Choosing Meaning-Preserving Embeddings for RAG: From Infinite Banach to Finite Practical Vector Spaces

Stan Miasnikov, September 2025
stanmiasnikov@gmail.com

Abstract

Most guides to embedding models for RAG emphasize empirical leaderboards and heuristics, leaving open the principled question of when an embedding is "large enough" and why one model should be preferred over another. In this paper, we aim to place embedding choice on a theoretical footing within the Recursive Consciousness (RC) framework [1]. Previously, we proved that a full & faithful infinite embedding into an infinite Banach-enriched category exists for any normed-category of semantics [2]. Now we show that practical finite embeddings used in language models and specifically in retrieval-augmented generation (RAG) act as bi-Lipschitz approximations whose operator distortion in the embedded meaning-transfer pipeline ($E \circ M \circ I$) yields *two-sided bounds* on Jensen–Shannon divergence between induced belief distributions. These bounds give *testable* criteria for meaning preservation, translate into quantitative prescriptions for dimension and context (with human-language scales as guidance). Beyond size, we catalogue structural properties: isotropy, stability, invariance/equivariance, compositionality, calibration, domain/temporal alignment, with diagnostics and mitigations that materially affect retrieval quality. Thus, unlike prior surveys, this work elevates embedding choice from heuristics to *principled design* with measurable tolerances for RAG/agentic and human/AI communication settings.

1 Introduction

In the RC framework, each agent A_i has a semantic category $\mathcal{C}_{\text{sem}}^{(i)}$ and communicates through a shared symbolic category $\mathcal{C}_{\text{sym}}$ via interpretation/meaning functors $I_i : \mathcal{C}_{\text{sem}}^{(i)} \to \mathcal{C}_{\text{sym}}$ and $M_i : \mathcal{C}_{\text{sym}} \to \mathcal{C}_{\text{sem}}^{(i)}$ [3, 2]. To compare meanings quantitatively we use an embedding functor E into a Banach/vector space, where geometric proximity proxies semantic proximity; this underpins the dyadic and group mutual-understanding scores that combine cosine geometry with Jensen-Shannon divergence (JS) [3, 4]. In RAG architectures, such embeddings serve as the bridge between semantic queries (e.g., user intent in $\mathcal{C}_{\text{sem}}^{(i)}$) and retrieved contexts, where distortion in E risks amplifying retrieval errors or semantic drift.

Motivation and gap. Recent RAG surveys and practitioner guides agree that *embedding choice dominates retrieval quality*, yet their guidance is largely empirical or heuristic, leaving unanswered the key design question: *how large is large enough, and why?* Comprehensive surveys consolidate architectures, training strategies, and robustness issues but stop short of *theoretical guarantees* that relate embedding properties to meaning preservation or belief agreement [5, 6, 7]. Practitioner articles and blogs provide useful checklists (model families, chunking, hybrid retrieval) but likewise lack principled bounds [8, 9, 10]. Consequently, practitioners often overfit to leaderboards or isolated benchmarks, without guarantees that a chosen embedding preserves the semantics required by their application.

Novelty and contributions. While the dimensional scaling $m = \Theta(\eta^{-2} \log N)$ (with a union bound over $\binom{N}{2}$ pairs) is classical Johnson-Lindenstrauss, this paper is, to our knowledge, the first to (i) derive and *use* a *two-sided, testable* information-theoretic bound that links the embedded channel deviation $\|E_m \circ (\Phi - \text{id})\|_{\text{op}}$ directly to the belief disagreement term D_{JS} in the RC metric *via* (8); (ii) turn an application budget τ on D_{JS} into concrete engineering choices by composing the geometric tolerance η (hence m through (12)) with a context-capacity budget W (13), yielding an *end-to-end sizing recipe* for embeddings and prompts in RAG/multi-agent pipelines; (iii) explicitly account for the practical *two-E* application and cross-gauge calibration by absorbing these effects into measurable constants $C_{\text{low}}^{(E)}$ and $C_{\text{high}}^{(E)}$ on the realized session set Σ; and (iv) propagate these bounds to dyadic and group mutual-understanding scores used in RC, making fidelity losses attributable between E, I, and M. In short, we elevate embedding selection from leaderboard heuristics to *principled, verifiable design rules*: specify τ, map it to an admissible $\|E_m \circ (\Phi - \mathrm{id})\|_{\mathrm{op}}$ by (8), choose η and thus m by (12), size W by (13), and validate the calibration constants empirically on Σ. Section 2 formalizes the ideal and finite embeddings and derives the two-sided JS bounds; Section 3 instantiates these bounds into practical sizing rules for m and W.

2 Embeddings as Functors and the Ideal Limit

2.1 The embedding functor $E_\infty : \mathcal{C}_{\mathrm{sem}} \to \mathbf{Ban}_\infty$

The embedding functor E_∞ concretizes the abstract semantic category into a $\mathbf{Ban}_\infty$ [1] space, aiming to preserve the semantic structure "as geometry". Its effectiveness hinges on accurately encoding meanings (e.g., cosine similarity correlating with human judgments). Poor embeddings (insufficient dimension, anisotropy, misalignment) distort norms, corrupting RC metrics. We justify the assumption that suitable E_∞ can be chosen.

$$\begin{tikzcd}
\mathcal{C}^{(i)}_{\mathrm{sem}} \arrow[r, bend left=10, "I_i"] \arrow[rr, bend left=40, "\Phi_{i\to j}=M_j\circ I_i"] \arrow[d, "E_\infty"'] & \mathcal{C}_{\mathrm{sym}} \arrow[r, bend left=10, "M_j"] & \mathcal{C}^{(j)}_{\mathrm{sem}} \arrow[d, "E_\infty"] \\
\mathbf{Ban}_\infty \arrow[rr, dashed, bend right=15, "\widehat{\Phi}_{i\to j}=E_\infty\circ M_j\circ I_i"'] & & \mathbf{Ban}_\infty
\end{tikzcd}$$

