51	2.50	1.66
03	Target Access	3.61	4.99	6.01	10.21
04	Activity From Actions	0.76	1.61	0.73	1.16
05	Keyed Access	2.77	2.11	2.12	2.86
06	Task-object engagement	2.45	1.17	1.73	2.15
07	Container Access Barrier	2.96	1.45	1.24	2.47
08	Shared Attention Cueing	2.86	1.10	3.95	2.64
09	Deictic cue interpretation	1.81	1.10	1.59	1.33
10	Joint Play Framing	0.82	0.41	0.50	0.22
11	Help-Motivation Gating	2.70	2.26	3.20	1.28
12	Object Acquisition	2.33	3.32	1.71	4.32
13	social intent inference	2.78	2.34	4.13	3.08
14	disentangled social helpfulness	1.61	1.21	2.36	1.34
15	Directed Greeting Engagement	1.71	5.96	6.73	7.78
16	Joint-Activity Alignment	1.55	1.21	2.17	2.71
17	Targeted attention inference	2.94	3.16	3.49	1.64
18	Access obstacle affordance	2.74	2.77	1.79	1.71
19	Object-task continuation	1.99	1.83	1.42	0.44
20	Container Access Affordance	3.72	2.33	3.13	3.60
21	Intention State Updating	4.71	7.02	8.54	7.86
22	Container Access Affordance	2.74	3.65	2.38	4.63
23	help-target inference	3.86	3.92	3.64	3.42
24	Dyadic partner necessity	1.78	1.44	0.71	0.99
25	Social-object orientation	1.94	0.48	0.81	0.23
26	Search Goal Updating	4.07	4.31	3.24	3.39
27	Conservative intent persistence	3.87	4.42	3.71	1.27
28	Two-Player Affordance	1.04	0.00	0.03	0.01
29	Trait Profile Inference	1.75	2.52	4.65	2.98
30	Contextual cue uptake	1.65	0.95	0.41	0.31
31	Directed Greeting Cue	1.81	2.76	1.59	1.03
32	Two-Player Play Affordance	2.04	1.55	1.23	1.91
33	Task Nonparticipation	3.40	3.70	3.09	5.61
34	Container access bottleneck	4.20	4.20	3.18	1.78
35	Placement task state	2.71	4.54	2.38	2.19
36	Cooperative task alignment	2.28	2.79	2.29	3.20
37	Help-Request Grounding	1.44	1.09	0.64	0.35
38	Joint Participation Framing	2.37	0.64	0.35	0.21
39	Directed social acknowledgment	0.89	1.02	0.95	0.35

Exploratory graph bootstrap. The supplied analysis independently resamples whole graphs within each source, 1,000 times with seed 20260926, to form pointwise 95% percentile intervals for equal-graph-share differences. Table S19 uses observed differences calculated from the saved equal-graph shares; its endpoints come directly from the supplied bootstrap CSV. The CSV's `difference_pp` field is the mean across bootstrap draws, which can differ slightly from the observed point estimate.

This bootstrap retains within-graph node dependence but does not jointly resample games, participants, or response repeats. Its intervals are exploratory and are neither multiplicity-adjusted nor interchangeable with the 5,000-resample source-game intervals of Appendix E.5. Equal graph weighting reduces the influence of large graphs but does not match task mix or remove source-specific measurement differences. The node-level chi-square output in the quick audit is not used as an inferential result because nodes within a graph or game need not be independent.

Table S19: Selected equal-graph-share differences relative to $\text{Human}_{\text{full}}$ (percentage points), with exploratory graph-bootstrap intervals.

Model	Reasoning type	Difference	95% interval
Astra	Directed Greeting Engagement	+4.63	[4.05, 5.25]
DeepSeek	Directed Greeting Engagement	+5.22	[4.74, 5.72]
Gemini	Directed Greeting Engagement	+6.22	[5.71, 6.73]
DeepSeek	Trait Profile Inference	+3.17	[2.84, 3.52]
Gemini	Target Access	+5.75	[5.17, 6.30]