Category-Theoretic Extension of Mutual Understanding to Group Communication

Stan Miasnikov, September 2025
stanmiasnikov@gmail.com

Abstract

We extend the Recursive Consciousness framework's mutual understanding metric [1] from two agents to an N-agent group (discussion). Using category theory, we model $N > 2$ semantic spaces connected via a single shared symbolic channel. Each conversation turn has exactly one incoming interpretation functor (I) into the symbolic channel and multiple outgoing meaning functors (M) to listeners. We derive a **group understanding** score that generalizes the pairwise metric by requiring alignment across all agents. The proposed formulation combines information-theoretic alignment (Jensen-Shannon divergence [2]), semantic similarity (embedding-based distance), and pragmatic convergence (iterative stability) for every pair of agents, aggregated in a non-compensatory way (multiplicatively). Normalization uses the strict geometric mean over pairs; turn weights are an explicit design choice. Two scenarios are treated: (i) agent-generated query and potentially evolving initial understanding, and (ii) external (to the group) query with static ground truth (GT).

1 Introduction

Understanding in a group context requires more than just pairwise agreement; it demands a shared **common ground** among all participants [3]. If even one agent misconstrues the query or topic, the group as a whole lacks true consensus. Building on our prior work modeling two-agent communication and mutual understanding, we now address the multi-agent case. We consider N agents $A_1, A_2, \ldots, A_N$ engaging in a *discussion* about a query or statement. The Recursive Consciousness (RC) framework [4] provides the setting: each agent A_i has its own semantic category $\mathcal{C}_{\text{sem}}^{(i)}$ (encapsulating that agent's internal meanings or knowledge state), and all agents share a common symbolic context category $\mathcal{C}_{\text{sym}}$ [1] through which communication occurs.

Formally, each agent has an **interpretation functor** $I_i : \mathcal{C}_{\text{sem}}^{(i)} \to \mathcal{C}_{\text{sym}}$ that encodes its semantic objects into $\mathcal{C}_{\text{sym}}$ (phrases, symbols, messages, actions, etc.), and a **meaning functor** $M_i : \mathcal{C}_{\text{sym}} \to \mathcal{C}_{\text{sem}}^{(i)}$ that decodes any message from the common channel into that agent's semantic space.

For each agent A_i, we assume an adjunction $(I_i \dashv M_i)$ between $\mathcal{C}_{\text{sem}}^{(i)}$ and $\mathcal{C}_{\text{sym}}$, as in the two-agent case introduced in [1]. Inter-agent communication is the composite $\Phi_{i\to j} = M_j \circ I_i : \mathcal{C}_{\text{sem}}^{(i)} \to \mathcal{C}_{\text{sem}}^{(j)}$ (not an adjunction in general). Now, however, $\mathcal{C}_{\text{sym}}$ acts as a hub for *all* communication: any agent's message is placed into $\mathcal{C}_{\text{sym}}$ and potentially received by all others via their respective M_j. There are no direct $A_i \to A_j$ semantic mappings; all exchanges occur through the shared context.

2 Group Understanding Metric

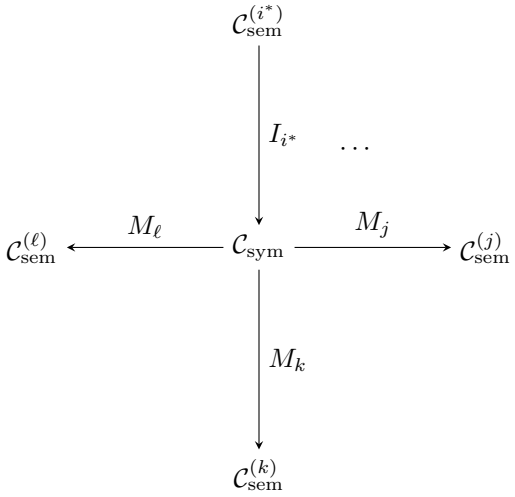
As in the two-agent case, information loss can occur: each I_i may be non-faithful (many distinct semantic ideas becoming one message) and M_j may not invert I_i perfectly, especially if the agents' internal categories differ. The question we address is how to quantify *group mutual understanding* - i.e., how well the entire group understands the query or topic being discussed. Intuitively, we want a metric U_{group} that is high (near 1) **only if all agents have essentially the same understanding** of the query or the external GT, and low if any agent is significantly out of sync. In other words, every agent's interpretation should align with the *intended meaning* and with every other agent's interpretation. This is related to the concept of *common knowledge*: not only must each agent understand the query, but they should also be aware (at least implicitly) that the others understand it. Our metric does not explicitly model higher-order knowledge; it requires all pairwise alignments to be strong, ensuring a common ground.

[1] $\mathcal{C}_{\text{sym}}$ denotes the shared symbolic/syntactic category, synonymous with $\mathcal{C}_{\text{out}}$ (syntactic output) in prior work.