Proposition 2.1 (Existence of faithful, norm-preserving embeddings into an infinite Banach space)**.** *Let* $\mathrm{Ob}(\mathcal{C}_{\mathrm{sem}})$ *be the metric space of semantic objects, where the distance function is induced by the category's normed structure. By the Kuratowski embedding theorem, there exists a faithful, isometric embedding functor*

$$E_\infty :\ \mathcal{C}_{\mathrm{sem}} \hookrightarrow \mathbf{Ban}_\infty$$

into an infinite-dimensional Banach space such that

$$\|x - y\|_{\mathcal{C}_{\mathrm{sem}}} = \|E_\infty(x) - E_\infty(y)\|_{\mathbf{Ban}_\infty} \quad \textit{for all } x, y,$$

This provides the ideal metric space into which semantic objects can be placed without distortion.

For proof see the Appendix [2].

Remark 2.2 (Epistemic limits)*.* From an LLM's internal perspective, E is epistemically full/faithful/norm-preserving: the model has no access to meanings outside its own embedding topology; all internal judgments are governed by E.

Remark 2.3 (Shared Architectures)*.* If two agents share the same E and input x, then $E(x)$ is identical for both. Divergence (if any) arises from M and downstream inference. Thus the RC mutual understanding metric attributes representational misalignment to M when E is shared, making E a controllable bottleneck.

Remark 2.4 (Finite approximations)*.* Strict isometries into finite $\mathbb{R}^m$ may not exist globally, but one can embed isometrically into ℓ^∞ (or a large $\mathbb{R}^N$) and then reduce dimension with controlled distortion (JL/bi-Lipschitz), yielding practical E_m.

2.2 Finite Embeddings and Two-Sided Information Bounds

We connect the ideal, full & faithful embedding E_∞ in an $\mathbf{Ban}_\infty$ setting to finite, practical embeddings $E_m \subset \mathbb{R}^m$ used in computations, and obtain two-sided control of information disagreement by structural distortion. The *upper* bound was established earlier in the Appendix; here we supply a complete *lower* bound and integrate Johnson-Lindenstrauss (JL)-type guarantees for the realized finite set of semantic objects.

[1] Here $\mathbf{Ban}_\infty$ denotes an infinite-dimensional Banach-enriched category (e.g., the category of real Banach spaces with bounded linear maps, or a Hilbert subcategory).

$$\begin{tikzcd}
\mathcal{C}_{\mathrm{sem}}^{(i)} \arrow[r, "I_i"] \arrow[rr, bend left, "\Phi_{i\to j}=M_j\circ I_i"] \arrow[d, "E_m"'] & \mathcal{C}_{\mathrm{sym}} \arrow[r, "M_j"] & \mathcal{C}_{\mathrm{sem}}^{(j)} \arrow[d, "E_m"] \\
\mathbb{R}^m \arrow[rr, dashed, bend right, "\widehat{\Phi}_{i\to j}=E_m\circ M_j\circ I_i"'] & & \mathbb{R}^m
\end{tikzcd}$$

Working set and JL-type finite embedding. Let $\Sigma \subset \mathrm{Ob}(\mathcal{C}_{\mathrm{sem}})$ be the finite set of semantic objects actually used in a session (e.g., agent restatements and hypothesis centroids). Assume (as in Prop. 2.1) that d_{sem} on Σ is induced by a Hilbertian norm via an (ideal) embedding $E_\infty : \Sigma \to \mathcal{H}$ with $\mathcal{H}$ a (possibly infinite-dimensional) real Hilbert space.[2] Then a (random) linear map $\Pi : \mathcal{H} \to \mathbb{R}^m$ exists such that, for any failure probability $\delta \in (0,1)$ and distortion parameter $\eta \in (0,1/2]$,

$$(1-\eta)\,\|E_\infty x - E_\infty y\|_2 \;\le\; \|\Pi E_\infty x - \Pi E_\infty y\|_2 \;\le\; (1+\eta)\,\|E_\infty x - E_\infty y\|_2 \quad \forall\, x,y \in \Sigma, \tag{1}$$

with

$$m \;\ge\; \frac{2\ln|\Sigma| \;+\; \ln(1/\delta)}{\frac{\eta^2}{2} - \frac{\eta^3}{3}} \;\;=\;\; \Theta\big(\eta^{-2}\log(|\Sigma|/\delta)\big), \tag{2}$$

with high probability [11, 12, 13, 14]. In practice we take $E_m := \Pi \circ E_\infty : \Sigma \to \mathbb{R}^m$ as the shared finite embedding for all agents, so that pairwise distances on Σ incur only the controlled distortion η.

Purely metric fallback. If one starts only with a metric d_{sem} on Σ (with no a priori Hilbertian structure), Bourgain's theorem embeds $(\Sigma, d_{\mathrm{sem}})$ into $(\ell_2, \|\cdot\|_2)$ with distortion $O(\log|\Sigma|)$ and dimension $O(\log^2|\Sigma|)$ [15] For the explicit $O(\log^2 n)$ target dimension refinement [16]); composing this with (1) gives a bi-Lipschitz finite embedding whose overall distortion is the product of the Bourgain factor and $(1 \pm \eta)$.

