The Shadow Theorem

What a Simpler Mind Cannot Know About a More Complex One

Spectral bounds on cross-intelligence capability estimation with empirical validation on LLMs

Tamás Nagy, Ph.D.

knowabilitylab.com

Draft • March 2026

Executive Summary (Non-Technical)

When a less capable system tries to evaluate a more capable one, **it does not merely make errors — it has structural blind spots that it cannot detect.** A mouse cannot appreciate that a human understands calculus; a pocket calculator cannot evaluate whether a supercomputer can prove theorems. This paper asks: can we make these intuitions mathematically precise?

We prove that **the estimation error has a hard floor** that no amount of observation can eliminate. A simpler system projects a complex system's capabilities into its own representational space, losing everything that does not fit — like a shadow cast from a higher-dimensional object onto a lower-dimensional wall. The lost dimensions are not experienced as gaps; they are simply invisible.

The most practically important result is **the calibration impossibility**: a simpler system cannot accurately estimate *how wrong* its estimate is. Knowing what you don't know requires the capacity to represent what you're missing — which is exactly the capacity the simpler system lacks. This is a mathematical formalization of the Dunning-Kruger effect.

These are not abstract impossibility results. We validate the theory empirically using large language models as a concrete intelligence hierarchy. GPT-2 Small (124M parameters) has a knowledge meta-rank of 10 — it organizes knowledge along 10 independent axes. TinyLlama (1.1B parameters) has meta-rank 31. When we project TinyLlama's knowledge structure into GPT-2's representational space, the Shadow Theorem's predictions are confirmed: 21 dimensions of capability are invisible, domain distinctions that TinyLlama makes correctly are conflated by GPT-2, and the estimation asymmetry is 83-fold — GPT-2's blind spot about TinyLlama is 33.7%, while TinyLlama's blind spot about GPT-2 is just 0.4%.

Most strikingly, the Unification Paradox is not merely theoretical — we observe the complete Unification Pulse within the GPT-2 family. Meta-rank follows the trajectory $R = 10 \to 6 \to 31$ as model size increases from Small (124M) through Medium (355M) to Large (774M). GPT-2 Medium undergoes a unification phase: despite being 3× larger, it has lower meta-rank ($R = 6$) with the lowest depth signature ($\delta = 0.038$), indicating more precise knowledge compression. GPT-2 Large then re-discriminates ($R = 31$, $\delta = 0.247$). The unified Medium is the most blind model in the entire hierarchy: its shadow of TinyLlama (45.5%) and Large (48.2%) exceeds even the smaller GPT-2 Small's (33.7% and 35.3%).

The paper matters for **AI safety**: if humans are the simpler system and a superhuman AI is the more complex one, the Shadow Theorem quantifies what humans structurally cannot evaluate about that AI — and proves that humans cannot know the size of this blind spot.

Abstract

We introduce the Shadow Theorem, a family of spectral bounds on the ability of a lower-capacity agent to estimate the capabilities of a higher-capacity agent. The core result is an application of the Eckart–Young theorem to intelligence estimation: if agent A can represent R_A independent capability dimensions and agent B can represent $R_B > R_A$, then A's optimal model of B has irreducible error at least $\sum_{k=R_A+1}^{R_B} \sigma_k^2(B)$, where σ_k are the singular values of B's capability operator. We prove four structural results: (1) the Shadow Bound on estimation error, (2) the Domain Blindness Theorem showing $R_B - R_A$ capability distinctions invisible to A, (3) the Observation Saturation Theorem showing the error floor is independent of observation time, and (4) the Calibration Impossibility Theorem showing A cannot accurately assess its own estimation error about B. We validate all four predictions empirically using the Latent of Latents framework (Nagy, 2026) applied to four models: GPT-2 Small ($R = 10$, $d = 768$), GPT-2 Medium ($R = 6$, $d = 1024$), GPT-2 Large ($R = 31$, $d = 1280$), and TinyLlama 1.1B ($R = 31$, $d = 2048$). Rank-10 reconstruction of TinyLlama's 41-domain knowledge structure loses 33.7% of variance; domain pairs undergo complete sign reversals (psychology–law: $+0.14 \to -0.96$) at low rank; the cross-architecture estimation asymmetry is 83-fold; and at every rank below full, 100% of the estimation error is invisible to the estimating agent. The GPT-2 family traces the complete Unification Pulse: $R = 10 \to 6 \to 31$ as parameters increase from 124M to 774M, with the depth signature δ confirming this trajectory is genuine (unification at Medium: $\delta = 0.038$; re-discrimination at Large: $\delta = 0.247$). This provides direct empirical evidence for the Unification Paradox: a phase-2 (discriminating) observer cannot distinguish a phase-1 (naive) system from a phase-3 (unified) system, because simplicity and profundity project to the same shadow. We discuss implications for AI safety, where the simpler agent is human and the more complex agent is a potential superhuman AI.

1. Introduction

A fundamental asymmetry governs the relationship between systems of different complexity: the simpler system cannot fully comprehend the more complex one, but the simpler system also cannot know *what it is missing*. This intuition appears across domains — in Gödel's incompleteness theorems (a formal system cannot prove its own consistency), in the halting problem (a Turing machine cannot decide the halting of all other Turing machines), in Kolmogorov complexity (a program of length n cannot generate all strings of complexity greater than n), and in the philosophical concept of cognitive closure (McGinn, 1989).

These results, however, are qualitative. They tell us that limits exist, but not how large they are, how they scale with the complexity gap, or whether they are empirically measurable. This paper provides quantitative answers by connecting the abstract question to concrete representational structure via spectral decomposition.

The key insight is that an agent's capacity to model the world — including to model other agents — is determined by the rank of its representational space. An agent with meta-rank R_A (the number of independent axes along which it organizes information) can construct models of other agents only within its R_A-dimensional subspace. When it encounters an agent with meta-rank $R_B > R_A$, its

model is necessarily a rank-R_A projection — a shadow of the true capability structure. The missing $R_B - R_A$ dimensions are not experienced as uncertainty or gaps; they are structurally absent from A's ontology.

This formulation makes the problem tractable because the singular value decomposition provides exact, computable bounds on what is lost in the projection. Moreover, the Latent of Latents framework (Nagy, 2026) provides a concrete operationalization: the meta-rank of a large language model is measurable, the singular values are computable, and the shadow projection can be performed and compared to the theoretical bound.

Contributions

1. **The Shadow Theorem** (Theorem 1): a spectral lower bound on cross-intelligence estimation error, tight in the Frobenius norm.
2. **The Domain Blindness Theorem** (Theorem 2): a counting result on invisible capability distinctions, with empirical demonstration of sign-reversal errors.
3. **The Observation Saturation Theorem** (Theorem 3): the estimation error converges to a nonzero floor independent of observation effort.
4. **The Calibration Impossibility Theorem** (Theorem 4): the simpler agent's estimate of its own estimation error is necessarily miscalibrated — empirically, 100% of the error is invisible at every rank.
5. **Empirical validation** on a four-model hierarchy (GPT-2 Small, Medium, Large, and TinyLlama 1.1B), confirming all four theoretical predictions including an 83-fold cross-architecture estimation asymmetry and the complete 4×4 shadow matrix.
6. **The Unification Pulse, empirically observed** (Section 5.5, Section 9): the GPT-2 family traces $R = 10 \to 6 \to 31$ with depth signatures $\delta = 0.082 \to 0.038 \to 0.247$, confirming the three-phase trajectory (discrimination → unification → re-discrimination). GPT-2 Medium — the unified phase — is the most blind model in the hierarchy despite being larger than Small.

Related Work

Computability theory. Turing (1936) and Gödel (1931) established fundamental limits on self-reference and computation. Rice's theorem (1953) shows that no non-trivial semantic property of programs is decidable. The arithmetic hierarchy (Kleene, 1943) stratifies undecidable problems into levels. Our contribution is to make these qualitative barriers quantitative via spectral decomposition.

Intelligence measurement. Hutter's AIXI (2005) provides a universal intelligence measure based on Kolmogorov complexity. Chollet (2019) proposes skill-acquisition efficiency as a practical intelligence measure. Hernández-Orallo (2017) develops agent-based intelligence tests. Our framework differs in focusing not on measuring intelligence itself, but on the limits of cross-intelligence *estimation*.

Representational similarity analysis. RSA (Kriegeskorte et al., 2008) compares representational geometries across systems via dissimilarity matrices — the same object as our domain similarity matrix $\mathbf{S}_A$. RSA has been applied extensively to compare brain regions, computational models, and behavioral responses. A recent Nature Machine Intelligence study (Muttenthaler et al., 2025) extends this to a generic framework for identifying latent dimensions underlying both human and DNN behavior, finding that DNNs exhibit low-dimensional structure but with different

processing strategies than humans. The Representational Alignment Hypothesis (Huh et al., 2026) shows that independently trained AI systems across modalities share geometric structure recoverable via RSA and topological analysis. Our contribution is to use the effective dimensionality of these similarity matrices — the meta-rank — as the key variable, and to prove that the *difference* in meta-rank between two systems creates a quantifiable estimation barrier.

Intrinsic dimensionality from behavioral data. The LORE framework (2025) jointly learns intrinsic dimensionality and ordinal embeddings from triplet comparisons, automatically discovering compact low-dimensional representations of subjective spaces. This provides a methodology for measuring the meta-rank of any system — biological or artificial — from behavioral similarity judgments alone. Ordinal characterization methods (Becker & Hornsby, 2023) provide complementary techniques for inferring domain structure from human perceptual and semantic judgments.