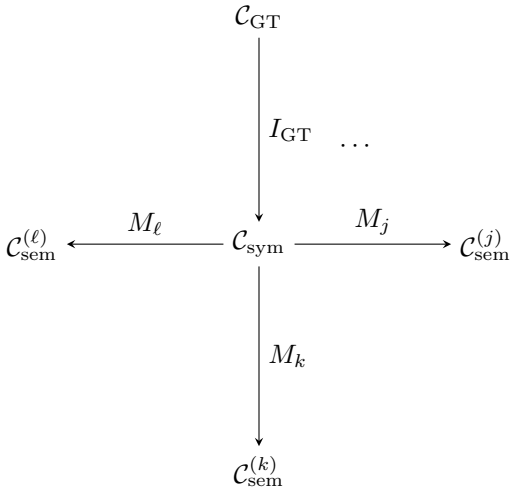
2.1 Setup and Diagrams

We assume one shared symbolic context category $\mathcal{C}_{\text{sym}}$ and N semantic categories $\{\mathcal{C}_{\text{sem}}^{(i)}\}_{i=1}^{N}$. At any given conversation *turn*, there is exactly one incoming I-arrow into $\mathcal{C}_{\text{sym}}$ (from the query source) and multiple outgoing M-arrows from $\mathcal{C}_{\text{sym}}$ to the listening agents. No direct arrows connect $\mathcal{C}_{\text{sem}}^{(i)}$ to $\mathcal{C}_{\text{sem}}^{(j)}$.

Case A (Agent-generated Query/Understanding, evolving):



Case B (External Query/GT, static):



2.2 Group Understanding Metric

Suppose the discussion progresses in T discrete turns (messages), where in each turn one agent communicates a message into the symbolic context, while other agents apply their meaning functors to interpret it. We denote $s_i^{(n)} \in \mathcal{C}_{\text{sem}}^{(i)}$ as the semantic state or interpretation held by agent A_i after the n-th message has been processed. (For example, $s_1^{(0)}$ might be agent A_1's initial query posed in semantic form, which is then encoded by I_1 and broadcast; $s_j^{(1)}$ for $j \neq 1$ would be each other agent's understanding of that query after turn 1, and so on). Each agent's knowledge or belief about the query can be represented as a probability distribution $P_i^{(n)}$ over relevant semantic possibilities (this comes from viewing each semantic object as a proposition in a Kripke frame, as in [1]).

2.2.1 Operationalizing Semantic States via Agent Restatements

To compute the semantic states $s_i^{(n)}$ and belief distributions $P_i^{(n)}$ in practice, we elicit explicit restatements from each agent after every conversation turn. Specifically, at the end of turn n, each agent A_i is prompted to restate its current understanding of the query or topic in natural language (e.g., "Summarize

your understanding of the query."). In practice we compute $\|s_i^{(n)} - s_i^{(n-1)}\|$ via the shared embedding: $\|E(\text{restatement}_i^{(n)}) - E(\text{restatement}_i^{(n-1)})\|_2$.

These restatements are processed via a shared embedding functor E (approximating the faithful norm-preserving embeddings from Proposition 1 in [6]), yielding vector representations. For $P_i^{(n)}$, we recommend a calibrated distribution over a shared hypothesis set $\mathcal{H}^{(n)} = \{h_1, \ldots, h_m\}$ of meaning candidates (e.g., paraphrase/cluster centroids), with

$$P_i^{(n)}(h_k) \propto \exp\big(\alpha \cos(\, E(\text{restatement}_i^{(n)}),\, E(h_k)\,)\big),$$

with $\alpha > 0$ a scaling factor (e.g., inverse temperature), followed by normalization and (optionally) temperature/Dirichlet smoothing to avoid zeros. This aligns D_{JS} with "belief over meanings" on a common support rather than token probabilities.

2.2.2 Pragmatic Retention

We define each agent's pragmatic retention at turn n as:

$$\kappa_i^{(n)} \;=\; \exp\Big(-\tfrac{1}{\tau}\,\|s_i^{(n)} - s_i^{(n-1)}\|\Big) \in [0,1],$$

where $\|s_i^{(n)} - s_i^{(n-1)}\|$ is the enriched Banach norm (or its embedding approximation via E) (cf. [6, Prop. 1]); and $\tau > 0$ is a temperature parameter (e.g., set by variation, typically $\tau \in [0.5, 2]$ in practice) controlling the sensitivity of retention to changes in semantic state. $\kappa_i^{(n)} \in [0,1]$ indicates how little agent A_i changed its understanding on that turn; $\kappa_i^{(n)} = 1$ means no change (perfect stability) and lower values mean a bigger update (suggesting prior misunderstanding or new insight).

2.2.3 Strict geometric per-pair aggregation

Let $K = \binom{N}{2}$ be the number of unordered pairs. For a given turn n, we define the *per-turn group understanding score* by the strict geometric mean over all pairwise fidelities, multiplied by a pragmatic stability factor:

$$\mathcal{U}_{\text{group}}^{(n)} \;=\; \left(\prod_{1 \le i < j \le N} u_{ij}^{(n)}\right)^{1/K} \cdot \left(\prod_{i=1}^{N} \kappa_i^{(n)}\right)^{1/N}. \tag{1}$$