Channels, embedding, and the "two-E" factor. For inter-agent meaning transfer $\Phi = M_j \circ I_i : \mathcal{C}_{\mathrm{sem}} \to \mathcal{C}_{\mathrm{sem}}$, the structural deviation we evaluate in practice is computed *after* embedding:

$$\delta \;:=\; \big\|\, E_m \circ \Phi \;-\; E_m \circ \mathrm{id}\,\big\|_{\mathrm{op};\,\mathrm{span}(E_m\Sigma)}. \tag{3}$$

If a single shared E_m is used, then $\delta \le \|E_m\|_{\mathrm{op}} \cdot \|\Phi - \mathrm{id}\|_{\mathrm{op}}$. If two calibrated embeddings $E_m^{(i)}$ (source gauge) and $E_m^{(j)}$ (target gauge) are usedinto the same ambient space, a triangle inequality gives

$$\|E_m^{(j)} \circ \Phi - E_m^{(i)}\| \;\le\; \|E_m^{(j)}\|\,\|\Phi - \mathrm{id}\| \;+\; \|E_m^{(j)} - E_m^{(i)}\|.$$

Under matched Lipschitz constants and small cross-gauge calibration gap, the right-hand side is at most a constant factor (at most ≈ 2) times $\|\Phi - \mathrm{id}\|$, which we absorb into C_{high} and C_{low} below. Intuitively: in practice we apply E to both $E \circ \Phi$ and $E \circ \mathrm{id}$, so any embedding distortion is incurred *twice*; we budget for this multiplicative effect in the constants.

2.2.1 Belief mapping and two-sided information bounds

Belief mapping in the embedded space. Fix a shared finite embedding E_m and write $x := E_m(s)$, $x' := E_m(\Phi s)$, with unit-normalized $e := x/\|x\|_2$ and $e' := x'/\|x'\|_2$. Given unit hypothesis directions $v_k := E_m(h_k)/\|E_m(h_k)\|_2$ for $k = 1,\dots,H$, collect them as the matrix $V = [v_1 \;\cdots\; v_H] \in \mathbb{R}^{m\times H}$. Define scores $s_k(e) = \alpha\langle e, v_k\rangle$ and posteriors $P(e) = \mathrm{softmax}(s(e))$ with temperature $\alpha > 0$. Assume standard smoothing so that all coordinates lie in $[\mu, 1-\mu]$ for some $\mu \in (0, \frac{1}{2}]$ on the working set Σ. Let

$$\kappa_V := \|V^\top\|_{2\to 2}, \qquad \sigma_V := \sigma_{\min}\big(V^\top|_{\mathrm{span}(E_m\Sigma)}\big).$$

[2] The standard Johnson–Lindenstrauss (JL) lemma applies to any finite subset of a Hilbert space; the ambient dimension of $\mathcal{H}$ is irrelevant for the bound [11, 14].

Theorem 2.5 (Upper quadratic control in the embedded space). *Let*

$$\delta_E \ := \ \big\| \, E_m \circ (\Phi - \mathrm{id}) \, \big\|_{\mathrm{op;\,span}(E_m \Sigma)} .$$

There exists a constant

$$C^{(E)}_{\mathrm{high}} \ = \ C^{(E)}_{\mathrm{high}}\big(\alpha, \mu, \|V^\top\|_{2\to 2}, C_{\mathrm{norm}}\big)$$

depending only on the temperature α, smoothing floor μ, the hypothesis geometry (via $\|V^\top\|_{2\to 2}$), and the Lipschitz constant $C_{\mathrm{norm}} \geq 1$ of $x \mapsto x/\|x\|_2$ on $E_m\Sigma$, such that for all $s \in \Sigma$,

$$D_{\mathrm{JS}}\big(P(e) \,\|\, P(e')\big) \ \leq \ C^{(E)}_{\mathrm{high}} \ \delta_E^2. \tag{4}$$

Proof. Step 1: smooth upper bound of JS by ℓ_2 distance on the simplex. Let $M = (P + Q)/2$. By definition,

$$D_{\mathrm{JS}}(P\|Q) = \tfrac{1}{2}\big(D_{\mathrm{KL}}(P\|M) + D_{\mathrm{KL}}(Q\|M)\big).$$

Using the standard inequality (with base-2 logarithms)

$$D_{\mathrm{KL}}(R\|S) \ \leq \ \frac{1}{\ln 2}\,\chi^2(R, S) \ = \ \frac{1}{\ln 2}\sum_i \frac{(R_i - S_i)^2}{S_i},$$

we obtain

$$D_{\mathrm{JS}}(P\|Q) \ \leq \ \frac{1}{2\ln 2}\sum_i \frac{(P_i - M_i)^2 + (Q_i - M_i)^2}{M_i}.$$

Since $P - M = \frac{1}{2}(P - Q)$ and $Q - M = -\frac{1}{2}(P - Q)$, the two numerators are equal and

$$D_{\mathrm{JS}}(P\|Q) \ \leq \ \frac{1}{4\ln 2}\sum_i \frac{(P_i - Q_i)^2}{M_i}.$$

Under the smoothing assumption $M_i \geq \mu$ for all i, this yields the quadratic bound

$$D_{\mathrm{JS}}(P\|Q) \ \leq \ \frac{1}{4\,\mu\,\ln 2}\,\|P - Q\|_2^2. \tag{5}$$

Step 2: Lipschitz control of $P(e)$ with respect to e. Write scores $s_k(e) = \alpha\langle e, v_k\rangle$, $s(e) \in \mathbb{R}^m$. The softmax Jacobian with respect to s is $J_{\mathrm{sm}}(s) = \mathrm{diag}(P) - PP^\top$, whose spectral norm is $\|J_{\mathrm{sm}}\|_{2\to 2} \leq \frac{1}{4}$ for any P (it is the covariance matrix of a categorical distribution with one trial). By the chain rule, $\nabla_e P = J_{\mathrm{sm}}(s)\cdot \alpha V^\top$, thus for any e, e',

$$\|P(e) - P(e')\|_2 \ \leq \ \sup_\xi \|\nabla_e P(\xi)\|_{2\to 2}\,\|e - e'\|_2 \ \leq \ \frac{\alpha}{4}\,\|V^\top\|_{2\to 2}\,\|e - e'\|_2 \ \leq \ \frac{\alpha\,\kappa_V}{4}\,\|e - e'\|_2.$$

