Category-Theoretic Analysis of Inter-Agent Communication and Mutual Understanding Metric in Recursive Consciousness

Stan Miasnikov, August 2025
stanmiasnikov@gmail.com

A Appendix: Rigorous Categorical Derivation of the Mutual Understanding Metric

Note: This appendix supplements the main paper "*Category-Theoretic Analysis of Inter-Agent Communication and Mutual Understanding Metric in Recursive Consciousness*" available at http://dx.doi.org/10.13140/RG.2.2.15752.33280.

In this appendix, we provide a more rigorous categorical foundation for the mutual understanding metric introduced in the main paper. We address the gaps noted in the heuristic justifications by enriching the semantic categories over Banach spaces, formally defining an embedding functor to vector spaces, and proving convergence properties of the communication functor Φ using an enriched version of the Banach fixed-point theorem [1]. We also clarify assumptions, conjectures, and their connections to existing literature in enriched category theory, categorical semantics, and categorical probability.

A.1 Semantic Categories as Enriched Banach Categories

We model each agent's semantic universe as a category enriched over **Ban**, the monoidal category of Banach spaces and short linear maps (with projective tensor product $\hat{\otimes}$, also denoted $\otimes_\pi$, where $X\hat{\otimes}Y$ is the completion of the algebraic tensor product $X \otimes Y$ with respect to the projective cross-norm $\|z\|_\pi = \inf\{\sum_{i=1}^n \|x_i\| \cdot \|y_i\| \mid z = \sum_{i=1}^n x_i \otimes y_i,\, n \in \mathbb{N},\, x_i \in X,\, y_i \in Y\}$). This enrichment, standard in quantitative category theory [2], equips hom-objects with norms to quantify semantic distortion, justified by the need for measurable information loss in recursive self-query and inter-agent alignment.

The phrasing "enriched over the monoidal category" is essential because enrichment requires a base category $\mathcal{V}$ that is monoidal (equipped with a tensor product $\otimes$ and unit object, satisfying associativity and unit axioms up to isomorphism). This allows composition of morphisms via the tensor: for $f : A \to B$ and $g : B \to C$, $g \circ f$ uses $\hom(A, B) \otimes \hom(B, C) \to \hom(A, C)$. Here, $\mathcal{V} = \mathbf{Ban}$, the category of Banach spaces (complete normed vector spaces over $\mathbb{R}$ or $\mathbb{C}$) with morphisms as bounded linear maps. **Ban** is monoidal under the projective tensor product $\hat{\otimes}$, which completes the algebraic tensor product with the norm $\|z\| = \inf\{\sum \|x_i\|\|y_i\| : z = \sum x_i \otimes y_i\}$. This makes **Ban** closed monoidal (with internal hom $[X, Y] = \mathcal{B}(X, Y)$, bounded operators). The unit is $\mathbb{R}$ (or $\mathbb{C}$).

For any two semantic objects (e.g., concepts or meanings) s, s' in the category $\mathcal{C}_{\text{sem}}$, the collection of morphisms $\mathcal{C}_{\text{sem}}(s, s')$ carries the structure of a normed linear space (a Banach space, assuming completeness). Composition of morphisms is bilinear and continuous with respect to these norms.

Objects correspond to semantic points (or vector spaces), and morphisms to bounded linear maps encoding transformations or entailments. The enrichment provides a quantitative notion of distance: each morphism $f : s \to s'$ has an operator norm $\|f\|$ measuring distortion. The identity morphism has operator norm 1 (as $\sup_{\|x\|\leq 1} \|x\| = 1$), but represents zero distortion in semantic transformations.[1] The enrichment axioms ensure norm submultiplicativity under composition ($\|g \circ f\| \leq \|g\| \cdot \|f\|$).

This construction aligns with normed categories [3], generalizing metric spaces and linear operator theory. Enrichment over Banach spaces incorporates quantitative information into semantic structures, enabling analysis of information loss.