Here:

$$d_{\text{sem}}(s_i, s_j) \;=\; \frac{1 - \cos\big(E(s_i), E(s_j)\big)}{2} \in [0,1], \qquad 1 - d_{\text{sem}} \;=\; \frac{1 + \cos\big(E(s_i), E(s_j)\big)}{2},$$

and

$$u_{ij}^{(n)} \;=\; \Big(1 - D_{\text{JS}}(P_i^{(n)} \,\|\, P_j^{(n)})\Big)\Big(1 - d_{\text{sem}}(s_i^{(n)}, s_j^{(n)})\Big).$$

We have $u_{ij}^{(n)} \in [0,1], u_{ij}^{(n)} = u_{ji}^{(n)}, u_{ii}^{(n)} = 1$ by convention. The product runs over all unordered pairs (i,j). Thus, for each pair $i < j$, if $P_i^{(n)}$ and $P_j^{(n)}$ agree (small D_{JS}), then $(1 - D_{\text{JS}})$ is near 1; if they disagree strongly, it approaches 0. We use D_{JS} with logarithm in base 2, so $D_{\text{JS}} \in [0,1]$. Similarly, if $s_i^{(n)}$ and $s_j^{(n)}$ are geometrically close (cosine ≈ 1), then $1 - d_{\text{sem}} \approx 1$; if they are *orthogonal* (cosine $= 0$), then $1 - d_{\text{sem}} = \frac{1}{2}$; and if they are *opposite/antipodal* (cosine ≈ -1), then $1 - d_{\text{sem}} = 0$. We assume the vectors $E(s)$ are ℓ_2-normalized, so $\cos(E(s_i), E(s_j)) \in [-1, 1]$. The factor $\big(\prod_i \kappa_i^{(n)}\big)^{1/N}$ is the geometric mean of per-agent stability and is near 1 only if *every* agent made small adjustments in that turn.

For $N = 2$ (so $K = 1$), $\mathcal{U}_{\text{group}}^{(n)}$ reduces to the dyadic product of information and semantic fidelity, multiplied by the geometric mean of κ's, i.e., the original two-agent mutual understanding per-turn score.

Why a geometric mean? The matrix of pairwise fidelities $[u_{ij}^{(n)}] \in [0,1]^{N\times N}$ can be viewed as a *fuzzy relation* on the agent set whose values live in $[0,1]$ equipped with the usual multiplicative structure. In such $[0,1]$-valued settings, combining independent alignment constraints corresponds to multiplying their strengths, so folding over all unordered pairs yields the product $\prod_{i<j} u_{ij}^{(n)}$. Passing to the log domain turns this multiplicative fold into an additive average,

$$\frac{1}{K}\sum_{i<j}\log u_{ij}^{(n)},$$

and exponentiating gives the symmetric, scale-free aggregate

$$\exp\Big(\tfrac{1}{K}\sum_{i<j}\log u_{ij}^{(n)}\Big) \;=\; \Big(\prod_{i<j} u_{ij}^{(n)}\Big)^{1/K}.$$

This choice is *non-compensatory*: any misaligned pair ($u_{ij}^{(n)} \approx 0$) drives the group score down, matching the conjunctive nature of consensus. The same multiplicative combination appears in categorical probability/Markov-kernel formalisms, where independent channels compose by multiplying likelihoods (e.g., [5]).

2.2.4 Normalization and log-additivity.

The Kth-root yields the strict geometric mean over pairs:

$$\left(\prod_{i<j} u_{ij}^{(n)}\right)^{1/K} \;=\; \exp\Big(\frac{1}{K}\sum_{i<j}\log u_{ij}^{(n)}\Big),$$

which aligns with log-additivity in the Banach/Hilbert embedding and avoids pair-count bias. Other monotone power rescalings are possible, but the strict pairwise geometric mean is the natural normalization at the pair level. When any $u_{ij}^{(n)} = 0$, the product is 0; the logarithmic form is interpreted in the standard limiting sense or implemented with a small ε-smoothing in practice (e.g. $\max(u_{ij}^{(n)}, \varepsilon)$, where, for example, $\varepsilon = 10^{-6}$).

2.2.5 External GT variant.

When a static ground truth (GT) exists with fixed meaning ($\kappa_{\mathrm{GT}} = 1$), we include the N GT-pairs along with the K agent-agent pairs. The per-turn score becomes:

$$\mathcal{U}_{\mathrm{group,GT}}^{(n)} \;=\; \left(\prod_{i=1}^{N} u_{i,\mathrm{GT}}^{(n)} \cdot \prod_{1\le i<j\le N} u_{ij}^{(n)}\right)^{1/(K+N)} \cdot \left(\prod_{i=1}^{N}\kappa_i^{(n)}\right)^{1/N}, \tag{2}$$

where

$$u_{i,\mathrm{GT}}^{(n)} \;=\; \Big(1 - D_{\mathrm{JS}}(P_i^{(n)} \,\|\, P_{\mathrm{GT}})\Big)\Big(1 - d_{\mathrm{sem}}(s_i^{(n)}, s_{\mathrm{GT}})\Big).$$

The $1/(K+N)$ normalizes over all (agent-agent + agent-GT) pairs, treating GT as an external 'agent' for balance. This enforces alignment to GT and (therefore) to one another.