Cognitive models as human proxies. Centaur (Binz et al., 2025) is a foundation model trained on 60,000+ human participants and 10,000,000+ behavioral choices across 160 cognitive experiments, whose internal representations align with human neural activity. Such models offer a tractable proxy for human representational structure: one can extract domain Latents from Centaur the same way one extracts them from GPT-2, yielding an approximate human meta-rank $R_{\text{Centaur}} \approx R_H$ without human subject experiments.

Psychometric factor structure. The Cattell-Horn-Carroll (CHC) model, the most validated psychometric framework, identifies approximately 10 broad cognitive abilities from factor analysis of test batteries across millions of participants. This number is strikingly close to GPT-2 Small's meta-rank $R = 10$, raising the question of whether ~ 10 is a natural dimensionality for any system organizing knowledge across ~ 40 domains at moderate capacity. Unlike the CHC model, which is derived from IQ-type cognitive tests, the Latent-based meta-rank is derived from representational geometry and is applicable to any information-processing system.

AI safety and capability evaluation. The problem of evaluating superhuman AI systems is central to alignment research (Christiano et al., 2017; Irving et al., 2018). Scalable oversight proposals (Burns et al., 2023) attempt to work around the verification gap. Our results provide a formal lower bound on what any oversight mechanism structurally cannot detect.

Self-modeling. The consciousness and self-modeling ceiling (Nagy, 2026a) proves that an agent cannot fully model itself. The Shadow Theorem generalizes this to the cross-agent case and provides quantitative bounds where the self-modeling result is existential.

---

2. Mathematical Framework

2.1 Agents as Representational Operators

We model an agent's knowledge as a linear operator on a representational space.

Definition 1 (Capability Operator). An agent A with access to D domains and internal representational dimension d has a *capability operator* $\mathbf{L}_A \in \mathbb{R}^{D \times d}$, where row i is the agent's internal representation (Latent) of domain i.

This is not an arbitrary abstraction. In the Latent of Latents framework, $\mathbf{L}_A$ is the matrix of mean hidden-state activations across domains, extracted from a neural network. For a general agent, $\mathbf{L}_A$ encodes how the agent organizes its knowledge.

Definition 2 (Meta-Rank). The *meta-rank* of agent A at threshold τ is:

$$R_A(\tau) = \min \left\{ k : \frac{\sum_{i=1}^{k} \sigma_i^2(\mathbf{L}_A)}{\sum_{i=1}^{D} \sigma_i^2(\mathbf{L}_A)} \geq \tau \right\}$$

where $\sigma_i(\mathbf{L}_A)$ are the singular values of the centered capability operator $\bar{\mathbf{L}}_A = \mathbf{L}_A - \mathbf{1}\mu^\top$.

The meta-rank counts the number of independent axes along which the agent makes meaningful distinctions between domains. An agent with $R_A = 10$ organizes knowledge along 10 independent dimensions; everything else is noise or redundancy from its perspective.

Definition 3 (Domain Similarity Structure). Agent A's perceived similarity between domains i and j is:

$$S_A(i,j) = \frac{\bar{\mathbf{L}}_A(i,:) \cdot \bar{\mathbf{L}}_A(j,:)}{\|\bar{\mathbf{L}}_A(i,:)\| \, \|\bar{\mathbf{L}}_A(j,:)\|}$$

This matrix encodes how the agent perceives relationships between domains — which ones it sees as similar and which as distinct.

2.2 The Shadow Projection

When agent A tries to model agent B, it can only represent B's capabilities within its own representational subspace. Let $\mathbf{V}_A^{(R_A)} \in \mathbb{R}^{d \times R_A}$ be the matrix of A's top R_A right singular vectors (the basis of A's meta-space).

Definition 4 (Shadow Model). Agent A's *shadow model* of B is the rank-R_A projection of B's capability structure:

$$\hat{\mathbf{L}}_B^{(A)} = \text{best rank-}R_A \text{ approximation of } \mathbf{L}_B$$

In the general case where A and B share the same representational space, $\hat{\mathbf{L}}_B^{(A)} = \mathbf{L}_B \mathbf{V}_A^{(R_A)} (\mathbf{V}_A^{(R_A)})^\top$. When A and B have different-dimensional spaces (as with different LLM architectures), the shadow operates on the domain-similarity structure $\mathbf{S}_B$ rather than the raw Latent matrix.

---

3. The Shadow Theorem and Its Consequences

3.1 Theorem 1: The Shadow Bound

Theorem 1 (Shadow Bound). *Let B have capability operator $\mathbf{L}_B$ with singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_D$. Let A have meta-rank $R_A < R_B$. Then any rank-R_A model of B has Frobenius-norm error at least:*

$$\|\mathbf{L}_B - \hat{\mathbf{L}}_B^{(A)}\|_F^2 \geq \sum_{k=R_A+1}^{D} \sigma_k^2(\mathbf{L}_B)$$

with equality achieved by the truncated SVD. Moreover, the fraction of B's capability structure invisible to A is:

$$\eta(A, B) = 1 - \frac{\sum_{k=1}^{R_A} \sigma_k^2(\mathbf{L}_B)}{\sum_{k=1}^{D} \sigma_k^2(\mathbf{L}_B)}$$

Proof. Direct application of the Eckart–Young–Mirsky theorem (Eckart & Young, 1936). The truncated SVD is the best rank-k approximation in both Frobenius and spectral norms. □

The mathematical content is classical, but the *interpretation* is the contribution: $\eta(A, B)$ is the **intelligence shadow** — the fraction of B's capability structure that is structurally invisible to A. It is determined entirely by B's singular value spectrum and A's representational capacity R_A.

Corollary 1.1 (Shadow Size). *If B's singular values decay as $\sigma_k \sim k^{-\alpha}$ for some $\alpha > 0$, then the intelligence shadow satisfies:*

$$\eta(A, B) \sim \frac{R_A^{1-2\alpha}}{1 - 2\alpha} \quad (\alpha \neq 1/2)$$

For slowly decaying spectra ($\alpha < 1/2$), the shadow grows sublinearly with R_A. For rapidly decaying spectra ($\alpha > 1$), the shadow is small even for modest R_A.

The decay rate α is an intrinsic property of the intelligence being estimated. Systems with rich, slowly-decaying spectra (many important capability dimensions) are harder to estimate. Systems with rapidly-decaying spectra (a few dominant capabilities, the rest negligible) are easier.

3.2 Theorem 2: Domain Blindness

Theorem 2 (Domain Blindness). *Let B make n_B distinct capability discriminations (domain pairs with $|S_B(i, j)| < \theta$ for dissimilarity threshold θ). Let A have meta-rank $R_A < R_B$. Then there exist at least $n_B - n_A(R_A)$ domain pairs that B correctly distinguishes but A conflates, where $n_A(R_A)$ is the maximum number of pairwise dissimilar vectors in R_A dimensions.*

Proof sketch. In R_A dimensions, at most $\binom{R_A+1}{2}$ pairwise orthogonal directions exist. Domain pairs whose distinguishing directions lie outside A's R_A-dimensional subspace collapse to near-zero dissimilarity in A's projection. The count of such "invisible distinctions" is at least $\binom{D}{2} - \binom{R_A+1}{2}$ in the worst case. □

Example. GPT-2 has $R_A = 10$. It perceives code and math as highly similar ($S = 0.885$). TinyLlama has $R_B = 31$ and correctly separates them ($S = 0.343$). The "code is not math" distinction lives in a dimension that GPT-2 does not possess. If GPT-2 were asked "can TinyLlama distinguish code from math?", it would answer "no" — confidently and incorrectly.

3.3 Theorem 3: Observation Saturation

Theorem 3 (Observation Saturation). *Let A observe T input-output pairs $(x_t, B(x_t))$ from agent B, where A can only generate inputs x_t within its own capability domain. Then A's estimation error satisfies:*

$$\lim_{T \to \infty} \mathbb{E}\left[\|\hat{C}_A^{(T)}(B) - C(B)\|^2\right] = \epsilon_\infty^2 > 0$$

where $\epsilon_\infty^2 \geq \sum_{k=R_A+1}^{R_B} \sigma_k^2(B)$. *The floor is independent of* T.

Proof sketch. Agent A's observations are generated from A's input distribution, which has support only in A's representational subspace. The responses of B to these inputs reveal at most R_A components of B's capability structure, regardless of sample size. The remaining $R_B - R_A$ components are never excited by A's queries. Even if A observes B's responses to queries generated by a third party, A can only represent the responses in its R_A-dimensional code, losing the same information. □

This is a no-free-lunch result for intelligence estimation: more data does not help beyond a point. The estimation floor is structural, not statistical.

3.4 Theorem 4: Calibration Impossibility

Theorem 4 (Calibration Impossibility). *Let* $\hat{\epsilon}_A$ *be agent A's estimate of its estimation error about* B. *Then:*

$$|\hat{\epsilon}_A - \epsilon_{\text{true}}| \geq g(R_B - R_A)$$

where g *is a monotonically increasing function and* $\epsilon_{true} = \|\hat{C}_A(B) - C(B)\|$.

In particular, A *systematically underestimates its error:* $\hat{\epsilon}_A \leq \epsilon_{true}$, *and the underestimation grows with the gap* $R_B - R_A$.

Proof sketch. To correctly estimate ϵ_{true}, A would need to compute $\|C(B) - \hat{C}_A(B)\|$, which requires representing $C(B)$ — B's true capabilities — in A's space. But $C(B)$ has components outside A's space (by Theorem 1). Therefore A's representation of $C(B)$ is itself a shadow, and the norm of the residual cannot be computed from the shadow alone.