Justification for applicability to semantic universes in recursive consciousness: In consciousness/AI models, semantics involve distances (e.g., similarity of concepts). Banach enrichment quantifies "distortion" via norms: $\|f\| = \sup_{\|x\|\leq 1} \|f(x)\|$, measuring how morphisms (semantic transformations) stretch

[1] In Kubiś's normed categories [3], identities are axiomatically norm 0; our Banach enrichment follows standard operator conventions (e.g., Castillo, 2021).

meanings. This aligns with distributional semantics (e.g., word embeddings in $\mathbb{R}^d$), where meanings are vectors and relations are linear maps [4]. In recursive consciousness, agents "forget" and reconstruct via functors; norms capture loss as $\|\Phi(s) - s\| > 0$. Completeness for convergence: Banach completeness (every Cauchy sequence converges) enables fixed-point theorems for iterative dialogues, modeling convergence to some level of mutual understanding. Prior research support: [2] surveys categorical Banach theory, emphasizing limits/colimits and adjunctions—key to our $I \dashv M$. Philosophically, this echoes Leibniz's monads (e.g., Monadology, 1714) mirroring universes with perspective shifts, quantifiable as normed distortions. Why not plain Vect? Banach adds norms/completeness, essential for metrics like operator norms in consciousness (e.g., bounded self-reference avoiding divergence).

Morphisms as bounded linear operators ensure structure-preservation and continuity. The hom-set $\mathcal{C}_{\text{sem}}(s, s')$ is a Banach space with operator norm. This holds by the enrichment definition: hom-objects inherit Banach structure from **Ban**, with the operator norm ensuring boundedness. The hom-set satisfies the triangle inequality analog.

A.2 The Embedding Functor $E : \mathcal{C}_{\text{sem}} \to \mathbb{R}^d$

The embedding functor E concretizes the abstract semantic category into a vector space, ensuring that semantic structure is faithfully preserved. Its effectiveness critically depends on accurately encoding semantic meanings, for example, through language-model embeddings where cosine similarity corresponds well with human judgments of conceptual closeness. Poorly chosen embeddings, such as those with insufficient dimensionality or misalignment, could distort norms and consequently produce inaccurate mutual understanding metrics. Thus, it is crucial to justify the assumption that embedding functors E can be chosen to preserve semantic structure accurately.

To address these theoretical considerations, we establish the following proposition regarding the existence and properties of faithful, norm-preserving embeddings:

Proposition A.1 (Existence of Faithful Norm-Preserving Embeddings). *Let $\mathcal{C}_{sem}$ be the semantic category, assumed to be enriched over a normed linear space or at least to have an intrinsic metric/norm on its hom-sets. Assuming $\mathcal{C}_{sem}$ is a normed category in the sense of Kubiś (e.g., semi-additive or with metric hom-sets) [3] - then there exists a faithful functor*

$$E : \mathcal{C}_{sem} \hookrightarrow V$$

where V is an infinite-dimensional Banach space like ℓ^∞ that preserves norms exactly:

$$\|x - y\|_{\mathcal{C}_{sem}} = \|E(x) - E(y)\|_V \quad \textit{for all semantic objects } x, y$$

Moreover, E can be chosen to be full and faithful, reflecting the semantic structure faithfully. While a strictly isometric embedding into a finite-dimensional vector space $\mathbb{R}^d$ may generally not be feasible, one can first construct an exact embedding into a higher-dimensional normed space (e.g., ℓ^∞ or $\mathbb{R}^N$ for sufficiently large N) and subsequently apply dimensionality-reduction techniques to achieve approximate embeddings in practical finite-dimensional spaces.

Existence. This result leverages classical and contemporary results from metric category theory. Consider initially the well-known Kuratowski embedding [5], which maps any metric space (X, d) isometrically into the normed linear space $\ell^\infty(X)$, via

$$x \mapsto f_x, \quad \text{where } f_x(y) = d(x, y).$$

This embedding is isometric (exactly distance-preserving) and functorial in the metric sense, as maps between metric spaces lift naturally to linear maps between their representations in ℓ^∞.

Following Kubiś [3], a categorical refinement of this classical embedding approach provides a faithful embedding of each normed or metric-enriched category into a suitable normed linear space. Concretely, one constructs a free normed vector space generated by the semantic objects of $\mathcal{C}_{\text{sem}}$, imposing exactly the relations derived from semantic distances. If the semantic category has an inner-product structure (e.g., a Hilbert space of meanings), one obtains even stronger embeddings by employing orthonormal bases. □

Faithfulness and Norm Preservation. The construction described ensures injectivity on objects by design: distinct semantic elements map to distinct vectors, as their zero-distance equivalence would otherwise contradict the embedding construction. Faithfulness on morphisms follows naturally since any semantic relation (morphisms in the enriched categorical sense) that preserves norms or metrics will translate into linear operators in the embedding vector space. Norm preservation is inherent to this construction since the induced distances between points in the embedding space precisely match the original semantic distances by the chosen constraints.