2.3 Discussion-level score and weights

Over T turns, choose nonnegative weights w_n with $\sum_{n=1}^{T} w_n = 1$, typically increasing to emphasize late-turn convergence (e.g., linear). The discussion-level scores are

$$U_{\mathrm{group}} \;=\; \sum_{n=1}^{T} w_n \mathcal{U}_{\mathrm{group}}^{(n)}, \qquad U_{\mathrm{group,GT}} \;=\; \sum_{n=1}^{T} w_n \mathcal{U}_{\mathrm{group,GT}}^{(n)}. \tag{3}$$

Note that while the per-turn aggregator is non-compensatory across agents, the across-turn aggregator in (3) is compensatory in time (by design); a geometric time-aggregator is possible if non-compensation across turns is desired.

Weight choice is a design decision. Absent additional structure, there is no general proof of optimality for a particular w_n; any monotone schedule with $\sum w_n = 1$ is admissible. The flexibility of the w_n choice allows the metric to be adapted for different analytical goals, such as prioritizing final-state consensus (recency weighting) versus evaluating the entire conversational process (uniform weighting). For example, uniform weights $w_n = \frac{1}{T}$ may be suitable for assessing the overall quality of a conversational process, while recency-weighted schemes are better for evaluating final-state consensus.

By construction, $U_{\text{group}} \in [0, 1]$ is non-compensatory across agents: a serious misalignment involving any one agent will inject a near-zero factor into the product in Eq. (1), pulling the score down. Only if *every* pair achieves high alignment (and all agents are individually stable) will $\mathcal{U}^{(n)}_{\text{group}}$ be high. Thus U_{group} captures the intuitive notion that the group understands the query well *iff* all members are on the same page.

2.3.1 Design choice rationale

We prefer the full product (Eq. 1) with strict geometric mean (pair-level normalization) over an additive average. The product is conjunctive (AND-like): it approaches 1 only when every pair aligns; any misaligned pair (factor $\ll 1$) lowers the result sharply. An additive average would be compensatory and could mask a single agent's deep misunderstanding. One can also consider the matrix $[u^{(n)}_{ij}]$ for all pairs (i, j) after turn T:

$$\mathbf{U}_{ij} = \sum_{n=1}^{T} w_n \, u^{(n)}_{ij}.$$

This matrix representation can provide additional insights into pairwise alignments and potential sources of misalignment within the group.

3 Theoretical Foundations and Implications

3.1 Categorical semantics and alignment

The enrichment of semantic categories over Banach spaces from our previous work carries over to the multi-agent case. Each $\mathcal{C}^{(i)}_{\text{sem}}$ is enriched, allowing distances $\|x - y\|_i$ within each agent's semantic space. By Proposition 1 of [6], we assume a (practical approximation of) faithful embedding functor $E_i : \mathcal{C}^{(i)}_{\text{sem}} \hookrightarrow \mathbf{Ban}$ for each agent's category, so that semantic differences correspond to vector-space distances (at least in a high-dimensional embedding space). We assume a shared embedding functor E used for all agents in metric computations; Proposition 1 in [6] justifies that each agent admits a faithful embedding, and in practice we instantiate a common E to compare agents in one space. We further assume that all agents share a comparable embedding representation for the *symbolic* content in $\mathcal{C}_{\text{sym}}$ (e.g., identical token embeddings). Under these assumptions, the semantic factor $1 - d_{\text{sem}}(s_i, s_j) = \frac{1+\cos(E(s_i),E(s_j))}{2}$ is a stable semantic proximity: it maps identity to 1, orthogonality to 0.5, and opposite to 0.

Meanwhile, Theorem 1 in the appendix of [6] (a Pinsker-type inequality) asserts that there exists a constant $C > 0$ such that

$$D_{\text{JS}}(P_i \,\|\, P_j) \;\geq\; C \left\| \left(E \circ M_j \circ I_i\right) - \left(E \circ \text{id}_{\mathcal{C}^{(i)}_{\text{sem}}}\right) \right\|^2_{\text{op}}$$

on the relevant subspace (i.e., after identifying semantics via the shared embedding E). This relates the Jensen-Shannon divergence between two agents' belief distributions to the squared norm of the deviation of the communication functor $\Phi_{i\to j} = M_j \circ I_i$ from the identity after embedding. Intuitively, if A_i could transmit meanings to A_j with zero distortion ($\Phi_{i\to j} = \text{id}$), then P_i and P_j would coincide and $D_{\text{JS}} = 0$.

3.2 Convergence and practical use

Let $S = \prod_{i=1}^{N} \mathcal{C}^{(i)}_{\text{sem}}$ with the product metric $\|s\|_\infty = \max_i \|s_i\|_i$. If the joint update $T\colon S \to S$ induced by (I_i, M_i) and turn-taking is a contraction on S, then by the Banach fixed-point theorem, there exists a unique fixed point $s^* \in S$ such that $T(s^*) = s^*$. In the ideal scenario, $s^*_1 = \cdots = s^*_N$ in content (though in different $\mathcal{C}^{(i)}_{\text{sem}}$), and $U_{\text{group}} \to 1$. In practice, contractivity may not strictly hold; however, emphasizing

later turns via w_n ensures that approximate convergence yields a high U_{group}. If not, a low score flags failed consensus (useful, e.g., for human-AI collaboration diagnostics).