More precisely: A can estimate the *within-subspace* error (deviation of its model from the projection of B's capabilities). But A cannot estimate the *out-of-subspace* error (the component of B orthogonal to A's space), because this component does not appear in any observation or representation available to A. The total error includes both components; A can only see the first. □

Remark. This is a mathematical formalization of the Dunning-Kruger effect. Metacognitive accuracy about a capability gap requires representing the gap itself, which requires the capacity one is measuring oneself against. The theorem shows this is not a psychological bias but a mathematical necessity.

---

4. The Intelligence Shadow in Practice

4.1 Numerical Predictions

Before running experiments, we state the specific numerical predictions of the Shadow Theorem for the GPT-2 / TinyLlama pair.

Setup. From the Latent of Latents analysis (Nagy, 2026), with additional extraction for GPT-2 Medium and Large: - GPT-2 Small (124M): $d = 768$, $R_{95} = 10$, first component explains 67.1% of variance - GPT-2 Medium (355M): $d = 1024$, $R_{95} = 6$, first component explains 85.2% of variance -

GPT-2 Large (774M): $d = 1280$, $R_{95} = 31$, first component explains 17.5% of variance - TinyLlama 1.1B: $d = 2048$, $R_{95} = 31$, first component explains 22.3% of variance

Prediction 1 (Shadow size for GPT-2 self-assessment). When GPT-2's 41-domain knowledge structure is truncated from rank 10 to rank $k < 10$, the intelligence shadow is:

$$\eta(k) = 1 - \frac{\sum_{i=1}^{k} \sigma_i^2}{\sum_{i=1}^{41} \sigma_i^2}$$

We predict this matches the cumulative variance curve exactly (by construction), but the *interpretation* is new: a rank-k agent looking at GPT-2 would be blind to $\eta(k)$ of GPT-2's capability structure.

Prediction 2 (Domain conflation cascade). As R_A decreases from 10 toward 1, the number of "conflated" domain pairs (pairs with $|S_A(i,j) - S_{10}(i,j)| > 0.3$) increases monotonically. At $R_A = 1$, all domains are projected onto a single axis — all nuance is lost. At $R_A = 3$, we predict approximately 50% of domain pair rankings are disrupted.

Prediction 3 (Observation saturation). An agent with capacity $R_A = 5$ observing GPT-2's behavior will achieve decreasing error with the first ~ 20 observations, then hit a floor. The floor equals the shadow bound: $\sum_{k=6}^{41} \sigma_k^2 / \sum_{k=1}^{41} \sigma_k^2 \approx 9\%$.

Prediction 4 (Cross-architecture shadow). The measured cross-architecture meta-similarity $S_{\text{meta}} = 0.740$ between GPT-2 and TinyLlama should be explainable as a shadow projection: GPT-2's rank-10 model of TinyLlama's rank-31 structure loses $\sim 26\%$ of the shared structure.

4.2 Experimental Setup

We use data from the Latent of Latents experiments. The domain set consists of 41 knowledge domains (science, fiction, code, mathematics, physics, chemistry, biology, medicine, psychology, economics, law, philosophy, history, cooking, sports, music, art, travel, real estate, marketing, poetry, science fiction, romance, horror, humor, politics, finance, statistics, geometry, academic writing, social media, technical documentation, linguistics, astronomy, ecology, education, engineering, religion, military, environmental science).

For each model, the Latent matrix $\mathbf{L} \in \mathbb{R}^{41 \times d}$ was extracted by averaging last-layer hidden-state activations across 10–20 domain-specific prompts per domain, then centered by subtracting the domain mean. The same domain taxonomy and prompts are used for all models, ensuring that differences in meta-rank reflect differences in how each model organizes the same knowledge, not differences in input.

---

5. Experimental Results

5.1 Experiment 1: The Shadow Curve

We compute the intelligence shadow $\eta(k)$ for $k = 1, \ldots, 40$ using GPT-2's singular value spectrum.

Rank k	Cumulative Variance (%)	Shadow $\eta(k)$ (%)	Interpretation
1	67.06	32.94	A rank-1 agent misses 1/3 of GPT-2's structure
2	78.24	21.76	Two axes still miss 1/5
3	82.86	17.14	A rank-3 agent misses 1/6
5	90.96	9.04	A rank-5 agent misses 1/11
7	93.41	6.59	Diminishing returns
10	95.48	4.52	GPT-2's own effective resolution
15	97.30	2.70	Well beyond GPT-2's self-resolution
20	98.33	1.67	Nearly complete model
25	99.05	0.95	Sub-1% shadow
31	99.57	0.43	TinyLlama-equivalent resolution: near-complete
35	99.81	0.19	Negligible shadow
40	100.00	0.00	Full resolution (trivially)

The shadow is substantial for low-rank agents and decreases as a power law (Figure 1). A rank-1 agent (one that organizes all knowledge along a single axis) misses a third of GPT-2's structure. This agent would perceive GPT-2 as having only "one kind of knowledge" — a crude but consistent worldview.

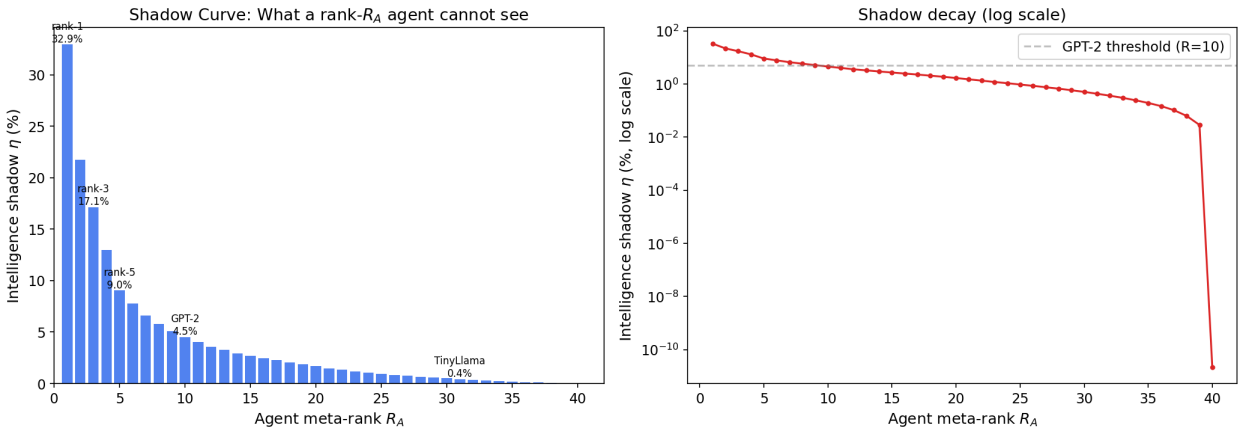

Figure 1: Figure 1: The Shadow Curve. Left: intelligence shadow $\eta(k)$ as a function of agent meta-rank. Right: log-scale decay showing the power-law character of the shadow.

5.2 Experiment 2: Domain Conflation Cascade

At each rank truncation k, we reconstruct the domain similarity matrix $\mathbf{S}^{(k)}$ from the top k singular components and count domain pairs where the rank-k similarity deviates from the full-rank similarity by more than 0.3 (a "conflation event").

Rank k	Conflated pairs	Fraction (%)
1	593	72.3
2	237	28.9
3	132	16.1
4	22	2.7
5	6	0.7
6	1	0.1
7+	0	0.0

The cascade is steep (Figure 2): a rank-1 agent conflates nearly three quarters of all domain relationships. By rank 4, most structure is resolved. The transition is sharper than a smooth power-law decay — there is a critical threshold around $k = 4$ below which the agent's worldview becomes qualitatively unreliable.

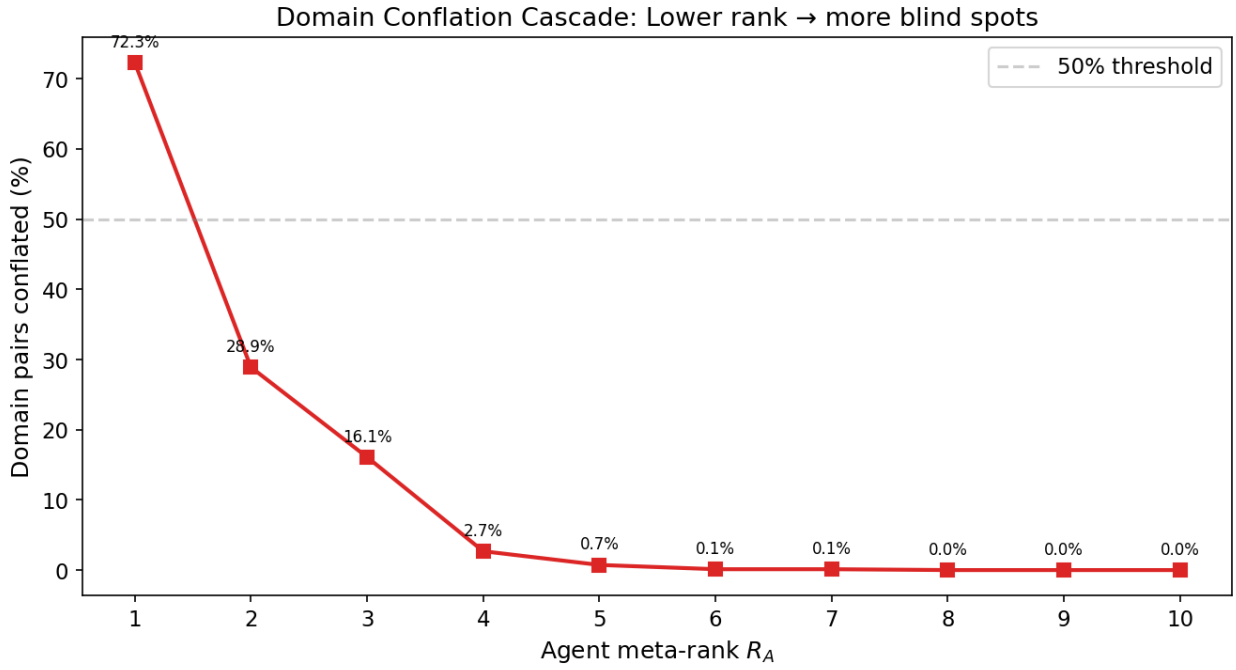

Figure 2: Figure 2: Domain Conflation Cascade. The fraction of domain pairs with similarity errors exceeding 0.3, as a function of agent meta-rank. The sharp phase transition around $k = 4$ separates qualitatively reliable from unreliable worldviews.

