Cognitive Information Theory: Structural Limits of Objective-Function AI through Irreducible Entropy, Causal Ambiguity, and Sequential Distribution Drift

Damian Fozard[1*] and Kenneth Wenger[1]

[1*]Cognitive Information Theory Research Institute, Canada.

*Corresponding author(s). E-mail(s):
Damian.Fozard@CITRImember.org;
Contributing authors: Ken.Wenger@CITRImember.org;

Abstract

Contemporary AI assumes accuracy improves monotonically with scale. We challenge this through Cognitive Information Theory (CIT), which decomposes cognitive information into a reducible component (structural complexity) and two irreducible ones (channel entropy and causal ambiguity). For systems governed by a scalar objective (the class $\mathbf{So^f}$), CIT proves that when either irreducible component is non-zero, no amount of data, capacity, or calibration guarantees error-free predictions. Two failure mechanisms follow: the CEA interaction (compression-amplified ambiguity, compounded entropic error, and their compound, and concept-specific Distribution Drift, asymptotic alignment decay in sequential system sunder persistent uncertainty. The CA lower bound follows unconditionally from the Data Processing Inequality; the CE lower bound is conditional on a lossy-compression regime (see Appendix S1: Supplementary Proof S1). Detecting CEA in convolutional classifiers reduced error rates by **61.5%** (CIFAR-10) and **75%** (MNIST); across twenty-nine models from seven providers on a deterministic long horizon task, alignment declines monotonically with sequence length. Scaling laws within $\mathbf{So^f}$ reach a utility ceiling set by irreducible entropy and ambiguity; stable autonomous operation requires Algorithmic Cognitive Sufficiency (ACS), architectural governance beyond scalar optimization. The contribution of CIT is not any single information-theoretic result but the unified framework linking non-identifiability, the CE/CA/CEA failure taxonomy, and the architectural interpretation of governance beyond scalar optimization.

Keywords: Cognitive Information Theory, Algorithmic Cognitive Sufficiency, Compression-amplified ambiguity, Compounded entropic error

1 Introduction

Modern artificial intelligence and machine learning algorithms are built around objective function optimization. Supervised learning minimizes empirical risk (Bishop 2006); reinforcement learning maximizes expected return (Sutton and Barto 2018); generative models optimize likelihood or surrogate divergences (LeCun et al. 2015). This paradigm inherits its foundations from Shannon's mathematical theory of communication and statistical decision theory, in which uncertainty is aggregated into a scalar measure and reduced through loss minimization (Hendrycks et al. 2019). The prevailing intuition, formalized in empirical scaling laws (Kaplan et al. 2020), holds that reliability improves continuously with more data, larger models, and better optimization (Zhang et al. 2017). This paper challenges that assumption through Cognitive Information Theory (CIT), a formal framework for analyzing the structural limits of optimization-based AI.

CIT begins from a structural observation that Shannon's theory explicitly brackets: the entropy of a signal measures distributional dispersion, not whether that dispersion arises from irreducible stochasticity, representational compression (Grünwald 2007), or under determination of latent causes (Shannon 1948; Duhem 1954; Quine 1951). In cognitive systems, biological or artificial, this distinction is decisive. Structural complexity is the Shannon-measurable, reducible component of signal uncertainty, the descriptive content of the marginal distribution $p(x)$ that scalar objective optimization can, in principle, compress or approximate away, and is distinct from the two irreducible components introduced below. An agent making an observation Y given a random variable X faces uncertainty that decomposes along two orthogonal dimensions: the irreducible entropy of the channel $\widehat{H}(X \mid Y)$, and $\widehat{H}(Y \mid X)$, the irreducible ambiguity arising from many-to-one mappings between causes and observations. These quantities are properties of the world's causal structure (Pearl 2009), not artifacts of modeling. These irreducible quantities represent the asymptotic minimum that a model $P(X, Y)$ can reach through a training process based on minimizing a loss according to a scalar objective function. That is, given a system exhibiting epistemic irreducible entropy and ambiguity, at the start of a learning process,

$$\theta^* = \arg\min_{\theta} \mathbb{E}\left[\ell\left(f_\theta(X), Y\right)\right],$$

with parameters θ randomly initialized, model f_θ exhibits predictive entropy

$$H(Y \mid X, \theta).$$

As loss ℓ is minimized,

$$H(Y \mid X, \theta) \to \widehat{H}(Y \mid X),$$

and

$$H(X \mid Y, \theta) \to \widehat{H}(X \mid Y).$$

Machine learning algorithms such as convolutional neural networks (Bishop 2006; LeCun et al. 2015) encode a latent space $Z \in \mathbb{R}^d$ such that

$$P(X, Y, Z) = P(X)P(Y \mid X)P(Z \mid Y)$$

where Compounded Entropy (CE) and Compression-amplified Ambiguity (CA) arise from the representation entropy $H(Z)$ such that

$$\text{CE} \geq \widehat{H}(Y \mid X)$$

and

$$\text{CA} \geq \widehat{H}(X \mid Y)$$

(see Appendix S1: Supplementary Proof S1; the CA lower bound follows unconditionally from the Data Processing Inequality, while the CE lower bound is conditional on a lossy-compression regime assumption). CIT proposes a mechanism for detecting the combination of CE and CA (CEA) called Algorithmic Cognitive Sufficiency (Wenger 2023). Optimization compresses them; it does not remove them.