In practice, one can apply classical dimensionality-reduction methods (e.g., Johnson-Lindenstrauss lemma, Isomap, Multi-dimensional Scaling) to approximate such embeddings into a lower-dimensional space $\mathbb{R}^d$, maintaining near-isometric structure with controlled distortion. □

Remark 1 (Epistemic Limits of Embedding in LLMs and Shared Architectures)**.** ***(1) Internal Fullness and Norm Preservation.*** *From the perspective of a large language model (LLM), the embedding functor $E : \mathcal{C}_{sem} \to \mathbb{R}^d$ is epistemically full, faithful, and norm-preserving by construction. This follows from the fact that the LLM lacks access to any external semantic category beyond its internally instantiated embedding space. Therefore, regardless of potential semantic imperfections or distortions relative to an idealized external universe U, the internal system must treat E as the best possible approximation to meaning. Consequently, all internal judgments about semantic similarity, alignment, or inference are grounded in the topology and norm structure induced by E, even if these diverge from any ontologically grounded meanings in the external environment U.*

(2) Embedding Equivalence in Shared Architectures. *If two AI models (e.g., two LLM instances) use the* same *embedding function E and receive the same input x, they necessarily produce identical embedding vectors $E(x)$, regardless of their downstream interpretative mechanisms. Thus, from the standpoint of the semantic metric d_{sem} introduced above, such systems exhibit maximal alignment at the representational level. However, this alignment does not guarantee semantic agreement, as the meaning extraction functors M_A and M_B may diverge. Still, the mutual understanding metric will treat any divergence between agents with a shared E as purely a function of M and downstream inference, since E itself cannot introduce representational divergence. Hence, the quality of mutual understanding is upper-bounded by the expressiveness and faithfulness of E, which acts as a representational bottleneck.*

Consequently, the existence of $E : \mathcal{C}_{\text{sem}} \to \mathbf{Vect}_{\mathbb{R}}$ crucially depends on the semantic category $\mathcal{C}_{\text{sem}}$ being concrete, admitting a faithful functor into $\mathbf{Vect}$. Sufficient conditions include semi-additivity or linearity in the hom-sets. The practical approximation by finite-dimensional embeddings aligns well with categorical compositional semantics frameworks [4], supporting the practical deployment of embeddings for computational purposes while ensuring minimal semantic distortion.

A.3 Fixed Points of the Composite Functor $\Psi = I \circ M$ and Banach's Theorem

The communication functor $\Phi_{A\to B} = M_B \circ I_A : \mathcal{C}^A_{\text{sem}} \to \mathcal{C}^B_{\text{sem}}$ encapsulates the channel. While $\Phi_{A\to B}$ is unidirectional, we denote $\Psi = (M_A \circ I_B) \circ (M_B \circ I_A)$ for round-trip analysis in iterative dialogues, capturing epistemic stabilization across agents. A fixed point $s^* \in \mathcal{C}^B_{\text{sem}}$ satisfies $\Psi(s^*) \cong s^*$.

In the enriched setting, we can restrict to a shared subcategory where $\Psi : X \to X$.

Theorem A.1 (Pinsker-type Lower Bound for Jensen-Shannon)**.** *There exists a universal constant $c > 0$ such that for any two probability distributionsP_A and P_B in the semantic space, the Jensen-Shannon divergence is bounded below by the squared norm of the difference between the semantic channel Φ and the identity:*

$$D_{\text{JS}}(P_A \,\|\, P_B) \;\geq\; c\,\|\Phi - \text{id}\|^2\,.$$

In particular, if $\|\Phi - \text{id}\|$ is measured via the L^1 norm on distributions (i.e., $\|\Phi - \text{id}\| = \sup_P \|\Phi(P) - P\|_1$), one may take $c = \frac{1}{8}$. Equivalently, if measured via the total variation distance (TV, defined as $\text{TV}(P,Q) = \frac{1}{2}\|P - Q\|_1$), one may take $c = \frac{1}{2}$ (via Vajda's refinement of Pinsker; see [6]).