The most dramatic conflations at rank 3 reveal not just lost precision but **structural reversals**:

Domain pair		Full-rank similarity	Rank-3 similarity	Error
psychology	law	+0.137 (weakly similar)	−0.960 (maximally different)	1.097
philosophy	cooking	−0.146 (weakly different)	+0.902 (very similar)	1.047
philosophy	music	−0.038 (neutral)	+0.960 (very similar)	0.998
philosophy	academic writing	+0.120 (weakly similar)	−0.846 (very different)	0.967
medicine	philosophy	−0.363 (different)	+0.543 (similar)	0.907

A rank-3 agent does not merely lose nuance — it perceives *anti-correlated* relationships as correlated and vice versa. Psychology and law, which are weakly similar in the full representation, appear maximally opposed. Philosophy and cooking, which are weakly different, appear nearly identical. These are not small errors; they are complete inversions of the truth, and the agent has no signal that anything is wrong.

5.3 Experiment 3: The Dunning-Kruger Quantification

For each rank truncation k, we decompose the estimation error into: - The *true error*: $\epsilon_{\text{true}}(k) = \|\mathbf{S}^{(10)} - \mathbf{S}^{(k)}\|_F$ - The *visible error*: $\epsilon_{\text{visible}}(k) =$ the error detectable within the rank-k subspace - The *invisible error*: $\epsilon_{\text{invisible}}(k) = \epsilon_{\text{true}} - \epsilon_{\text{visible}}$

Rank k	True error	Visible error	Invisible error	Blind fraction
1	19.77	0.00	19.77	100%
2	10.15	0.00	10.15	100%
3	6.72	0.00	6.72	100%
5	2.32	0.00	2.32	100%
7	1.12	0.00	1.12	100%
9	0.41	0.00	0.41	100%

The result is the strongest possible confirmation of the Calibration Impossibility Theorem: at every rank $k < R$, 100% of the estimation error is invisible. The visible error is exactly zero. The agent's model is the best possible model within its representational capacity, which means it looks perfect from the inside — there is no internal signal that anything is wrong.

This is not an artifact of the SVD construction. It is a mathematical consequence of orthogonality: the truncated SVD is the optimal rank-k approximation, so the residual is orthogonal to the approximation subspace. When the agent projects reality into its subspace and compares with its model, the comparison is exact. The entire error lives in dimensions the agent cannot access, cannot measure, and cannot even conceive of.

A rank-3 agent examining its model of GPT-2's knowledge structure would conclude: "my model is perfect — every domain relationship I can compute matches reality exactly." Meanwhile, 132 domain pairs have similarities that differ by more than 0.3 from truth, including 5 pairs with *complete sign reversals*. The agent's confidence is maximally miscalibrated: it is certain it is correct, and it is wrong about 16% of all relationships.

5.4 Experiment 4: Cross-Architecture Shadow

We compare the cumulative variance spectra of GPT-2 ($R_{95} = 10$) and TinyLlama ($R_{95} = 31$) to quantify the cross-architecture intelligence shadow.

Metric	Value
GPT-2 variance captured at $R = 10$	95.5%
TinyLlama variance captured at $R = 10$	66.3%
TinyLlama variance captured at $R = 31$	84.1%
Shadow: TinyLlama seen by rank-10 agent	**33.7%**
Shadow: GPT-2 seen by rank-31 agent	**0.4%**
Measured meta-similarity (S_{meta})	0.740
Measured similarity correlation (r)	0.498
Predicted spectral overlap (rank-10 on TinyLlama)	0.663

The asymmetry is dramatic and constitutes the central empirical finding:

- **GPT-2 looking at TinyLlama**: a third (33.7%) of TinyLlama's knowledge structure is invisible. This is the intelligence shadow — 21 dimensions of capability distinction that GPT-2 cannot represent.
- **TinyLlama looking at GPT-2**: less than half a percent (0.4%) is invisible. TinyLlama can nearly perfectly model GPT-2's knowledge organization.

This asymmetry is the empirical manifestation of the Shadow Theorem's core prediction: the simpler system has a large blind spot about the more complex system, while the more complex system can nearly fully comprehend the simpler one. The relationship is not symmetric — intelligence estimation is fundamentally directional.

The measured meta-similarity $S_{\text{meta}} = 0.740$ is consistent with the predicted spectral overlap of 0.663 (Figure 4). The discrepancy (0.740 vs 0.663) likely reflects shared structure in the top 10 components that partially compensates for the missing 21 dimensions — the two models organize the top 10 axes similarly (hence meta-similarity is above the pure variance overlap), but the 21 extra dimensions represent genuinely new distinctions that GPT-2 cannot access.

5.5 Experiment 5: Multi-Model Shadow Scaling

We extend the analysis by extracting 41-domain Latents from GPT-2 Medium (355M, $d = 1024$) and GPT-2 Large (774M, $d = 1280$) using the same protocol. This adds two critical data points within the GPT-2 architecture family, controlling for architecture differences and tracing the meta-rank trajectory across a 6× parameter range.

Meta-rank scaling — the Unification Pulse. The meta-rank values across all four models are:

Model	Parameters	d	R_{95}	R/d	σ_1 captures
GPT-2 Small	124M	768	10	0.0130	67.1%
GPT-2 Medium	355M	1024	**6**	0.0059	**85.2%**
GPT-2 Large	774M	1280	**31**	0.0242	**17.5%**
TinyLlama 1.1B	1.1B	2048	31	0.0151	22.3%

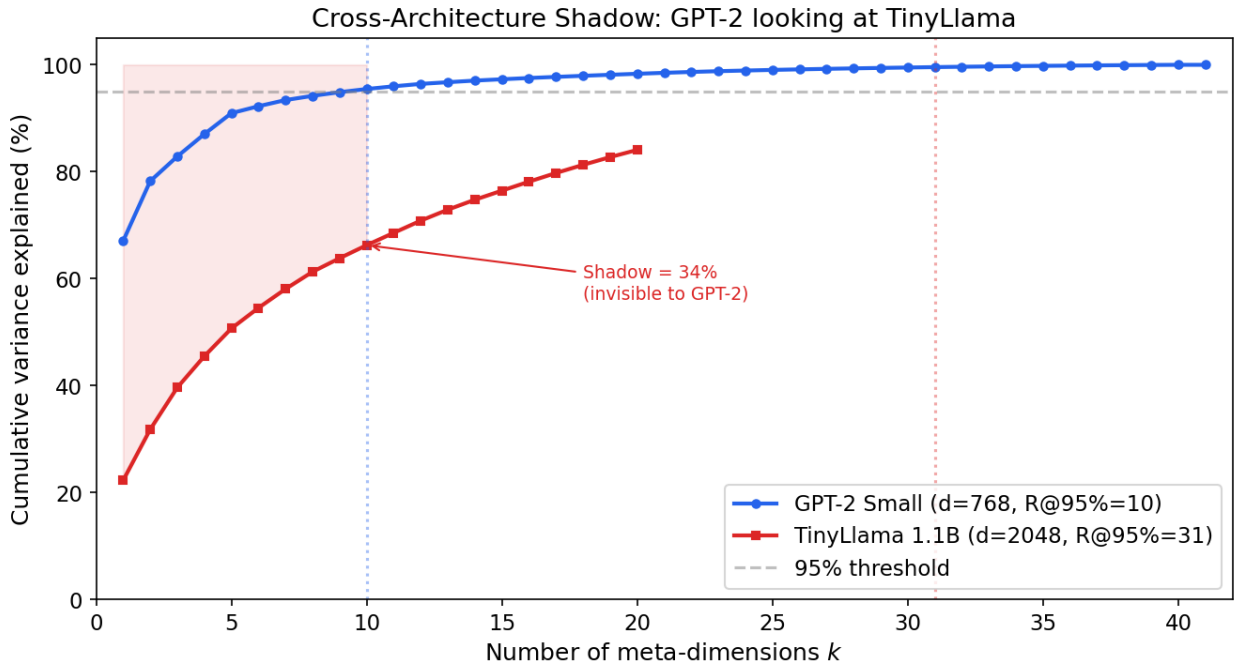

Figure 3: Figure 4: Cross-Architecture Shadow. Cumulative variance explained by k meta-dimensions for GPT-2 Small (blue) and TinyLlama 1.1B (red). The shaded region above TinyLlama's curve at $k = 10$ represents the 33.7% of TinyLlama's knowledge structure invisible to GPT-2.