4 Conclusion and Future Work

We presented a categorical extension of the mutual understanding metric to multi-agent dialogue via a single symbolic channel. The per-turn group score is the strict geometric mean of pairwise information - semantic fidelities, times a geometric mean of pragmatic stabilities, and supports both agent-initiated (evolving) and external-GT (static) cases.

Computational complexity. The number of pairs is $K = \binom{N}{2} = O(N^2)$; per-turn cost is $O(K\, c_{\text{pair}})$, where c_{pair} is the cost of one $u_{ij}^{(n)}$. For large N, one can estimate the mean of $\log u_{ij}^{(n)}$ by mini-batching pairs, sparsify candidates via k-NN in embedding space, or maintain streaming/exponential-moving averages across turns.

Structural diagnostics. Beyond a single score, the matrix $[u_{ij}^{(n)}]$ forms a weighted graph: community detection (or spectral clustering) reveals aligned subgroups; agents with low average $\frac{1}{N-1}\sum_{j\neq i} \log u_{ij}^{(n)}$ flag persistent misalignment; trends in $\Delta \log u_{ij}^{(n)}$ highlight improving or degrading relations.

Hierarchical organization. For $N \gg 1$, cluster agents (e.g., by $d_{ij} = -\log u_{ij}^{(n)}$), compute within- and between-cluster scores, and aggregate them geometrically. This preserves the non-compensatory design while reducing computation and enabling modular deliberation.

5 Related Work

Our score builds on standard components - Jensen-Shannon divergence [2] and cosine geometry - yet differs in how they are combined and aggregated: rather than averaging toward a barycenter, we take a strict geometric product over all pairs, yielding a non-compensatory, per-turn group score in $[0, 1]$. Classical accounts of agreement as common knowledge [3] motivate the requirement that all pairs align, rather than permitting trade-offs across agents. Importantly, the equation here is *derived within the RC framework*: the single symbolic hub with interpretation/meaning functors and the appendix's Pinsker-type bound linking D_{JS} to deviation of $M \circ I$ from the identity motivate the multiplicative fidelity form [1, 6]. Although this derivation proceeds from RC rather than consensus dynamics or Bayesian pooling, the resulting score is nevertheless compatible with information- geometric treatments [2] and common-knowledge arguments [3].

A Experimental Design and Results

We report two empirical evaluations of the group-understanding metric introduced in Eqs. (1)-(2). The goal is not to benchmark particular models, but to illustrate how the metric behaves under controlled conversational dynamics and how the pairwise structure supports diagnosis.[2]

Across all runs, per-turn scores are computed exactly as in the main text: for each turn n, form the pairwise fidelities $u_{ij}^{(n)} = (1 - D_{\text{JS}}(P_i^{(n)} \parallel P_j^{(n)}))(1 - d_{\text{sem}}(s_i^{(n)}, s_j^{(n)}))$, take their strict geometric mean over pairs, and multiply by the geometric mean of the per-agent pragmatic retentions $\kappa_i^{(n)}$ (Eq. (1)). When a fixed ground truth (GT) is present, include the N agent-GT pairs and normalize over $K + N$ (Eq. (2)). Per-turn scores are aggregated with weights $\{w_n\}$, $\sum_n w_n = 1$ (Eq. (3)). All embeddings are ℓ_2-normalized; a small ε (e.g. 10^{-6}) guards the log domain when a fidelity is extremely small.

A.1 Protocols

We evaluate two configurations matching the use cases in the main text.

[2] Code and additional figures are provided in the supplementary material [9].

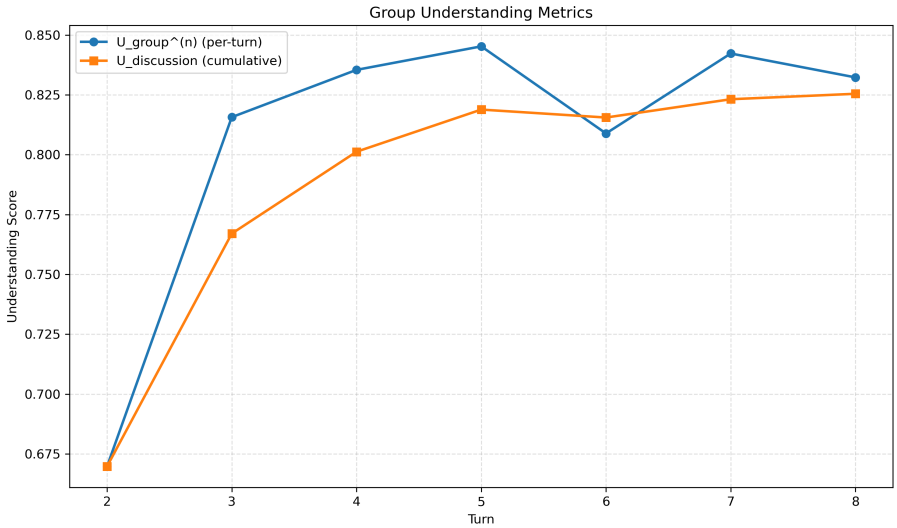


Figure 1: Case A (agent-generated query): per-turn group score $\mathcal{U}^{(n)}_{\text{group}}$ and cumulative discussion score U_{group} across turns. Early rises typically plateau; dips often coincide with large revisions penalized by κ.