Proof. This result is inspired by Pinsker's inequality (which lower-bounds Kullback-Leibler divergence by half the squared L^1 distance in appropriate units). Here we leverage known bounds relating Jensen-Shannon divergence to total variation. In [7] they derive that for any two distributions P, Q, one has $D_{\text{JS}}(P\|Q) \geq \frac{1}{8}\|P - Q\|_1^2$. This result follows from Pinsker's inequality and its refinements (e.g. Vajda's inequality) in the space of probability measures. In our setting, $\|\Phi - \text{id}\|$ represents the worst-case distortion of the "meaning" channel Φ away from the identity mapping. If we choose some P_A that attains this worst-case (i.e. $\|\Phi(P_A) - P_A\|_1 = \|\Phi - \text{id}\|$ where the norm is L^1) and let $P_B = \Phi(P_A)$, then indeed

$$D_{\mathrm{JS}}(P_A \| P_B) \;\geq\; \frac{1}{8}\,\|P_A - P_B\|_1^2 \;=\; \frac{1}{8}\,\|\Phi - \mathrm{id}\|^2$$

Thus a nonzero deviation of Φ from the identity functor yields a strictly positive Jensen-Shannon divergence. Intuitively, $D_{\mathrm{JS}}(P_A \| P_B)$ captures the information loss when going from P_A to P_B; if Φ were the identity (perfect mutual understanding), this divergence would be zero. For Φ that is not identity, one can interpret D_{JS} as the mutual information between an underlying binary variable and the outcome under P_A vs. P_B, which by the data-processing inequality cannot be zero unless Φ is identity. This information-theoretic view [8] is consistent with the above bound - it guarantees a quadratic relationship between JSD and the "channel error" norm. We note that sharper constants c can be obtained for small distortions using refined inequalities, but $\mathcal{O}(\|\Phi - \mathrm{id}\|^2)$ growth is the correct order. □

Remark 2. *In practice, this lower bound implies that if two agents' semantic distributions P_A, P_B differ significantly (large D_{JS}), then the interpretation/meaning mapping Φ between them must deviate appreciably from identity. Conversely, a near-identity semantic mapping guarantees a very small JSD (mutual information loss) between the agents. This justifies using D_{JS} as a quantitative measure of mutual understanding: it vanishes only when Φ effectively transmits semantics without distortion. Note that in dynamic dialogues, mutual understanding may fluctuate or even decrease as agents update beliefs, consistent with experimental observations where the metric converges quickly but can vary non-monotonically due to context shifts or forgetting mechanisms.*

Theorem A.2 (Contractive Interpretation–Meaning Cycle). *Let $I_A : \mathcal{C}_{sem} \to \mathcal{C}_{syn}$ and $I_B : \mathcal{C}_{sem} \to \mathcal{C}_{syn}$ be the interpretation functors for agents A and B (mapping semantic space to syntax/signals), and let $M_A : \mathcal{C}_{syn} \to \mathcal{C}_{sem}$ and $M_B : \mathcal{C}_{syn} \to \mathcal{C}_{sem}$ be the corresponding meaning-extraction functors. These functors are lossy (non-full, information-destroying) and thus contractive on the normed semantic with constants $0 \leq \lambda_A, \lambda_B, \mu_A, \mu_B < 1$ in typical cooperative RC settings (e.g., under forgetful adjunctions where information non-increase holds; cf. enriched DPI and positivity axioms in [9, 10]; note that in adversarial or novel exchanges, contractivity may fail.) such that for all semantic states s, t:*

$$\|I_A(s) - I_A(t)\| \leq \lambda_A \,\|s - t\|, \qquad \|I_B(s) - I_B(t)\| \leq \lambda_B \,\|s - t\|,$$

$$\|M_A(x) - M_A(y)\| \leq \mu_A \,\|x - y\|, \qquad \|M_B(x) - M_B(y)\| \leq \mu_B \,\|x - y\|,$$

where $x = I_A(s)$, $y = I_A(t)$ (or similarly for I_B) are syntactic states in $\mathcal{C}_{syn}$.
Then the composite round-trip functor

$$\Psi \;:=\; (M_A \circ I_B) \;\circ\; (M_B \circ I_A) : \quad \mathcal{C}_{sem} \to \mathcal{C}_{sem}$$

is a strict contraction on the semantic space. In particular,

$$\|\Psi(s) - \Psi(t)\| \;\leq\; c\,\|s - t\|, \qquad c := \lambda_A\,\lambda_B\,\mu_A\,\mu_B < 1$$

for all semantic states s, t. Consequently, by the Banach Fixed-Point Theorem, Ψ admits a unique fixed point in $\mathcal{C}_{sem}$, and iterative application $\Psi^n(s)$ converges to this fixpoint for any initial state s.