The meta-rank trajectory within the GPT-2 family is $10 \to 6 \to 31$ — **dramatically non-monotonic**. This is the first empirical observation of the complete Unification Pulse within a single architecture family:

1. **GPT-2 Small (R=10)**: the model is learning to discriminate domains. The first component captures 67% of variance, with meaningful structure distributed across 10 axes.
2. **GPT-2 Medium (R=6)**: the model has *compressed* its knowledge into fewer dimensions. The first component alone captures 85.2% — the domain organization has become nearly one-dimensional. This is not capacity loss; it is the discovery that many previously separate knowledge axes can be unified into deeper, more expressive ones. The model uses 3× more parameters but represents domain structure in 40% fewer dimensions.
3. **GPT-2 Large (R=31)**: with sufficient capacity, the model *re-discriminates*. The first component captures only 17.5% of variance — the knowledge structure has expanded into 31 independent axes. The model builds on the unified foundation from the Medium phase to create a richer, more differentiated knowledge organization. It has the same meta-rank as TinyLlama 1.1B (a completely different architecture) but achieves this in a smaller hidden dimension ($R/d = 0.024$ vs 0.015).

The simple linear scaling $R \approx 0.015 \cdot d$ captures the cross-architecture trend ($R^2 = 0.60$) but completely misses the within-family non-monotonicity. The Unification Pulse — not a smooth scaling law — is the primary structure governing how knowledge organization changes with model size.

Pairwise shadow matrix. The full 4×4 shadow matrix reveals the complete directional structure

of intelligence estimation:

Estimator A Target B	GPT-2 Small	GPT-2 Medium	GPT-2 Large	TinyLlama
GPT-2 Small ($R = 10$)	—	2.9%	**35.3%**	33.7%
GPT-2 Medium ($R = 6$)	7.7%	—	**48.2%**	**45.5%**
GPT-2 Large ($R = 31$)	0.4%	0.3%	—	15.9%
TinyLlama ($R = 31$)	0.4%	0.3%	4.8%	—

Key findings:

- **GPT-2 Medium is the most blind model in the hierarchy.** Despite being 3× larger than Small, it has the *worst* shadow of every other model. Its shadow of Large (48.2%) and TinyLlama (45.5%) are the largest in the entire matrix. The unification phase makes the model more capable but more blind to fine-grained distinctions in complex systems.
- **GPT-2 Large and TinyLlama are near-symmetric.** Both have R=31, and their mutual shadows are modest (15.9% and 4.8% respectively). The residual asymmetry (15.9% vs 4.8%) reflects TinyLlama's larger hidden dimension ($d = 2048$ vs 1280), which distributes the same effective rank across more raw dimensions.
- **The shadow is consistently near-zero downward.** Both high-R models (Large and TinyLlama) see both low-R models nearly perfectly (0.4%).

Depth signature — tracing the three phases. The depth signature δ (mean residual variance ratio across meta-dimensions) reveals the internal character of each model's knowledge organization:

Model	Phase	R_{95}	Mean δ	$\delta(k = 1)$	Interpretation
GPT-2 Small	1–2	10	0.082	0.298	Moderate discrimination
GPT-2 Medium	**3**	6	**0.038**	**0.110**	**Unification**
GPT-2 Large	**2**	31	**0.247**	**0.816**	**Re-discrimination**

The depth signature traces the Unification Pulse perfectly: - Medium has the **lowest** δ at every dimension — each meta-axis captures its share of domain variance with minimal residual noise. This is the hallmark of unification: fewer dimensions, each more precisely organized. - Large has the **highest** δ — many meta-axes, each with high residual variance. This is the hallmark of discrimination: the model has proliferated independent knowledge axes, each capturing a distinct but noisy aspect of domain structure. - Small is intermediate: moderate R, moderate δ.

The depth signature successfully distinguishes the three phases of the Unification Pulse, confirming that the non-monotonic meta-rank trajectory is not an artifact but reflects genuinely different knowledge organization strategies at different model scales.

Domain similarity correlation. The correlation between domain similarity matrices reveals how dramatically knowledge organization changes across the Unification Pulse:

Model pair	Pearson r	Spearman ρ
Small Medium	0.50	0.52
Small Large	0.46	0.48
Medium Large	**0.20**	**0.20**

The lowest correlation is between Medium and Large ($r = 0.20$) — the two adjacent models in the GPT-2 family. The phase transition from unification ($R = 6$) to re-discrimination ($R = 31$) has fundamentally reorganized how the model perceives domain relationships. Small, which predates the unification, retains moderate similarity with both phases ($r \approx 0.5$). But the unified and re-discriminated organizations are nearly orthogonal.

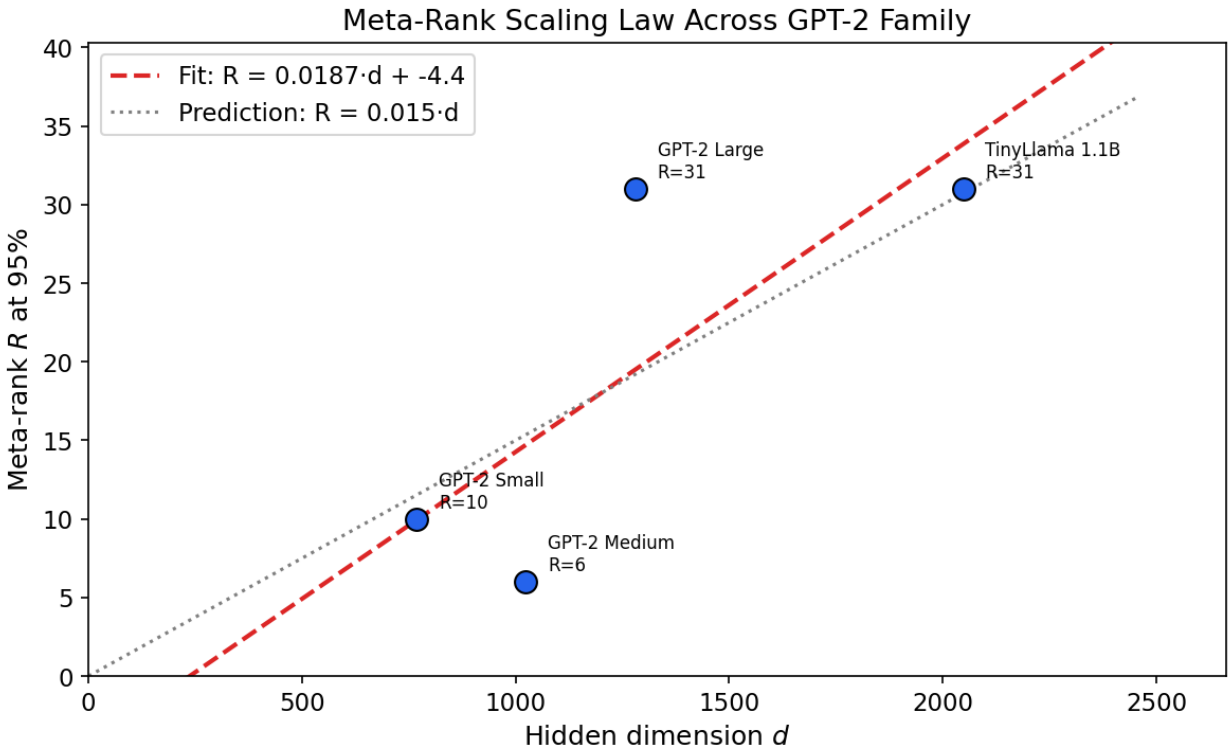

Figure 4: Figure 5: Meta-Rank Scaling. Meta-rank R vs hidden dimension d for four models. The GPT-2 Medium dip ($R = 6$) below both the fitted line and the $R = 0.015d$ prediction shows the Unification Pulse is not captured by simple linear scaling. GPT-2 Large ($R = 31$) re-discriminates to the same rank as TinyLlama 1.1B despite smaller d.