Step 3: from embedded channel deviation to $\|e - e'\|_2$. Let $C_{\mathrm{norm}} \geq 1$ be a Lipschitz constant of the normalization map $x \mapsto x/\|x\|_2$ on $E_m\Sigma$ (finite, since $\|E_m(s)\|_2$ is bounded away from 0 on Σ). Then

$$\|e - e'\|_2 \ \leq \ C_{\mathrm{norm}}\,\|E_m(\Phi s) - E_m(s)\|_2 \ \leq \ C_{\mathrm{norm}}\,\delta_E,$$

by the definition of δ_E as the operator norm of $E_m \circ (\Phi - \mathrm{id})$ restricted to $\mathrm{span}(E_m\Sigma)$.

Step 4: combine. Plugging the Lipschitz bound into (5) yields

$$D_{\mathrm{JS}}\big(P(e) \,\|\, P(e')\big) \ \leq \ \frac{1}{4\mu\,\ln 2}\left(\frac{\alpha\,\kappa_V}{4}\,C_{\mathrm{norm}}\,\delta_E\right)^2 \ = \ \frac{\alpha^2\kappa_V^2}{64\,\mu\,\ln 2}\,C^2_{\mathrm{norm}}\,\delta_E^2, \tag{6}$$

which is (4) with $C^{(E)}_{\mathrm{high}} = \frac{\alpha^2\kappa_V^2}{64\,\mu\,\ln 2}\,C^2_{\mathrm{norm}}$. □

Lower quadratic control in the embedded space (by reference). By the Pinsker-type lower bound established (see Theorem [LOWER-JS] in [2]), together with the co-Lipschitz behaviour of (i) the softmax map on the calibrated region and (ii) the unit normalization on $E_m\Sigma$, there exists a constant

$$C_{\text{low}}^{(E)} \;=\; C_{\text{low}}^{(E)}\big(\alpha, \mu, \sigma_V, c_{\text{norm}}\big) \;>\; 0,$$

depending on α, μ, a restricted minimal singular value σ_V of $V^\top$ on $\operatorname{span}(E_m\Sigma)$, and a co-Lipschitz constant $c_{\text{norm}} \in (0,1]$ of $x \mapsto x/\|x\|_2$ on $E_m\Sigma$, such that

$$D_{\text{JS}}\big(P(e) \,\|\, P(e')\big) \;\geq\; C_{\text{low}}^{(E)}\;\delta_E^2. \tag{7}$$

(Concretely, one admissible instantiation is $C_{\text{low}}^{(E)} = \frac{\alpha^2\,\sigma_V^2}{128}\,c_{\text{norm}}^2$ under the standard binary logistic sensitivity; sharper μ-explicit variants are possible but not required here.)

Sandwich on the realized session set (embedded form). From (4)-(7) we obtain

$$\boxed{C_{\text{low}}^{(E)}\;\big\|E_m \circ (\Phi - \text{id})\big\|_{\text{op}}^2 \;\;\leq\;\; D_{\text{JS}}\big(P(e) \,\|\, P(e')\big) \;\;\leq\;\; C_{\text{high}}^{(E)}\;\big\|E_m \circ (\Phi - \text{id})\big\|_{\text{op}}^2} \tag{8}$$

on the working set Σ. If one prefers to express the right-hand side in the semantic norm using a JL-type $(1 \pm \eta)$ bi-Lipschitz guarantee on Σ, the constants adjust by at most factors of $(1 \pm \eta)^2$.

It can also be written as:

$$\sqrt{\frac{D_{\text{JS}}}{C_{\text{high}}^{(E)}}} \;\leq\; \big\|E_m \circ (\Phi - \text{id})\big\|_{\text{op}} \;\leq\; \sqrt{\frac{D_{\text{JS}}}{C_{\text{low}}^{(E)}}}. \tag{9}$$

Remark 2.6 (When does $D_{\text{JS}} = 0$?). With an ideal full & faithful $E_\infty : \mathcal{C}_{\text{sem}} \hookrightarrow \mathbf{Ban}$, we have $D_{\text{JS}} = 0$ iff the entire channel is lossless on the relevant subspace: $M \circ I = \text{id}$ (no loss in M or I) *and* E_∞ is faithful. Thus, even perfect E_∞ does not force $D_{\text{JS}} = 0$ in RC; any non-ideal M or I keeps $D_{\text{JS}} > 0$, correctly attributing disagreement to the communication channel rather than to the embedding.

Corollary 2.7 (Control of per-pair fidelity and the group score). With $u_{ij}^{(n)} = (1 - D_{\text{JS}}) \cdot \frac{1+\cos}{2}$ (cosine on unit-normalized E_m), (8) implies

$$\Big(1 - C_{\text{high}}^{(E)}\;\big\|E_m \circ (\Phi - \text{id})\big\|_{\text{op}}^2\Big) \cdot \frac{1+\cos}{2} \;\;\leq\;\; u_{ij}^{(n)} \;\;\leq\;\; \Big(1 - C_{\text{low}}^{(E)}\;\big\|E_m \circ (\Phi - \text{id})\big\|_{\text{op}}^2\Big) \cdot \frac{1+\cos}{2}, \tag{10}$$

and the strict geometric mean across pairs yields the corresponding bounds for $\mathcal{U}_{\text{group}}^{(n)}$ [4].

3 How Large is "Large Enough"?

We now turn the two-sided embedded bound (8) into *actionable prescriptions* for choosing (a) the embedding dimension m and (b) the context window size (tokens) in RAG/multi-agent settings. The aim is to keep D_{JS} below an application budget τ for all relevant pairs, with high probability.