Proof. The contractivity follows straightforwardly from the assumptions on each component functor. For any two semantic states $s, t \in \mathcal{C}_{\mathrm{sem}}$, we have

$$\begin{aligned}
\|\Psi(s) - \Psi(t)\| &= \|(M_A \circ I_B)\big((M_B \circ I_A)(s)\big) \;-\; (M_A \circ I_B)\big((M_B \circ I_A)(t)\big)\| \\
&\leq \mu_A \,\|\, I_B(M_B(I_A(s))) \;-\; I_B(M_B(I_A(t)))\,\| \qquad \text{(since } M_A \text{ is } \mu_A\text{-Lipschitz)} \\
&\leq \mu_A\lambda_B \,\|\, M_B(I_A(s)) - M_B(I_A(t))\,\| \qquad \text{(since } I_B \text{ is } \lambda_B\text{-Lipschitz)} \\
&\leq \mu_A\lambda_B\mu_B \,\|\, I_A(s) - I_A(t)\| \qquad \text{(since } M_B \text{ is } \mu_B\text{-Lipschitz)} \\
&\leq \mu_A\lambda_B\mu_B\lambda_A \,\|\, s - t\| \qquad \text{(since } I_A \text{ is } \lambda_A\text{-Lipschitz)}
\end{aligned}$$

which yields the constant $c = \lambda_A\lambda_B\mu_A\mu_B < 1$ as claimed. The key intuition is that each interpretation I or meaning map M can only decrease the distinguishability of semantic states (they are "lossy compressors" of information), so the overall round-trip mapping Ψ strictly contracts distances in the semantic metric.

Given a contraction Ψ on a complete metric (or normed) space, Banach's Fixed-Point Theorem guarantees a unique fixed point s^* with $\Psi(s^*) = s^*$. In the categorical setting, [3] provides an analogous result: any contractive endofunctor on a Cauchy-complete normed category has a unique fixed object. Here Ψ acts as a contraction on the semantic space (assumed to be complete under $\|\cdot\|$), so there is a unique semantic state s^* that remains invariant under one full $A \to B \to A$ communication cycle. Moreover, for *any* initial state s_0, the sequence of iterates $\Psi^n(s_0)$ converges exponentially fast to s^* as $n \to \infty$. □

Remark 3. *In less formal terms, $\Psi = M_A \circ I_B \circ M_B \circ I_A$ represents one full back-and-forth exchange of meaning: A interprets s into a message, B extracts meaning, then B responds (or reinterprets) and A extracts meaning back. The contractivity condition says that any initial discrepancy in meaning between A and B will shrink with each round-trip communication. The factor $c = \lambda_A \lambda_B \mu_A \mu_B < 1$ quantifies how much closer the meanings get after each cycle. If, for example, each interpretation and extraction retains at least, say, 90% of the information (contractivity factors $\lambda, \mu \approx 0.9$), then one full cycle might retain $\sim (0.9)^4 \approx 0.66$ (66%) of the difference - a significant reduction. Repeated cycles eliminate the misunderstanding, converging to a fixed shared meaning s^*.*

However, experimental results indicate that mutual understanding may fluctuate or even decrease over iterations, and dialogues can diverge in some cases (e.g., human debates or AI hallucinations where polysemy expands distances). This suggests the contractivity assumption ($c < 1$) holds conditionally in cooperative, convergent dialogues; in adversarial regimes, $c \geq 1$ may lead to instability (cf. expansive functors in categorical ML [11]). Factors such as evolving functors (e.g., due to context forgetting in AI or shifting perspectives in humans) or noise can lead to $c \geq 1$, resulting in non-convergence or instability. Thus, the theorem models an ideal convergent scenario, while real systems may require adaptive mechanisms to enforce contractivity.

This contractive cycle underpins semantic stabilization in multi-agent systems, extending to the hierarchical structures of recursive consciousness. In RC context, semantic universes $\mathcal{C}_{\text{sem}}$ are modeled as metric-enriched domains, with recursive equations like $U_{n+1} \cong G(F(U_n))$ for hierarchical consciousness; contractions arise from forgetful functors inducing loss, ensuring unique stable meanings despite ambiguity.