Case A: Agent-generated query / evolving internal understanding.

1. A designated *speaker* produces a query and a private initial understanding; only the query is broadcast.
2. All agents independently write private understanding statements.
3. Agents ask clarifying questions to the speaker in turn; the speaker answers publicly via the shared symbolic channel. This conversation is available to all agents.
4. After each Q&A, all agents privately update their understanding.
5. Only the private understanding statements are used to compute $\mathcal{U}^{(n)}_{\text{group}}$.

Case B: External query with static GT.

1. A query and an external GT understanding statement are given. The external source is treated as another agent but is static ($\kappa_{\text{GT}} = 1$) and does not answer questions [3]
2. All agents independently write private understanding statements.
3. Clarifying questions are asked by randomly chosen agents and answered by randomly chosen agents.
4. After each Q&A, all agents privately update their understanding.
5. The per-turn score $\mathcal{U}^{(n)}_{\text{group,GT}}$ includes the static GT embedding (Eq. (2)).

In both cases we visualize: (i) the per-turn group score together with the running, weight-averaged discussion score; (ii) heatmaps of the pairwise $u^{(n)}_{ij}$ across turns; and (iii) the final-turn pairwise matrix for quick diagnosis.

A.2 Case A: Results and Analysis

Qualitative trends. With homogeneous agents (similar model family and decoding), $\mathcal{U}^{(n)}_{\text{group}}$ usually increases rapidly in the first few Q&A rounds and then plateaus. The cumulative score U_{group} rises more smoothly because of the weights $\{w_n\}$. Pairwise heatmaps show off-diagonals brightening as private understandings converge; pragmatic retentions $\kappa^{(n)}_i$ are lowest early (sharp revisions) and approach 1 as updates become incremental, naturally amplifying late-turn consensus.

[3] Updated configuration is possible where external dataset has predefined answers to predefined queries.

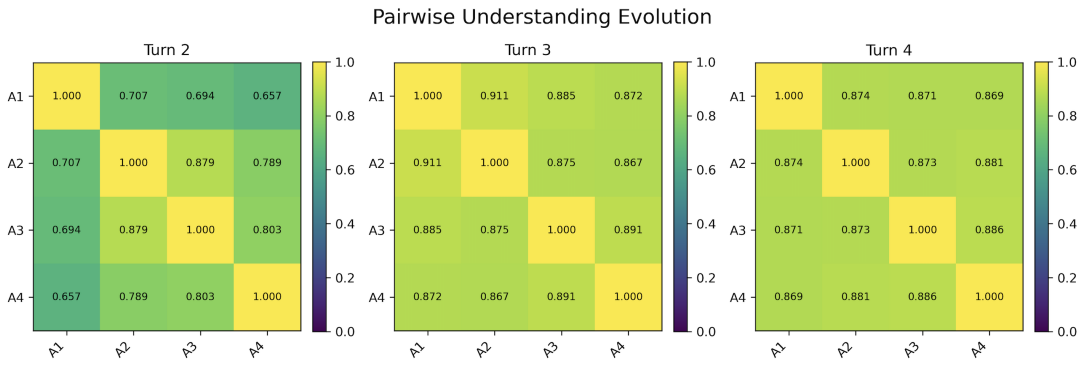


Figure 2: Case A: evolution of pairwise fidelities $u_{ij}^{(n)}$. Bright off-diagonal blocks indicate growing internal consensus; localized darker bands mark persistent bottlenecks.

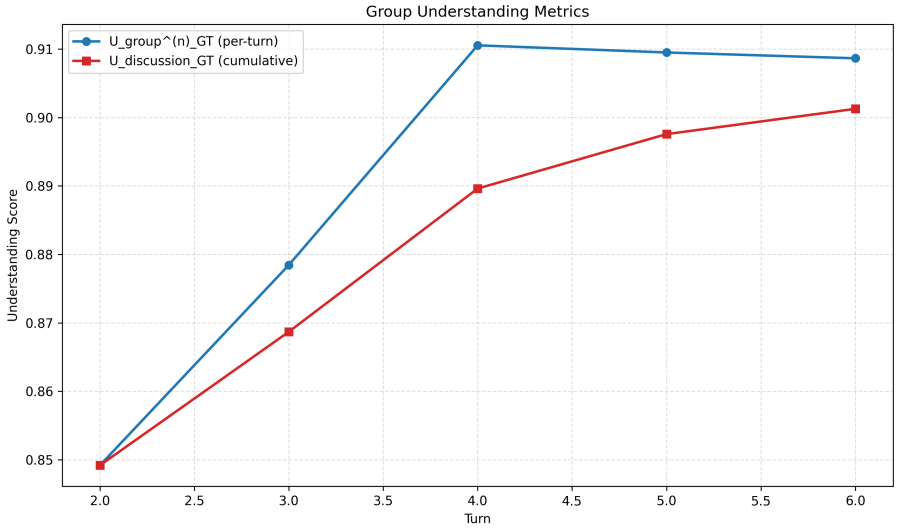


Figure 3: Case B (external GT): per-turn and cumulative scores with GT pairs included. A gap between inter-agent alignment and alignment-to-GT is common early and narrows as clarifications accrue.