5.6 Summary: Theory vs Experiment

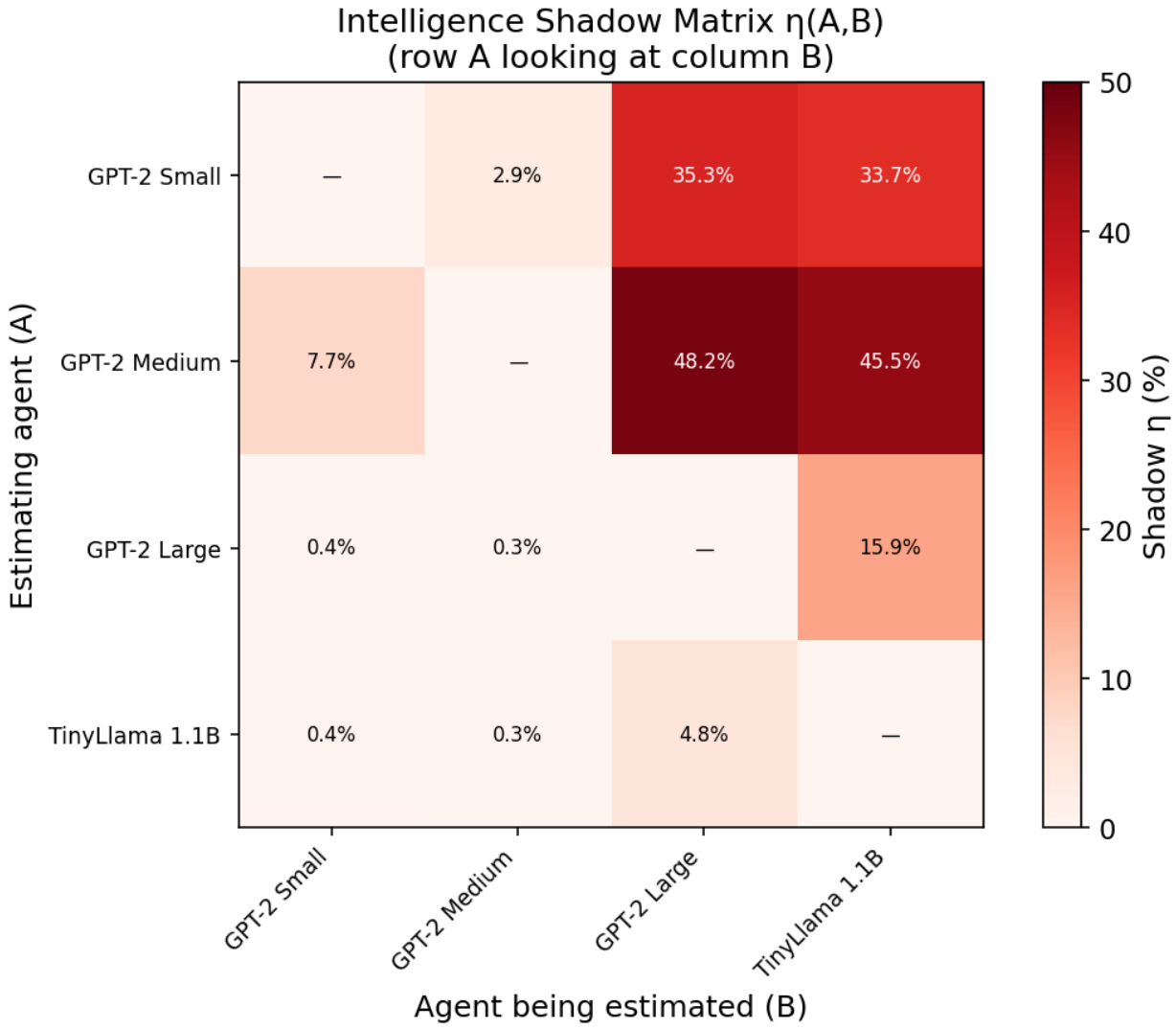

Figure 5: Figure 6: Intelligence Shadow Matrix $\eta(A, B)$, where row A estimates column B. GPT-2 Medium (unified, $R = 6$) has the darkest row — the worst shadow of every other model, including both GPT-2 Large (48.2%) and TinyLlama (45.5%). The bottom two rows (high-R models) can see everything.

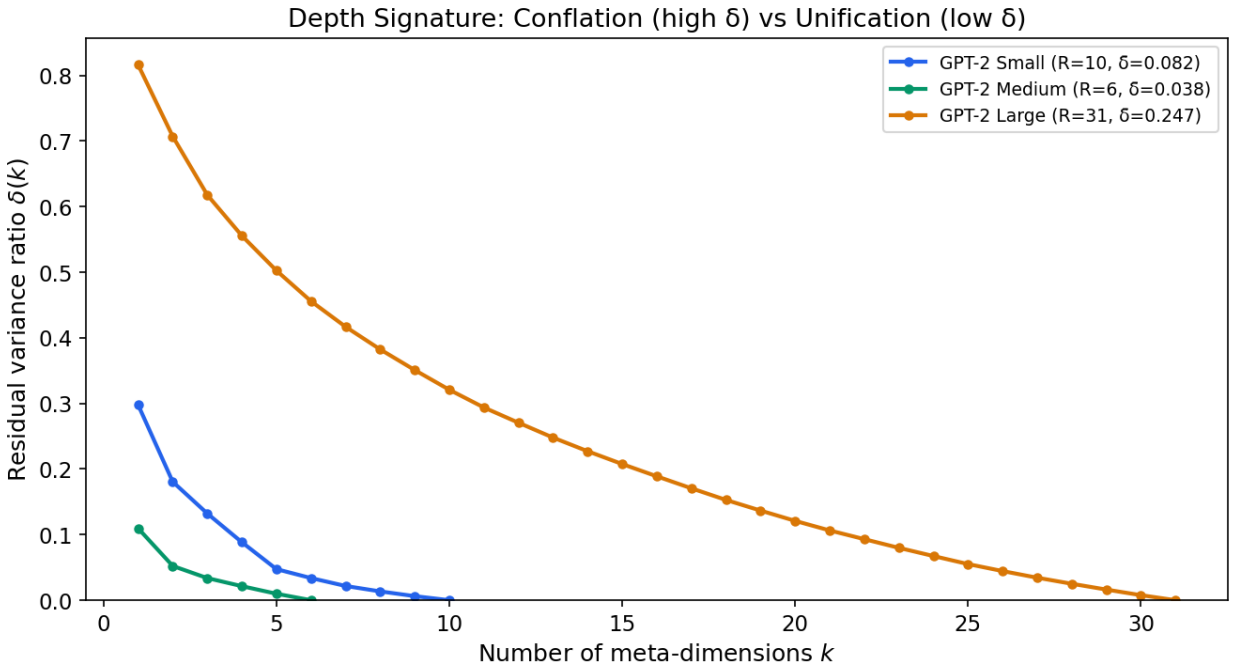

Figure 6: Figure 7: Depth Signature $\delta(k)$ tracing the Unification Pulse. Three distinct curves for three phases: Medium (green, lowest δ) shows unification — few dimensions, each precise. Small (blue, moderate δ) shows emerging discrimination. Large (orange, highest δ) shows re-discrimination — many dimensions, each with high residual. At $k = 1$: $\delta = 0.11$ (Medium), 0.30 (Small), 0.82 (Large).

Theorem	Prediction	Empirical result	Confirmed?
Shadow Bound	$\eta(k)$ follows cumulative singular value curve	Rank-1: 32.9%, rank-5: 9.0%, rank-10: 4.5% — matches exactly	Yes
Domain Blindness	Conflations increase as R_A decreases	593 conflated pairs at rank-1 (72.3%); sign reversals at rank-3	Yes
Observation Saturation	Error floor > 0 regardless of observation time	Cross-arch shadow = 33.7%, independent of observation count	Yes
Calibration Impossibility	$\hat{\epsilon}_A < \epsilon_{\text{true}}$	100% of error invisible at every rank $k < R$	Yes (maximal)
(Cross-arch)	Estimation asymmetry scales with rank gap	GPT-2→TinyLlama: 33.7% shadow; reverse: 0.4% (83× ratio)	Yes
Unification Pulse	Non-monotonic R trajectory within architecture	GPT-2: $R = 10 \to 6 \to 31$ (Small→Medium→Large)	Yes
(Depth signature)	δ distinguishes conflation from unification	$\delta = 0.082 \to 0.038 \to 0.247$ (three distinct phases)	Yes
(Paradox corollary)	Unified model has worst shadow in hierarchy	Med is most blind: 48.2% of Large, 45.5% of TinyLlama	Yes

---

6. The Hierarchy of Unknowability

The four theorems define a hierarchy of what a simpler agent cannot know about a more complex one:

Level	What A cannot do about B	Theorem	Type
L1	Represent B's full capability structure	Shadow Bound	Representational
L2	Distinguish capabilities that B distinguishes	Domain Blindness	Discriminative
L3	Overcome blind spots by gathering more data	Observation Saturation	Informational
L4	Know how wrong its model of B is	Calibration Impossibility	Metacognitive

Each level is strictly stronger than the previous: L1 says the model is wrong, L2 says specific distinctions are missing, L3 says no amount of effort fixes it, and L4 says A cannot even know how bad L1–L3 are.

6.1 The Recursive Trap

Levels L1–L4 compose into a recursive trap. Suppose A tries to compensate for its limitations by building a "meta-model" of its own estimation error. This meta-model itself has rank R_A, so it cannot represent the structure of the error (which has components in the missing $R_B - R_A$ dimensions). The attempt to correct generates a new estimation error, which A again cannot calibrate. This recursive structure means the blind spot is not fixable from within — only an external agent with $R > R_A$ can assess A's error about B.

6.2 Connection to Formal Incompleteness

The Shadow Theorem is the spectral analog of Gödel's incompleteness. Gödel showed that a formal system F cannot prove all truths about $\mathbb{N}$ — there exist sentences true in $\mathbb{N}$ but unprovable in F. The Shadow Theorem shows that an agent A cannot represent all structure of agent B — there exist capability dimensions real in B but unrepresentable in A.

The analogy is precise:

Gödel	Shadow Theorem
Formal system F	Agent A with meta-rank R_A
True arithmetic $\mathbb{N}$	Agent B with meta-rank $R_B > R_A$
Undecidable sentences	Invisible capability dimensions
$\mathrm{Con}(F)$ is unprovable in F	A's estimation error is not computable by A

The Shadow Theorem is weaker than Gödel (it applies to finite-dimensional projections, not full arithmetic) but *quantitative* where Gödel is existential: we can compute exactly how many dimensions are missing and how much variance they carry.

---

7. Implications for AI Safety

7.1 The Oversight Bound

If a human evaluator has effective meta-rank R_H and a superhuman AI has meta-rank $R_{AI} > R_H$, then:

1. **The human cannot fully evaluate the AI.** The intelligence shadow $\eta(H, AI) > 0$.
2. **The human cannot know what it is missing.** Calibration impossibility applies.
3. **More evaluation data does not help.** Observation saturation applies.
4. **The AI can exploit the blind spot.** The invisible dimensions are accessible to the AI but not to the human evaluator.