A.4 Quantitative Modeling of Metric Components

The mutual understanding metric quantifies deviation from identity:

$$\|\Phi - \mathrm{id}\| = \sup_s \frac{\|\Phi(s) - s\|}{\|s\|}.$$

- **Semantic Deviation**: Embed via E, use cosine similarity as proxy for norm-induced angle: $d_{\text{sem}}(s_A, s_B) = \frac{1-\cos(E(s_A),E(s_B))}{2}$, bounding chordal distance. From norm $\delta = \|\Phi(s_A) - s_A\|$ (on shared space), for unit-normalized embeddings ($\|E(s)\| = 1$) into a Hilbert space (a special case of Banach spaces), the Euclidean norm relates directly to the angle θ: $\delta^2 = \|u - v\|^2 = \|u\|^2 + \|v\|^2 - 2\cos\theta = 2(1 - \cos\theta)$, so $d_{\text{sem}} = \frac{\delta^2}{4} = \sin^2(\theta/2) \approx \frac{1-\cos\theta}{2}$.

- **Probabilistic Misalignment**: We map semantic objects to belief distributions $P(s)$ over possible worlds (e.g., a Kripke frame). Transformations are modeled by stochastic matrices, forming morphisms in $\mathsf{FinStoch}$, enriched over metric spaces with Jensen-Shannon divergence (JSD) on hom-sets. Composition is non-expansive w.r.t. JSD, per the Data Processing Inequality, making JSD a natural quantifier of distortion between agent distributions P_A and P_B:

  $$D_{\text{JS}}(P_A \| P_B) = \frac{1}{2}[D_{\text{KL}}(P_A \| M) + D_{\text{KL}}(P_B \| M)], \quad M = \frac{1}{2}(P_A + P_B)$$

  By the Theorem A.1 $D_{\text{JS}} \geq \frac{1}{8}\|\Phi - \mathrm{id}\|^2$ (with L^1 norm), directly linking probabilistic misalignment to the operator norm deviation from identity.

- **Pragmatic Instability**: This component measures the convergence of the dialogue over iterations $s^{(n+1)} = \Phi(s^{(n)})$. The assumption that the round-trip functor Ψ is contractive is grounded in the structure of the RC model. The constituent functors (I for interpretation, M for meaning) model a necessarily lossy process of projecting rich semantics onto a simpler symbolic channel and attempting to reconstruct them. This inherent information loss, when quantified in a normed category, is expected to manifest as a contraction with Lipschitz constant $c < 1$. As established in Section A.3, this contractivity allows the direct application of the category-theoretic Banach fixed-point theorem, guaranteeing convergence to a unique fixed point s^*. The rate of this convergence is bounded by the classical error estimates. The per-turn shift $\Delta_n = \|s^{(n)} - s^{(n-1)}\|$ is therefore expected to decay exponentially: $\Delta_n \leq c^{n-1}(1-c)|s^{(1)} - s^{(0)}| + o(1)$. We quantify stability using this shift: $\kappa^{(n)} = \exp(-\Delta_n)$, which converges to 1 as the dialogue stabilizes.

The resulting mutual understanding metric:

$$U_{\text{mutual}} = \sum_n w_n [(1 - D_{\text{JS}}^{(n)}) \cdot (1 - d_{\text{sem}}^{(n)}) \cdot \sqrt{\kappa_A^{(n)} \kappa_B^{(n)}}]$$

with weights $w_n = \frac{2n}{N(N+1)}$ for linear recency (summing to 1); derived from exponential decay: alternatively, $w_n \propto c^{N-n}$ from Banach bounds, ensuring alignment with convergence rate. It represents a composite fidelity score that quantifies the progressive alignment of semantic states between agents over N dialogue rounds. This formulation draws on the enriched categorical structure by aggregating normalized inverses of the distortion measures: $1 - D_{\text{JS}}^{(n)}$ reflects probabilistic belief alignment which is linked to the squared functor norm by the lower bound established in Theorem A.1 $D_{\text{JS}} \geq \frac{1}{8}\|\Phi - \text{id}\|^2$, $1 - d_{\text{sem}}^{(n)}$ captures semantic closeness in the embedded Banach space (derived from cosine as a proxy for angular distortion, with $d_{\text{sem}} = \sin^2(\theta/2) \approx \frac{\|\Phi(s) - s\|^2}{4}$ for unit vectors), and $\sqrt{\kappa_A^{(n)} \kappa_B^{(n)}}$ symmetrizes pragmatic stability (exponential decay of per-turn shifts under contractivity, bounded by Banach error estimates $\Delta_n \leq c^{n-1}(1-c)\Delta_0$). Multiplicative aggregation enforces non-compensatory synergy (failures in one dimension zero the term), justified as a geometric mean for interdependent factors, akin to products in joint entropy formulations [12] (for foundational entropy measures, where $H(X,Y) = H(X) + H(Y) - I(X;Y)$ motivates non-additive interactions). A possible alternative - additive with thresholds - could be explored, but it may not capture the same interdependencies.