Dimension via Johnson–Lindenstrauss (JL). Let Σ be the finite working set of semantic objects concurrently in play (restatements, retrieved snippets, hypothesis centroids), with $|\Sigma| = N$. For any two points $x, y \in \Sigma$, let $u = E_\infty(x) - E_\infty(y)$ be the vector connecting them. For $0 < \eta \leq \frac{1}{2}$ and failure probability $\delta \in (0,1)$, a JL projection $\Pi : \mathcal{H} \to \mathbb{R}^m$ preserves all pairwise distances on Σ within $(1 \pm \eta)$ [3] provided [11, 12, 13, 14]. The probability of significant distortion is exponentially small:

$$\mathbb{P}\left(\big|\|\Pi u\|_2^2 - \|u\|_2^2\big| > \eta\|u\|_2^2\right) \leq 2\exp(-c\,\eta^2 m) \tag{11}$$

We will set constant $c \in [\frac{1}{4}, 1]$ to $c = \frac{1}{4}$, a standard "conservative" value that arises from common proofs (e.g., using Gaussian projectors). To ensure *all* pairs in Σ are preserved, we sum the failure probabilities for each pair. There are $\binom{N}{2} = \frac{N(N-1)}{2} < \frac{N^2}{2}$ pairs. A simple union bound gives the total failure probability:

[3] There's a distinction between the norm-distortion parameter ϵ and the distance-distortion parameter η, where $\epsilon \approx 2\eta$. We simplified this by using η throughout, which is common in such explanations.

N	η	K	s	δ	m_{min}	W_{min}
500	0.25	5	100	10^{-2}	1,090	2,775
1,000	0.20	10	200	10^{-2}	1,842	4,500
1,000	0.20	20	300	10^{-2}	1,924	9,100
2,500	0.15	40	300	10^{-2}	3,600	16,000
5,000	0.15	50	300	10^{-2}	3,848	19,450

Table 1: For indices up to $N \leq 10^3$, $m \approx 2$k with $W_{\min} \approx 10$-16k comfortably covers $K \in [5, 20]$ at $s \leq 300$, so an 8k or 16k context is sufficient. For larger corpora ($N \approx 2.5$k-5k) and broader fan-in ($K \in [40, 50]$ at $s \approx 300$), $m \approx 3.6$-4.0k and $W_{\min} \approx 19$-24k; a 32k context is advisable to keep the truncation component of $\|E_m \circ (\Phi - \mathrm{id})\|$ small.

$$P(\text{any pair fails}) \leq \binom{N}{2} \cdot 2\exp(-c\eta^2 m) < N^2 \exp(-c\eta^2 m)$$

Solving for m: we want this total failure probability to be less than our desired threshold δ:

$$N^2 \exp(-c\eta^2 m) < \delta$$

Taking a logarithm, substituting the conservative constant $c = \frac{1}{4}$, and solving for m we have:

$$\frac{1}{4}\eta^2 m > 2\ln N + \ln(1/\delta) \implies m > \frac{8\ln N + 4\ln(1/\delta)}{\eta^2}$$

We get the final lower bound for the embedding dimension m:

$$m > \frac{8\,\ln(N/\sqrt{\delta})}{\eta^2}\,. \tag{12}$$

In our tables we instantiate (12) with a conservative constant (absorbing the "two-E" application and calibration slack) to produce working lower bounds $m_{\min}$ at $\delta = 10^{-2}$.

Context-window sizing (RAG/multi-agent). Let K be the retrieval fan-in (top-K), s the average chunk size (tokens), q the query/system overhead, and $\rho \in [0.1, 0.2]$ a metadata/separator overhead factor. A simple, robust capacity budget is

$$W_{\min} \;\approx\; q \;+\; (1+\rho)\,K\,s \;+\; g, \tag{13}$$

where g is a guard band for prompt scaffolding and generation (e.g., $g \simeq 1{-}2$k tokens for long-form answers; increase for multi-agent turns). In words: you should be able to fit the whole retrieved slate plus overheads *and* leave headroom. If W is materially smaller than $W_{\min}$, the effective channel Φ truncates relevant context, increasing $\|E_m \circ (\Phi - \mathrm{id})\|$ and, by (8), the D_{JS} term.

Estimating minimal embedding dimensions. The table below uses the target failure probability ($\delta = 10^{-2}$), the JL tolerance η, and practical RAG knobs (K and average chunk size s). We evaluate (12) to obtain $m_{\min}$ and the context budget rule (cf. (13)) with typical $q = 200$, $\rho = 0.15$, $g = 2000$ to obtain $W_{\min}$.

Tying size to the D_{JS} budget. Given an application target τ for the information term, invert (8):

$$\big\|E_m \circ (\Phi - \mathrm{id})\big\|_{\mathrm{op}} \;\leq\; \sqrt{\tfrac{\tau}{C^{(E)}_{\mathrm{high}}}}.$$

Choose η (hence m via (12)) so that JL distortion on Σ does not exceed this threshold, and choose W via (13) so that context truncation does not push the channel over it. Practically, you tune (α, μ) (softmax temperature/smoothing) and the hypothesis frame V to keep $C^{(E)}_{\mathrm{high}}$ moderate; then *dimension* and *window* are the two main levers that keep D_{JS} within budget.