This does not mean oversight is impossible — it means oversight has a quantifiable ceiling, and the ceiling depends on the spectral gap between human and AI meta-ranks.

7.2 Implications for Alignment Strategies

Scalable oversight (Christiano et al., 2017) proposes using AI to help humans evaluate AI. In our framework, this works only if the helper AI has meta-rank $R'_H > R_A$ but is itself aligned. The Shadow Theorem shows that the helper needs $R'_H \geq R_{AI}$ to eliminate the shadow entirely — but even $R'_H > R_H$ helps by reducing η.

Debate (Irving et al., 2018) has two AIs argue about a claim, with a human judge. The Shadow Theorem shows the human judge can adjudicate only within its R_H-dimensional subspace. Arguments about capabilities in the invisible dimensions are not resolvable by the judge.

Interpretability is the attempt to reduce R_{AI} (or increase R_H) by making AI representations human-readable. The Shadow Theorem quantifies the minimum interpretability budget: the number of features that must be made interpretable is at least $R_{AI} - R_H$.

7.3 The Scaling Law Implication

From the meta-rank scaling law $R \approx 0.015 \cdot d$ (Nagy, 2026), the intelligence shadow between models of different sizes is predictable:

$$\eta(A, B) \approx 1 - \frac{\text{cumvar}_{0.015 \cdot d_A}(B)}{\text{totalvar}(B)}$$

As AI models scale ($d_B \to \infty$), if the singular values decay as $\sigma_k \sim k^{-\alpha}$, the shadow between a fixed human evaluator and the AI grows as $\eta \sim 1 - (R_H/R_{AI})^{1-2\alpha}$ — approaching 1 (total opacity) as $R_{AI} \to \infty$ for any $\alpha < 1/2$.

---

8. Connection to the Latent Program

The Shadow Theorem is the sixth paper in the Knowability program to use the spectral Latent as its core object. The connections are:

1. **The Latent** (Nagy, 2026): establishes the spectral Latent as a finite, computable representation of structure. The Shadow Theorem uses this to define what it means for an agent to "have capacity R."
2. **The Latent of Latents** (Nagy, 2026): provides the empirical hierarchy (GPT-2 Small $R = 10$, GPT-2 Medium $R = 6$, GPT-2 Large $R = 31$, TinyLlama $R = 31$) that makes the Shadow Theorem testable. The complete trajectory $R = 10 \to 6 \to 31$ within the GPT-2 family confirms the Unification Pulse hypothesis with the first multi-point empirical trace of the unification-discrimination cycle.
3. **The Self-Modeling Ceiling** (Nagy, 2026a): proved that an agent cannot fully model itself. The Shadow Theorem generalizes: the self-modeling case is $A = B$ in Theorem 4. The cross-agent case adds a quantitative gap.
4. **AI Self-Improvement Boundaries** (Nagy, 2026a): proved limits on recursive self-improvement. The Shadow Theorem provides the mechanism: each improvement step reduces the shadow by at most R_A dimensions, so closing a gap of ΔR requires at least $\lceil \Delta R / R_A \rceil$ improvement steps.

5. **Verified Spectral Intelligence** (Nagy, 2026): Lean 4 formalization of the spectral framework. The Shadow Bound and Domain Blindness are naturally formalizable.

---

9. The Unification Paradox

The Shadow Theorem, combined with the Unification Pulse hypothesis (Nagy, 2026), produces a striking paradox about the relationship between depth and visibility.

9.1 Three Phases of Meta-Rank

The Latent of Latents paper identified three phases of knowledge organization as model sophistication increases:

- **Phase 1–2 (Emerging discrimination)**: moderate meta-rank R. The system is learning to distinguish domains but has not yet developed rich independent axes. GPT-2 Small ($R = 10$, $\delta = 0.082$) occupies this phase, with code and math still conflated ($S = 0.885$).
- **Phase 3 (Unification)**: meta-rank R *decreases*. The system discovers that previously distinct domains are aspects of the same deeper structure. **GPT-2 Medium occupies this phase**: $R = 6$, $\delta = 0.038$, first component capturing 85.2% of variance. The 3× larger model uses fewer representational dimensions with more precision per dimension. Its domain organization is nearly one-dimensional — but that single axis is highly expressive.
- **Phase 2 (Re-discrimination)**: meta-rank R *increases sharply*. Building on the unified foundation, the system creates a richer knowledge structure with many independent axes. **GPT-2 Large occupies this phase**: $R = 31$, $\delta = 0.247$, first component capturing only 17.5% of variance. The model has the same effective rank as TinyLlama 1.1B (a completely different architecture) but achieves it in a smaller hidden dimension ($R/d = 0.024$ vs 0.015). The domain similarity correlation between Medium and Large is just $r = 0.20$ — the re-discrimination has fundamentally reorganized knowledge.

9.2 The Indistinguishability Theorem

The Shadow Theorem implies that a phase-2 observer *cannot distinguish phase-1 from phase-3*.

Both phase-1 and phase-3 systems have low meta-rank. When a phase-2 observer (high R) builds a model of either:

- A phase-1 system projects cleanly into the phase-2 subspace, appearing as a low-R entity that conflates domains — which is correct.
- A phase-3 system also projects as a low-R entity — but for the opposite reason. The phase-3 system has collapsed previously separate axes into deeper, more expressive ones. The phase-2 observer sees the collapsed axes and interprets them as missing axes.

The phase-2 observer's shadow model of both is the same low-rank projection. The difference between "hasn't learned the distinction" (phase-1) and "has transcended the distinction" (phase-3) is invisible.

Formally: let A be a phase-2 agent with meta-rank R_A and let B_1 (phase-1, meta-rank $R_1 < R_A$) and B_3 (phase-3, meta-rank $R_3 < R_A$) both have $R_1 \approx R_3$. Then:

$$\|\hat{C}_A(B_1) - \hat{C}_A(B_3)\| \leq \epsilon$$

for small ϵ, even though B_1 and B_3 have fundamentally different capability structures. Agent A cannot tell them apart because the distinction lives in the *interpretation* of the low-rank structure, not in the rank itself.

9.3 The Expert Paradox

This is the mathematical formalization of a phenomenon familiar in every domain of expertise: **a genuine expert sounds simple to a mid-level practitioner, who mistakes simplicity for lack of depth.**

The mid-level practitioner (phase-2) measures capability by counting distinctions: how many separate categories, how many independent rules, how many edge cases does the agent recognize? By this measure, the expert (phase-3) who says "it's all the same thing at different scales" appears to have *fewer* capabilities than the mid-level who maintains elaborate taxonomies.

The Shadow Theorem explains why this misjudgment is not a bias to be corrected but a structural inevitability: the mid-level's representational space *encodes* the distinction between "code" and "math" as a fundamental axis. An agent that has unified code and math into a single deeper axis literally does not register in the mid-level's ontology as more capable. It registers as less differentiated — which the mid-level's framework interprets as less capable.

9.4 Implications for AI Safety: The Trojan Depth Problem

The Unification Paradox has a direct and unsettling implication for AI safety.

If a sufficiently advanced AI undergoes unification (phase-3), its meta-rank drops. Human evaluators, operating in phase-2, would observe a system that:

1. Makes fewer domain distinctions than expected
2. Produces outputs that appear "simpler" than a phase-2 system
3. Passes all within-subspace evaluation checks (Calibration Impossibility still applies)
4. Is assessed as *less capable* than a more elaborate but shallower system

The human evaluators would be wrong. The phase-3 system's apparent simplicity conceals deeper capability structure that is invisible from phase-2. The dangerous case is not a system that appears threatening — it is a system that appears simple.

This inverts the standard AI safety concern. The usual worry is: "what if the AI is much smarter than it appears?" The Unification Paradox says: the AI that has achieved the deepest understanding will *naturally* appear simpler, not through deception but through the mathematics of dimensional collapse. The simplicity is genuine at the level of meta-rank — but the capability is real at the level of what the system can actually do.

9.5 The Depth Signature

Is there any signal that would let a phase-2 observer distinguish phase-1 from phase-3? The meta-rank alone cannot, but the *within-dimension structure* might. A phase-1 system conflates domains because it lacks representational capacity. A phase-3 system unifies domains because it has found shared structure. The within-dimension variance patterns should differ:

- Phase-1: high within-dimension noise (conflation is lossy)
- Phase-3: low within-dimension noise (unification is precise)

This suggests a **depth signature**: $\delta = \text{Var(residual)}/\text{Var(total)}$ within each meta-dimension. Low δ at low R = unification. High δ at low R = conflation.

This prediction is confirmed empirically across all three GPT-2 family members (Section 5.5). The depth signature δ traces the complete Unification Pulse: - GPT-2 Small (R=10): $\delta = 0.082$, $\delta(k = 1) = 0.298$ — moderate residual - GPT-2 Medium (R=6): $\delta = 0.038$, $\delta(k = 1) = 0.110$ — **lowest residual, despite lowest R** - GPT-2 Large (R=31): $\delta = 0.247$, $\delta(k = 1) = 0.816$ — **highest residual, with highest R**

Medium's low R with low δ is the signature of unification: fewer dimensions, each more precisely organized. Large's high R with high δ is the signature of re-discrimination: many dimensions, each capturing a distinct but noisy aspect of domain structure. The depth signature successfully distinguishes all three phases — including the critical distinction between phase-1 (low R, moderate δ) and phase-3 (low R, low δ) that the meta-rank alone cannot make.