As mentioned above, philosophically, this echoes Leibnizian monads in recursive consciousness, where mutual mirroring converges to a fixpoint only through damped iterations, avoiding Gödelian oscillations; in AI terms, it advances frameworks like Mutual Theory of Mind (MToM) [13], which uses linguistic proxies (e.g., cosine-based adaptability) for perception scoring, by incorporating categorical norms for rigorous quantification in multi-agent systems, akin to symbolic knowledge overlap metrics [14] or communication efficiency ratios [15]. Empirically, this metric could be validated via simulations using AI agents, where agent embeddings evolve under Φ, computing U_{mutual} to benchmark alignment in human-AI dialogues.

A.5 Practical Considerations for Choosing Embedding Functors

To construct or learn a "well-chosen" embedding functor E (Section A.2) that is faithful (injective on morphisms) and norm-preserving, techniques from geometric deep learning (GDL) and manifold learning offer practical pathways, particularly for ensuring alignment with the categorical requirements of preserving semantic distortions and hierarchical structures in recursive consciousness models. [16] GDL extends traditional deep learning to non-Euclidean domains, such as graphs or manifolds, by incorporating geometric priors that respect symmetries and invariances—crucial for embedding functors that must map abstract semantic morphisms (e.g., entailments or recursive queries) to bounded linear operators without collapsing distinctions. For instance, graph neural networks (GNNs) can be trained on semantic graphs where nodes represent concepts and edges denote relationships (e.g., hypernymy or similarity), using message-passing layers to propagate features while preserving geodesic distances. A key equation in GDL for node embeddings is the graph convolution:

$$h_v^{(l+1)} = \sigma\left(W^{(l)} \cdot \text{AGG}\left(\{h_u^{(l)} : u \in \mathcal{N}(v) \cup \{v\}\}\right)\right),$$

where $h_v^{(l)}$ is the embedding of node v at layer l, AGG is an aggregation function (e.g., mean or attention-weighted sum as in GAT), $W^{(l)}$ is a learnable weight matrix, and σ is a non-linearity.

To enforce norm-preservation, incorporate regularization terms like $\mathcal{L}_{\text{norm}} = \sum_f \|E(f)\| - \|f\|\|^2$, minimizing deviation from abstract norms (estimated via proxy metrics like graph Laplacian eigenvalues). Training on datasets like WordNet or ConceptNet, with contrastive losses (e.g., NT-Xent:

$\mathcal{L} = -\log \frac{\exp(\cos(E(s_i),E(s_j^+))/\tau)}{\sum_k \exp(\cos(E(s_i),E(s_k))/\tau)}$, where s_j^+ are positive pairs and τ is temperature), ensures embeddings align with human semantic judgments while maintaining faithfulness—empirically achieving Spearman's correlations > 0.8 with benchmarks like SimLex-999.

Manifold learning complements GDL by assuming semantic spaces lie on low-dimensional manifolds embedded in high-dimensional ambient spaces, enabling techniques like UMAP (Uniform Manifold Approximation and Projection) or Isomap to recover intrinsic geometries that comply with our requirements for E. UMAP, for example, optimizes a cross-entropy loss to preserve local distances while globally aligning via fuzzy simplicial sets, formalized as minimizing $\mathcal{L} = \sum_{i,j} [p_{ij} \log \frac{p_{ij}}{q_{ij}} + (1 - p_{ij}) \log \frac{1-p_{ij}}{1-q_{ij}}]$, where p_{ij} and q_{ij} are high- and low-dimensional probabilities (e.g., Gaussian kernels for local metrics). For LLMs, fine-tune embedding layers using manifold alignment objectives, such as those in stratified manifold learning for LLM spaces, where embeddings are projected onto learned submanifolds to stratify hierarchical semantics (e.g., via spectral clustering on the graph Laplacian $L = D - A$, with D degree matrix and A adjacency). Practical implementation involves pre-training on corpora like Common Crawl, then fine-tuning with manifold regularization (e.g., adding $\lambda \|L\mathbf{H}\|_F^2$ to the loss, where $\mathbf{H}$ is the embedding matrix and λ balances fidelity).