A.3 Case B: Results and Analysis (External GT)

Coherence vs. correctness. A frequent pattern is rapid coalescence around a shared interpretation that is *not* initially well-aligned with GT: the agent-agent block is bright while the GT row/column is darker. Because the GT pairs enter multiplicatively, the per-turn $\mathcal{U}_{\text{group,GT}}^{(n)}$ remains below what a GT-agnostic score would report—precisely distinguishing *coherence* (internal agreement) from *correctness* (agreement with an external standard).

Convergence and saturation. When the GT is attainable, $\mathcal{U}_{\text{group,GT}}^{(n)}$ tends to rise and then plateau sooner than in Case A, as the GT anchor constrains plausible meanings. Residual variance often reflects asymmetric uptake of clarifications (a question that pulls one agent toward GT can nudge others slightly away).

A.4 Cross-cutting Observations

Heterogeneity and drift. With heterogeneous agents (different models, decoding regimes, or priors) we often see a mild late-turn decline after an initial rise. Some pairs stabilize while others drift, and the non-compensatory geometric mean drops due to a few hard pairs—usefully flagging true pockets of misalignment that an arithmetic average might hide.

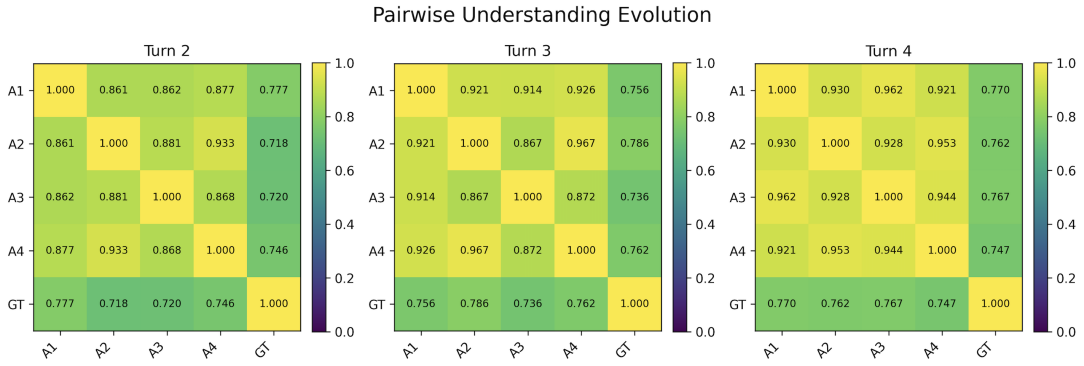


Figure 4: Case B: evolution of pairwise $u_{ij}^{(n)}$ including a GT row and column. A bright agent–agent block with a darker GT stripe indicates a coherent but GT-misaligned group; brightening of the GT stripe over time signals successful correction.

Role of the pragmatic factor. The retention term $\left(\prod_i \kappa_i^{(n)}\right)^{1/N}$ penalizes turns with large individual revisions and rewards stable consolidation. Smaller τ yields smoother, more stability-sensitive curves; larger τ makes the score more reactive.

Information vs. geometry. D_{JS} (belief agreement) and cosine (semantic proximity) often move together but not always. The multiplicative combination requires both to be high; either near zero pulls the pair score down, correctly signaling a substantive mismatch.

A.5 Practical Uses

- **Conversation steering and turn allocation.** Identify the weakest pair(s) (lowest $u_{ij}^{(n)}$) and prioritize questions from/to those agents for *active repair*.
- **Model/agent selection.** Use the final $[u_{ij}^{(T)}]$ to pick aligned subgroups or detect outliers that depress the geometric mean— enabling hierarchical deliberation.
- **GT vetting.** Compare the agent-agent block to the agent-GT columns/rows across turns to flag unreliable GTs or query/GT mismatches.
- **Stopping criteria.** When both $\mathcal{U}_{\mathrm{group}}^{(n)}$ and the dispersion of off-diagonal $u_{ij}^{(n)}$ stabilize, further Q&A is unlikely to materially improve consensus.
- **Ablation and fairness checks.** Swap agents, temperatures, or prompts and observe targeted effects on pairs; the non-compensatory design surfaces harmful asymmetries immediately.

A.6 Relation to Aumann's Agreement Theorem and Approximate Common Priors

Our findings do not contradict Aumann's Agreement Theorem [3]; rather, they illustrate what happens when a key premise - the *common prior assumption* - is violated. In the RC setting, agents typically inhabit *non-equivalent* (hence non-isomorphic) semantic categories $\mathcal{C}_{\mathrm{sem}}^{(i)}$, which we interpret as structurally different world models. This is a category-theoretic formalization of "no common prior." Under such heterogeneity, persistent interpretive gaps and rational disagreement are exactly what Aumann's framework allows once common priors are dropped.