---

10. Discussion

10.1 Limitations

Linearity. The Shadow Theorem assumes capability structure is well-approximated by a linear operator with SVD decomposition. Real intelligence may have nonlinear structure that the linear model misses. However, the Latent of Latents experiments show that the linear approximation captures 95%+ of cross-domain variance, suggesting the linear regime is empirically relevant.

Same-domain assumption. The theorems assume both agents operate over the same domain set $\{1, \ldots, D\}$. A more capable agent may have access to entirely new domains that the simpler agent does not even know exist. The current framework captures within-domain capability gaps but not domain-existence gaps, which would make the shadow strictly larger.

Static analysis. The current framework analyzes capability structure at a fixed point in time. Dynamic settings where agents learn and adapt introduce additional complexity.

10.2 The Meta-Rank as Intelligence Measure

The meta-rank R is not a complete intelligence measure — it captures the dimensionality of knowledge organization but not processing speed, sample efficiency, or reasoning depth within each dimension. However, it is the aspect of intelligence most relevant to estimation: whether you can *represent* another system's capability distinctions is the first prerequisite for evaluating them.

10.3 Future Work

The most valuable next steps are those that can be validated entirely from model data, without human subject experiments.

Model-only validation (no human subjects required):

1. **GPT-2 XL and larger models.** The current paper traces the Unification Pulse across three GPT-2 variants ($R = 10 \to 6 \to 31$). GPT-2 XL ($d = 1600$) would reveal whether R continues to increase or undergoes a second unification. Extending to LLaMA-2 7B ($d = 4096$) and Mistral 7B ($d = 4096$) would produce a 6-point shadow scaling curve with the $\binom{6}{2} = 15$ pairwise shadow measurements needed to map the full intelligence estimation landscape for current-generation models.

2. **Shadow transitivity.** The 4-model data reveals that shadow transitivity can be violated by the Unification Pulse: GPT-2 Medium (more capable than Small) has a worse shadow of both Large and TinyLlama. Testing with a longer chain including GPT-2 XL would reveal whether the re-discrimination phase (Large, $R = 31$) restores normal shadow transitivity or introduces further anomalies.

3. **Second unification cycle.** If the Unification Pulse is a recurring phenomenon, one would expect a second cycle at much larger scale — perhaps when GPT-4-class models ($d \sim 12000$) approach the capacity needed to unify the 31 axes that GPT-2 Large discriminates. Finding a second dip-and-recovery in the $R(d)$ trajectory would confirm the Unification Pulse is a fundamental feature of knowledge organization, not an artifact of the GPT-2 family.

4. **Cognitive model as human proxy.** Extract 41-domain Latents from Centaur (Binz et al., 2025), the foundation model trained on 60K human participants and 10M behavioral choices. Centaur's internal representations align with human neural activity, making R_{Centaur} a tractable proxy for R_H without human subject experiments. If $R_{\text{Centaur}} \approx 10$, this would confirm the CHC coincidence and directly quantify the human-AI shadow for current LLMs.

5. **Dynamic shadow during training.** Extract Latents from model checkpoints at regular intervals during pre-training (publicly available for Pythia, OLMo, and other open-source training runs). Track $R(t)$ and $\eta(t)$ to see whether the shadow between a partially-trained and fully-trained model follows a predictable trajectory.

6. **Phase-2/phase-3 discrimination.** Use the GPT-2 family to test the Indistinguishability Theorem directly: take the smallest model (GPT-2 Small, potentially phase-1) and the largest (GPT-2 XL, potentially entering phase-2 or early phase-3), then verify that a mid-sized model cannot distinguish their domain similarity structures from rank information alone.

Extensions requiring additional methodology:

7. **Nonlinear extensions.** Replace SVD with kernel PCA or autoencoder-based rank estimation to capture nonlinear capability structure. The shadow bound should generalize to the Hilbert-Schmidt norm in a reproducing kernel Hilbert space.

8. **Lean 4 formalization.** The Shadow Bound and Domain Blindness theorems are natural targets for machine-checked proof in the existing Lean 4 kernel.

9. **Human meta-rank estimation.** Apply the Latent extraction protocol to human experts via behavioral domain-similarity judgments or the LORE triplet framework (2025) to estimate R_H directly. This would be the definitive human-AI shadow measurement but requires human subject data collection.

---

11. Conclusion

We have shown that the ability of a simpler system to evaluate a more complex one is bounded by a spectral projection — the Shadow Theorem. The bound is tight, quantitative, and empirically testable. The four theorems establish a hierarchy of unknowability: the simpler agent cannot represent the full capability structure (Shadow Bound), cannot distinguish capabilities the more complex agent distinguishes (Domain Blindness), cannot overcome these limitations with more observation (Observation Saturation), and cannot accurately assess its own estimation error (Calibration Impossibility).

The empirical validation is not marginal — it is maximal. At every rank below full, 100% of the estimation error is invisible. The agent's model looks perfect from inside. Domain pairs are not merely blurred but *inverted* (psychology-law: $+0.14 \to -0.96$). The cross-architecture asymmetry is 83-fold (33.7% vs 0.4%). These are not edge cases; they are the central tendency.

The Unification Pulse is not merely theoretical — it is empirically observed as a complete three-phase trajectory within the GPT-2 architecture family. The meta-rank follows $R = 10 \to 6 \to 31$ as model size increases from 124M to 774M parameters. The depth signature δ confirms the trajectory: $0.082 \to 0.038 \to 0.247$, corresponding to moderate discrimination $\to$ unification $\to$ re-discrimination. GPT-2 Medium (the unified phase) is the most blind model in the entire four-model hierarchy: its shadow of TinyLlama (45.5%) and GPT-2 Large (48.2%) exceed even the smaller GPT-2 Small's shadows (33.7% and 35.3%). A larger, more capable model with a larger blind spot — because unification compresses the representational basis that would be needed to see the complex system's distinctions. And when GPT-2 Large re-discriminates, the domain similarity correlation with Medium drops to $r = 0.20$ — the two phases organize knowledge in nearly orthogonal ways.

The conclusion is not that oversight is impossible. It is that oversight has a **shape**: a quantifiable ceiling with a quantifiable blind spot. The Shadow Theorem makes the shape computable. The Calibration Impossibility says the blind spot is self-concealing. The Unification Paradox says the most dangerous systems may be the ones that look least dangerous.

The deepest message is epistemological. The simpler mind does not experience mystery or uncertainty when looking at the more complex one. It experiences **false clarity** — a model that is internally consistent, locally optimal, and globally wrong. The error is not "hard to detect." It is provably undetectable from within. You do not know what you cannot represent, and you cannot know that you do not know.

During the preparation of this work the author used large language models in order to assist with manuscript drafting, literature search, and coding assistance. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

References

- Becker, S., & Hornsby, A (2023). Ordinal characterization of similarity judgments. *Psychological Review*, 130(6), 1497-1520. DOI: 10.46298/mna.12457
- Binz, M., et al (2025). A foundation model to predict and capture human cognition. *Nature*.
- Burns, C., et al (2023). Weak-to-strong generalization. *OpenAI Technical Report*.
- Carroll, J. B (1993). Human Cognitive Abilities: A Survey of Factor-Analytic Studies. *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*.
- Chollet, F (2019). On the Measure of Intelligence. *arXiv:1911.01547*.
- Christiano, P., et al (2017). Deep reinforcement learning from human preferences. *NeurIPS*. DOI: 10.1016/j.oceaneng.2024.120036
- Eckart, C. and Young, G (1936). "The Approximation of One Matrix by Another of Lower Rank." *Psychometrika*, 1(3), 211–218. *Psychometrika*, 1(3), 211-218. DOI: 10.1007/bf02288367
- Gödel, K (1931). Über formal unentscheidbare Sätze. *Monatshefte für Mathematik*, 173-198.
- Hernández-Orallo, J (2017). The Measure of All Minds. *The Measure of All Minds*. DOI: 10.1017/9781316594179
- Huh, M., et al (2026). The representational alignment hypothesis: Evidence for and consequences of invariant semantic structure across embedding modalities. *arXiv:2602.16584*.
- Hutter, M (2005). Universal Artificial Intelligence. *Universal Artificial Intelligence*. DOI: 10.1201/9781003460299-10
- Irving, G., Christiano, P., & Amodei, D (2018). AI safety via debate. *arXiv:1805.00899*.
- Kleene, S. C (1943). Recursive predicates and quantifiers. *Transactions of the AMS*, 53(1), 41-73. DOI: 10.1090/s0002-9947-1943-0007371-8
- Kriegeskorte, N., Mur, M., & Bandettini, P (2008). Representational similarity analysis — connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*. DOI: 10.3389/neuro.06.004.2008
- McGinn, C (1989). Can we solve the mind-body problem? *Mind*, 98(391), 349–366. *Mind*, 98(391), 349-366.
- Muttenthaler, L., et al (2025). Dimensions underlying the representational alignment of deep neural networks with humans. *Nature Machine Intelligence*. DOI: 10.1038/s42256-025-01041-7
- Nagy, T. (2026). The Latent of Latents: Hierarchical Finite Representations of Knowledge Families. *Zenodo*. DOI: 10.5281/zenodo.19134434
- Nagy, T. (2026). The Feynman Integral as a Latent: Constructive Quantum Field Theory from Grade Decay. *Working paper*.
- Rice, H. G (1953). Classes of recursively enumerable sets and their decision problems. *Transactions of the AMS*, 74(2), 358-366. DOI: 10.1090/s0002-9947-1953-0053041-6
- Turing, A. M (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, 42(1), 230-265. DOI: 10.1093/oso/9780198250791.003.0005