model_name	**dim**	**DataFit**	**BTI**	$\boldsymbol{\delta_{\text{op}}}$	$\boldsymbol{\eta_{\text{JL}}}$	**CCS**	$\overline{\textbf{cos}}$	$\boldsymbol{N_{\max,\eta=0.2}}$	$\boldsymbol{N_{\max,\eta=0.15}}$
text-embedding-3-large	3072	0.710	0.750	1.406	0.161	0.553	0.423	468,557	565
text-embedding-3-small	1536	0.599	0.749	1.433	0.227	0.517	0.348	216	7
SFR-Embedding-Mistral	4096	0.521	0.919	1.400	0.139	0.404	0.670	78,406,305	10,070
Qwen3-Embedding-8B	4096	0.370	0.820	1.404	0.139	0.186	0.493	78,406,305	10,070
Qwen3-Embedding-4B	2560	0.337	0.762	1.411	0.176	0.259	0.557	36,221	133
granite-embedding-30m-english	384	0.077	0.710	4.552	0.454	0.318	0.550	0	0
granite-embedding-small-english-r2	384	0.040	0.681	4.236	0.454	0.277	0.743	0	0

Table 2: **QA (TriviaQA)** — Embedding capacity, channel deviation, and fit. Column meanings: *DataFit* is the bounded score in $(0, 1]$ (higher is better); *BTI* $\in (0, 1]$ measures two-sided band tightness (higher is tighter); δ_{op} is the channel-deviation operator norm on the realized span (lower is better); η_{JL} is the JL tolerance at (m, N, δ) (lower is better); *CCS* is the channel-correlation score $\text{CCS} := \frac{1}{2}(1 + \max(0, \rho)) \in [0, 1]$ with $\rho = \text{corr}(D_{\text{JS}}, \|\Delta\|_2^2)$ on the realized set (higher is better); $\overline{\cos}$ is the mean cosine (diagnostic); $N_{\max}$ values are the maximum corpus sizes supported at target tolerances $\eta\!=\!0.20, 0.15$ under the current m (larger is better; zeros indicate insufficient capacity at that tolerance).

model_name	**dim**	**DataFit**	**BTI**	$\boldsymbol{\delta_{\text{op}}}$	$\boldsymbol{\eta_{\text{JL}}}$	**CCS**	$\overline{\textbf{cos}}$	$\boldsymbol{N_{\max,\eta=0.2}}$	$\boldsymbol{N_{\max,\eta=0.15}}$
text-embedding-3-large	3072	0.773	0.705	1.406	0.161	0.641	0.367	468,557	565
SFR-Embedding-Mistral	4096	0.650	0.866	1.400	0.139	0.530	0.619	78,406,305	10,070
text-embedding-3-small	1536	0.636	0.716	1.427	0.227	0.583	0.331	216	7
Qwen3-Embedding-8B	4096	0.602	0.784	1.404	0.139	0.342	0.425	78,406,305	10,070
Qwen3-Embedding-4B	2560	0.453	0.700	1.409	0.176	0.354	0.485	36,221	133
granite-embedding-30m-english	384	0.122	0.733	4.399	0.454	0.501	0.566	0	0
granite-embedding-small-english-r2	384	0.063	0.685	4.221	0.454	0.453	0.751	0	0

Table 3: **QA (HotpotQA)** — Embedding capacity, channel deviation, and fit. Column meanings as in Table 2.

4 From Theory to Practical Evaluation

The bounds in §2.2 give rise to concrete, *testable* quantities that we estimate on a target workload to judge whether a finite embedding E_m is "large enough" and well-calibrated: (i) a *geometric capacity* tolerance η_{JL} (Johnson-Lindenstrauss) that captures how tightly distances concentrate at the realized (m, N, δ) scale (cf. (12)-(11)); (ii) a *channel deviation* norm $\delta_{\text{op}} := \big\|E_m \circ (\Phi - \text{id})\big\|_{\text{op; span}(E_m\Sigma)}$ that measures distortion of the meaning-transfer pipeline on the session span (cf. (8)); and (iii) a *band tightness* statistic summarizing the two-sided, data-estimated JS-geometry relation in (8)-(9).

Band tightness (BTI). Let $R := \text{JS}/\|\Delta\|_2^2$ denote the per-pair ratio formed with *unit-normalized* embedding differences Δ (for cross-model comparability). We estimate a robust two-sided band by the quantiles $C_{\text{low}}^{(E)} := Q_{0.05}(R)$ and $C_{\text{high}}^{(E)} := Q_{0.95}(R)$ on the realized set Σ. To obtain a *bounded, scale-free* tightness we use the geometric/arithmetical mean index

$$\text{BTI} \;:=\; \frac{2\sqrt{C_{\text{low}}^{(E)}\,C_{\text{high}}^{(E)}}}{C_{\text{low}}^{(E)} + C_{\text{high}}^{(E)}} \;\in\; (0, 1],$$

which equals 1 iff the band collapses ($C_{\text{low}}^{(E)}\!=\!C_{\text{high}}^{(E)}$) and decreases smoothly as the band loosens. This choice is symmetric in the bounds, insensitive to overall scale, and less punitive than linear forms when $C_{\text{low}}^{(E)}$ is small yet the band is not egregiously wide.

DataFit (bounded, task-aware metric). We compress the theory-linked quantities into a single *DataFit* score in $(0, 1]$ for each model and workload. Besides BTI, η_{JL}, and δ_{op}, we include a *task factor* $T \in [0, 1]$ that reflects how the observed geometry couples to information in that task. For **paraphrase** (near-identity channel) we take $T = \text{CCS}$, where $\text{CCS} := \frac{1}{2}(1 + \rho) \in [0, 1]$ and ρ is the (Spearman or Pearson) correlation between D_{JS} and $\|\Delta\|_2^2$ across pairs; this rewards a stable bounds "sandwich" relation. For **QA** (lossy channel) we also discourage question-answer *collapse* by multiplying with a cohesion penalty $(1 - \overline{\cos})$; here $\overline{\cos}$ is the mean cosine between the question and an *answer-in-context* representation (answer sentence or small window), which is more stable than a bare short span. With $k = 2.0$ and a tiny $\varepsilon = 10^{-12}$ for numerical safety, we define