Beyond the exact theorem, epistemic logic/game-theoretic work studies *approximate* variants: common knowledge can be approximated by common belief [10], and robustness of *"agreeing to disagree"* phenomena has been analyzed under relaxed premises on knowledge and priors. These results support a quantitative perspective: when agents' priors (or type spaces) are only ε-apart in an appropriate metric and there is (high-order) mutual awareness of posteriors, the posteriors themselves must be $\delta(\varepsilon)$-close for some modulus of continuity.[4]

[4] We use this only as a conceptual guide here; pinning down the exact modulus requires assumptions on the type space and update kernels.

A quantitative bridge via RC. Let the *structural dissimilarity* between A_i and A_j be measured by the induced operator distortion in our shared embedding:

$$d_{\text{struct}}(i,j) \;:\; \big\| (E \circ M_j \circ I_i) - (E \circ \mathrm{id}_{\mathcal{C}_{\text{sem}}^{(i)}}) \big\|_{\text{op}}.$$

By the Pinsker-type bound in §3 (Theoretical Foundations), there exists $C > 0$ such that

$$D_{\text{JS}}(P_i^{(n)} \,\|\, P_j^{(n)}) \;\geq\; C\, d_{\text{struct}}(i,j)^2,$$

on the relevant subspace (base-2 logs; constants absorb norm/log bases). Thus, structural mismatch enforces a *minimum* informational disagreement. In particular:

Proposition A.1 (Lower bound on pairwise fidelity from structural distortion)**.** *For each turn n and pair (i,j),*

$$u_{ij}^{(n)} = \big(1 - D_{\text{JS}}(P_i^{(n)} \,\|\, P_j^{(n)})\big) \cdot \frac{1 + \cos(Es_i^{(n)}, Es_j^{(n)})}{2} \;\leq\; \Big(1 - C\, d_{\text{struct}}(i,j)^2\Big) \cdot \frac{1 + \cos(Es_i^{(n)}, Es_j^{(n)})}{2}.$$

Consequently, for any discussion turn, $\mathcal{U}_{\text{group}}^{(n)}$ cannot exceed a ceiling determined by the hardest (largest d_{struct}) pairs—exactly the non-compensatory behavior we observe. This gives the metric predictive behavior: even before a conversation starts, estimates of $d_{\text{struct}}(i,j)$ bound the best-case alignment attainable by ideal Bayesian exchange.

Toward predictive two-sided laws. The bound above is one-sided (a floor on D_{JS}). With additional regularity, e.g., Lipschitz decoding M_i, bounded Jacobians of E, or an ε-equivalence between semantic categories - one expects an upper bound of the form $D_{\text{JS}} \leq g(\varepsilon)$, yielding $f(\varepsilon) \leq D_{\text{JS}} \leq g(\varepsilon)$ and hence δ-closeness of posteriors whenever priors/models are ε-close. Establishing such two-sided laws inside RC, and relating d_{struct} to ε in standard type-space metrics, is a natural direction for future work.

A.7 Limitations and Reproducibility Notes

The metric inherits inductive biases of the embedding model $E(\cdot)$ and of the hypothesis set used for $P_i^{(n)}$. Using a single embedding model across agents (we used OpenAI's *text-embedding-3-large* model) improves internal consistency but can attenuate stylistic differences in heterogeneous groups. Re-running with alternative embeddings and normalized prompts is advisable before drawing strong comparative conclusions. Results also vary with the underlying language models (we tested several OpenAI and local/open-source variants). Finally, because the geometric mean is intentionally brittle to outliers (strict non-compensation), scalar scores should be read together with the pairwise matrices.

References

[1] Miasnikov, S. (2025) Category-Theoretic Analysis of Inter-Agent Communication and Mutual Understanding Metric in Recursive Consciousness *Preprint* `http://dx.doi.org/10.13140/RG.2.2.15752.33280`

[2] Lin, J. (1991) Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory* `https://ieeexplore.ieee.org/document/61115`

[3] Aumann, R. J. (1976) Agreeing to Disagree. *The Annals of Statistics* `https://people.hec.edu/lovo/wp-content/uploads/sites/28/2019/03/Agreing-to-disagree.pdf`

[4] Miasnikov, S. (2025) Recursive Consciousness: Modeling Minds in Forgetful Systems. *Preprint* `http://dx.doi.org/10.13140/RG.2.2.26969.22884`

[5] Fritz, T. (2020) A synthetic approach to Markov kernels. `https://arxiv.org/pdf/1908.07021.pdf`

[6] Miasnikov, S. (2025) Appendix: Rigorous Categorical Derivation of Mutual Understanding Metric. *Preprint* `http://dx.doi.org/10.13140/RG.2.2.15752.33280`

[7] Mac Lane, S. (1971). Categories for the Working Mathematician. Springer-Verlag. `https://link.springer.com/book/10.1007/978-1-4757-4721-8`

[8] Blackburn, P., de Rijke, M., & Venema, Y. (2001). Modal Logic. Cambridge University Press. https://www.amazon.com/Cambridge-Tracts-Theoretical-Computer-Science/dp/0521527147

[9] Miasnikov, S. (2025). Recursive Consciousness Repository. *github* https://github.com/phatware/recursive-consciousness

[10] Monderer, D., & Samet, D. (1989). Approximating Common Knowledge with Common Beliefs. *Games and Economic Behavior* https://kylewoodward.com/blog-data/pdfs/references/monderer+samet-games-and-economic-behavior-1989A.pdf