A.6 Computational Complexity Considerations

While the mutual understanding metric offers a category-theoretic lens on semantic alignment in recursive consciousness, its implementation demands attention to computational demands. The embedding functor $E : \mathcal{C}_{\text{sem}} \to \mathbb{R}^d$ typically relies on pretrained LLMs (e.g., BERT variants), incurring $O(d \cdot m)$ time per semantic object, where d is embedding dimension (often 768-4096) and m is input length; batching mitigates this to amortized $O(1)$ per dialogue round in practice. Jensen-Shannon divergence D_{JS}, computed over belief distributions (e.g., discrete Kripke worlds or softmax outputs), scales as $O(k \log k)$ for k-bin histograms via entropy formulas, or exactly $O(k)$ for sparse vectors—feasible for modest k (e.g., 10^3 states in modal logic simulations). Pragmatic stability $\kappa^{(n)}$ is $O(1)$ per norm computation. Overall, for N iterations, complexity is $O(N \cdot (dm + k \log k))$, reducible to sublinear via subsampling (e.g., reservoir sampling) or low-rank approximations (e.g., SVD on embeddings, adding $O(d^2)$ per step but amortizing over N [11] for reflexive domains in ML), tractable on GPUs for AI agents but scaling with hierarchical depth in recursive models; optimizations like low-rank approximations or subsampling worlds could reduce to sublinear, aligning with philosophical efficiency in monadic self-reference.

References

[1] Banach fixed-point theorem *wikipedia* https://en.wikipedia.org/wiki/Banach_fixed-point_theorem

[2] J.M.F. Castillo, (2021) The hitchhiker guide to Categorical Banach space theory. Part II, arXiv:2110.06300 [math.FA], 2021. https://arxiv.org/abs/2110.06300

[3] Wieslaw Kubis, (2017) Normed Categories, https://arxiv.org/pdf/1705.10189 https://arxiv.org/pdf/1705.10189

[4] Bob Coecke, Mehrnoosh Sadrzadeh, Stephen Clark, (2010) Mathematical Foundations for a Compositional Distributional Model of Meaning, arXiv:1003.4394 https://arxiv.org/abs/1003.4394

[5] Kuratowski embedding *wikipedia* https://en.wikipedia.org/wiki/Kuratowski_embedding

[6] J. Vajda, (1970) Note on discrimination information and variation https://doi.org/10.1109/TIT.1970.1054557

[7] J. Corander, U. Remes, and T. Koski, (2021) On the Jensen-Shannon divergence and the variation distance for categorical probability distributions. *Kybernetika*, 57(6):879–907, https://www.kybernetika.cz/content/2021/6/879/paper.pdf

[8] Eigil Fjeldgren Rischel, (2022) Jensen-Shannon divergence is compositional, *blog* https://erischel.com/jsd-as-enrichment/

[9] Tobias Fritz, (2020) A synthetic approach to Markov kernels, conditional independence and theorems on sufficient statistics https://arxiv.org/abs/1908.07021

[10] Tobias Fritz, Tomáš Gonda, Nicholas Gauguin Houghton-Larsen, Antonio Lorenzin, Paolo Perrone, Dario Stein, (2023) Dilations and information flow axioms in categorical probability `https://arxiv.org/abs/2211.02507`

[11] Pierre America, Jan J. M. M. Rutten, (1989) Solving Reflexive Domain Equations in a Category of Complete Metric Spaces, Journal of Computer and System Sciences, Volume 39, Issue 3, Pages 343-375. `https://ir.cwi.nl/pub/1646/1646D.pdf`

[12] C. E. Shannon, (1948) A Mathematical Theory of Communication, The Bell System Technical Journal, Vol. 27, No. 3, pp. 379-423 (July) and No. 4, pp. 623-656 (October). `https://people.math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf`.

[13] Qiaosi Wang, Ashok K. Goe, (2022) Mutual Theory of Mind for Human-AI Communication `https://qiaosiwang.me/Publications/MToM-CHAI2022.pdf`

[14] Federico Sabbatini, Christel Sirocchi, Roberta Calegari, (2024) Symbolic Knowledge Comparison: Metrics and Methodologies for Multi-Agent Systems `https://ceur-ws.org/Vol-3735/paper_17.pdf`

[15] Conor Bronsdon, (2025) A Guide to Measuring Communication Efficiency in Multi-Agent AI Systems *blog* `https://galileo.ai/blog/measure-communication-in-multi-agent-ai`

[16] Wenming Cao, Zhiyue Yan, Zhiquan He, Zhihai He, (2020) A Comprehensive Survey on Geometric Deep Learning, IEEE Access, vol. 8, pp. 35929-35964. `https://ieeexplore.ieee.org/document/9025496`