$$\text{DataFit}_{\text{Para}} \;=\; 1 - \exp\!\left(-\; k \cdot \frac{\text{BTI}}{\eta_{\text{JL}}\big(1 + \delta_{\text{op}}\big) + \varepsilon} \cdot \text{CCS} \right),$$

model_name	**dim**	**DataFit**	**BTI**	$\boldsymbol{\delta_{\mathrm{op}}}$	$\boldsymbol{\eta_{\mathrm{JL}}}$	**CCS**	$\overline{\mathbf{cos}}$	$\boldsymbol{N_{\max,\eta=0.2}}$	$\boldsymbol{N_{\max,\eta=0.15}}$
SFR-Embedding-Mistral	4096	0.780	0.454	1.961	0.139	0.686	0.850	78,406,305	10,070
Qwen3-Embedding-8B	4096	0.764	0.430	1.960	0.139	0.690	0.816	78,406,305	10,070
Qwen3-Embedding-4B	2560	0.679	0.450	1.961	0.176	0.658	0.810	36,221	133
text-embedding-3-large	3072	0.466	0.198	1.960	0.161	0.754	0.720	468,557	565
text-embedding-3-small	1536	0.445	0.258	1.961	0.227	0.768	0.729	216	7
granite-embedding-small-english-r2	384	0.351	0.595	2.971	0.454	0.655	0.923	0	0
granite-embedding-30m-english	384	0.320	0.594	3.506	0.454	0.665	0.880	0	0

Table 4: **Paraphrase (STS-B)** — Embedding capacity, channel deviation, and fit. Column meanings as in Table 2.

model_name	**dim**	**DataFit**	**BTI**	$\boldsymbol{\delta_{\mathrm{op}}}$	$\boldsymbol{\eta_{\mathrm{JL}}}$	**CCS**	$\overline{\mathbf{cos}}$	$\boldsymbol{N_{\max,\eta=0.2}}$	$\boldsymbol{N_{\max,\eta=0.15}}$
Qwen3-Embedding-8B	4096	0.715	0.456	1.842	0.139	0.544	0.847	78,406,305	10,070
SFR-Embedding-Mistral	4096	0.640	0.407	1.778	0.139	0.485	0.885	78,406,305	10,070
text-embedding-3-large	3072	0.433	0.290	1.789	0.161	0.438	0.802	468,557	565
text-embedding-3-small	1536	0.382	0.370	1.848	0.227	0.421	0.800	216	7
Qwen3-Embedding-4B	2560	0.355	0.199	1.870	0.176	0.555	0.820	36,221	133
granite-embedding-small-english-r2	384	0.310	0.656	2.717	0.454	0.477	0.938	0	0
granite-embedding-30m-english	384	0.260	0.622	3.191	0.454	0.460	0.907	0	0

Table 5: **Paraphrase (MRPC)** — Embedding capacity, channel deviation, and fit. Column meanings as in Table 2.

$$\mathrm{DataFit}_{\mathrm{QA}} \;=\; 1-\exp\!\left(-\;k\cdot\frac{\mathrm{BTI}}{\eta_{\mathrm{JL}}\big(1+\delta_{\mathrm{op}}\big)+\varepsilon}\cdot\mathrm{CCS}\cdot\big(1-\overline{\cos}\big)\right).$$

Motivation. In both cases the exponential form yields a bounded, monotone map that increases with tighter information bands (larger BTI) and stronger geometry → information coupling (larger CCS), while decreasing with weaker geometric capacity or greater channel distortion (larger η_{JL} or δ_{op}). For paraphrase we omit raw cosine because lexical novelty can legitimately reduce cosine without violating meaning preservation; CCS captures the *functional* link prescribed by (8). For QA we explicitly penalize question-answer directional collapse through $(1-\overline{\cos})$, which is measured against answer-in-context to reflect downstream RAG use.

Protocol and constants. We hold the following knobs fixed across models for fair comparison: $H = 64$ hypothesis directions (shared frame), a smoothing floor $\mu = 0.00039$, $N = 1000$ randomly sampled pairs per task (or per corpus shard), and a JL failure budget $\delta = 10^{-2}$ when solving for η_{JL} from (12). The operator norm δ_{op} is computed on *unit-normalized* (or whitened) embeddings via a ridge-fitted linear map on the realized span to remove scale/anisotropy artifacts. The band $(C^{(E)}_{\mathrm{low}}, C^{(E)}_{\mathrm{high}})$ is always estimated from ratios formed with unit deltas.

Sampling and dataset dependence. All quantities are *dataset-dependent*: they reflect the realized geometry and channel on the target corpus. We therefore report DataFit from $N = 1000$ random pairs, however we recommend estimating on the full dataset (or reporting bootstrap confidence intervals over seeds) for deployment-grade alignment.

Dataset structure and adaptations. Our QA evaluation assumes the ”answer” unit is semantically rich (typically a multi-sentence passage), matching RAG deployments where models retrieve and reason over paragraphs rather than bare strings. Under this assumption, the band $(C^{(E)}_{\mathrm{low}}, C^{(E)}_{\mathrm{high}})$, the deviation δ_{op}, and the tolerance η_{JL} are estimated on geometry that mirrors real retrieval. Datasets with short or templated targets (e.g., extractive spans in SQuAD or near-duplicate pairs in QQP) can distort these measurements unless adapted. When such datasets are used, we recommend: (i) representing answers by the sentence (or a small window) containing the span (*answer-in-context*) rather than the bare span; (ii) computing δ_{op} on unit-normalized (or whitened) embeddings; and (iii) using the QA cohesion factor $(1-\overline{\cos})$ with the answer-in-context vector for very short answers. Multi-sentence QA sets such as WikiHop, HotpotQA, or TriviaQA, and paraphrase sets such as MRPC and STSB, align naturally with our setup.

Reading the tables. Across both QA corpora (HotpotQA, TriviaQA), larger OpenAI models lead or are competitive, driven by strong *CCS* (geometry→information coupling) and moderate η_{JL}, with *SFR* also strong thanks to consistently tight *BTI.* In paraphrase (STS-B, MRPC), *Qwen3-8B* and *SFR* rank highly as their bands remain tight and the channel deviation δ_{op} stays controlled; OpenAI models show solid CCS but somewhat looser bands on STS-B, which the DataFit penalizes as predicted by the theory. Granite baselines exhibit large η_{JL} and δ_{op}, producing low *DataFit*; their $N_{\max}$ entries at target tolerances are 0, indicating insufficient capacity at those η thresholds under current m. Overall, the rankings align with our bounds: DataFit increases with tighter bands (higher BTI) and stronger CCS, and decreases with geometric/operational stress (higher η_{JL} or δ_{op}).

5 Conclusion

This paper placed embedding selection for RAG on a principled footing within the Recursive Consciousness framework. Starting from an ideal full & faithful embedding into a Banach-enriched setting, we connected practical finite embeddings $E_m : \mathcal{C}_{\text{sem}} \to \mathbb{R}^m$ to two-sided, *testable* information bounds that **link the channel-deviation norm $\delta_{\text{op}} := \|E_m \circ (\Phi - \text{id})\|_{\text{op}}$ and the belief-disagreement** D_{JS} (cf. (8)-(9)), allowing an application budget on D_{JS} to be converted into an admissible δ_{op} (and vice versa). Together with Johnson-Lindenstrauss (JL) capacity control, these results yield explicit tolerances that translate directly into design knobs: target D_{JS} budgets, admissible geometric distortion η_{JL} (hence dimension m), and context-window sizing.

On the evaluation side, we introduced a bounded, task-aware *DataFit* score that aggregates three theory-grounded ingredients: band tightness (BTI) from the empirical JS-geometry sandwich, geometric capacity via η_{JL}, and operational distortion via δ_{op}, modulated by a task coupling factor (CCS, and cohesion for QA). The accompanying script [17] computes these metrics from specific datasets, producing rankings that align with our predictions.

Practically, our prescription is as follows: (i) specify an information budget τ; (ii) bound $\|E_m \circ (\Phi - \text{id})\|_{\text{op}}$ *via* the two-sided inequality; (iii) choose η_{JL} and thus m to respect the budget on the realized set; (iv) size the context window to avoid truncation-induced channel error; and (v) validate BTI/CCS on the target corpus. This elevates model choice from leaderboard heuristics to verifiable engineering with measurable tolerances.

Limitations and extensions include tightening constants for non-Hilbertian semantics, robustness under domain shift and anisotropy, multilingual and cross-modal gauges, and group-level guarantees under evolving corpora. Future work will integrate active calibration (adaptive hypothesis frame V, dynamic context-capacity budget W), retrieval policy co-design, and out-of-distribution monitors, sharpening the connection between the theory and high-stakes RAG deployment.

References

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

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

[3] 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`

[4] Miasnikov, S. (2025). Category-Theoretic Extension of Mutual Understanding to Group Communication. *Preprint.* `http://dx.doi.org/10.13140/RG.2.2.16746.99527`

[5] Sharma, C. (2025). Retrieval-Augmented Generation: A Comprehensive Survey of Architectures, Enhancements, and Robustness Frontiers. *arXiv:2506.00054.* `https://arxiv.org/abs/2506.00054`

[6] Fan, W., *et al.* (2024). A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. *Preprint.* `https://dl.acm.org/doi/10.1145/3637528.3671470`

[7] Caspari, L., *et al.* (2024). Beyond Benchmarks: Evaluating Embedding Model Similarity for Retrieval Augmented Generation Systems. *arXiv:2407.08275.* `https://arxiv.org/abs/2407.08275.`

[8] Enterprise Bot (2024). Choose the Best Embedding Model for Your Retrieval-Augmented Generation (RAG) System. *Enterprise Bot Blog.* https://www.enterprisebot.ai/blog/choose-the-best-embedding-model-for-your-retrieval-augmented-generation-rag-system.

[9] Harsoor, S. (2025). The Complete Guide to Embeddings and RAG: From Theory to Production. *Medium.* https://medium.com/@sharanharsoor/the-complete-guide-to-embeddings-and-rag-from-theory-to-production-758a16d747ac.

[10] Latimer, C. (2024). Mastering Advanced RAG Techniques: Elevating Your AI Capabilities. *Vectorize Blog.* https://vectorize.io/blog/mastering-advanced-rag-techniques-elevating-your-ai-capabilities.

[11] Johnson, W. B., & Lindenstrauss, J. (1984). Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics.* https://stanford.edu/class/cs114/readings/JL-Johnson.pdf.

[12] Dasgupta, S., & Gupta, A. (2003). An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures & Algorithms.* https://cseweb.ucsd.edu/~dasgupta/papers/jl.pdf.

[13] Achlioptas, D. (2003). Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences* https://doi.org/10.1016/S0022-0000(03)00025-4

[14] Vershynin, R. (2018). High-Dimensional Probability: An Introduction with Applications in Data Science. *Cambridge University Press.* https://doi.org/10.1017/9781108231596

[15] Bourgain, J. (1985). On Lipschitz embedding of finite metric spaces in Hilbert space. *Israel Journal of Mathematics.* https://link.springer.com/article/10.1007/BF02776078.

[16] Indyk, P., Matoušek, J., & Sidiropoulos, A. (2017). Low-distortion embeddings of finite metric spaces. *Handbook of Discrete and Computational Geometry chapter preprint*: https://www.csun.edu/~ctoth/Handbook/chap8.pdf.

[17] Miasnikov, S. (2025). Choosing Meaning-Preserving Embeddings for RAG. *github* https://github.com/phatware/embedding