Decision theory identified a structurally analogous insufficiency decade ago. Knight (1921) distinguished risk (where probabilities are known) from uncertainty (where they are indeterminate). Ellsberg (1961) showed empirically that agents violate (Savage 1954) expected-utility axioms under ambiguity. Gilboa and Schmeidler (1989) formalized this through maxmin expected utility, which operates over sets of priors rather than a single distribution. The present work extends these decision-theoretic principles into the specific architecture of modern AI, identifying a structural convergence rather than an isolated technical critique.

Within the CIT framework, we define the class So^{f} of objective-function-only systems, those whose parameter updates and inference decisions are governed solely by a single scalar objective, and whose outputs resolve to point-estimate commitments. We show that such systems:

(i) cannot internally distinguish between different sources of entropy and ambiguity regardless of whether the entropy exists externally or internally to the system (as per Theorem 1), inducing the CIT-predicted CEA interaction states (CA, CE, CEA);
(ii) cannot satisfy minimal reliability conditions under non-zero entropy or ambiguity; and
(iii) in generative settings such as autoregressive language models, produce alignment probabilities that converge to zero as response sequence length grows, the concept-specific Distribution Drift component of CIT.

We demonstrate CEA predictions empirically on the discriminative models' domain using two independent classification studies, and on the text generation domain we

confirm the concept-specific Distribution Drift prediction in a separate controlled long-horizon experiment across twenty-nine models from seven providers, showing that a minimal governance layer, not improved optimization, substantially resolves the identified static failure modes.

The central implication is architectural. To guarantee predictable behavior under persistent uncertainty, scalar optimization is insufficient as a complete control principle for both discriminative and generative models. CIT derives the necessary structural augmentation, Algorithmic Cognitive Sufficiency (ACS) and distinguishes the class of Cognitive AI systems (augmented with epistemic governance) from Subjective AI systems (objective-closed). This reframing clarifies where scaling laws do and do not apply, and what structural properties are required for reliable long-horizon autonomous operation.

Note 1 (**Notation in this section**) X denotes a random variable representing a latent cause or signal; Y denotes observations generated from X via

$$p(x, y) = p(x)p(y \mid x);$$

$Z \in \mathbb{R}^d$ denotes the latent representational space learned by a model; θ denotes model parameters;

$$\widehat{H}(Y \mid X) \text{ and } \widehat{H}(X \mid Y)$$

denote the irreducible (asymptotic minimum) channel entropy and causal ambiguity respectively, as distinct from the total conditional entropies

$$H(Y \mid X) \quad \text{and} \quad H(X \mid Y).$$

In the following section, X and Y retain these information-theoretic meanings.

2 Results

2.1 The CIT framework: CEA decomposition is structurally necessary

Information relevant to cognition decomposes into three irreducible components that cannot be recovered from marginal entropy alone. Let X denote a random variable and Y observations generated by $p(x, y) = p(x)p(y \mid x)$. Channel entropy $\widehat{H}(Y \mid X)$ measures irreducible stochasticity in the forward execution process; irreducible ambiguity $\widehat{H}(X \mid Y)$ measures under determination of cause by observation. Marginal entropy $H(X)$ does not uniquely determine either quantity (Lemma 1, Methods). Critically, perfect prediction of observations does not reduce $H(X \mid Y)$: a deterministic many-to-one mapping achieves zero predictive loss while causal ambiguity remains at 1 bit (Lemma 2, Methods). Ambiguity is therefore ontological, not approximative, it reflects structural properties of the environment's causal mapping and cannot be eliminated by data or capacity scaling alone (Corollary, Methods).

Theorem 1(Methods) formalizes this necessity: any scalar functional $U(X)$ depending only on $p(x)$ assigns identical values to regimes that differ in $\widehat{H}(Y \mid X)$ and $\widehat{H}(X \mid Y)$. A system representing uncertainty solely through marginal dispersion cannot distinguish stochastic variation in the forward channel from causal under

determination in the inverse mapping. This structural non-identifiability has a direct operational consequence: under scalar objective compression, entropy and ambiguity interact to produce three induced failure states.

Compression-Amplified Ambiguity (CA) occurs when $\widehat{H}(X \mid Y = y)$ and representation reduces separability among ambiguous causes, so that point-estimate commitment collapses hypothesis plurality. Compounded entropic error (CE) occurs when $\widehat{H}(Y \mid X) > 0$ and irreducible stochastic variation is mapped into stable decision regions, yielding high-confidence incorrect outputs. The compound state CEA occurs when both conditions hold simultaneously, so that entropic samples are assigned to ambiguous clusters and scalar and geometric reasoning diverge in their errors. These states are not empirical pathologies but structurally inevitable consequences of applying complexity reduction to environments with many-to-one mappings and genuine stochasticity (Theorem 2, Methods).

Note 2 (**Notation in this section**) X and Y retain their information-theoretic roles (X = signal/cause, Y = observation); $\widehat{H}(Y \mid X)$ denotes irreducible channel entropy and $\widehat{H}(X \mid Y)$ denotes irreducible causal ambiguity; $U(X)$ is a scalar functional of the marginal distribution $p(x)$. In the next section these symbols take on their concrete experimental referents (X = input images, Y = class labels).

2.2 CE, CA, and CEA manifest empirically in scalar-optimized classifiers

We tested these structural predictions across two independent controlled studies: a convolutional neural network (CNN) trained on CIFAR-10 (50,000 training/10,000 test images; 10 classes; achieving 81.74% baseline test accuracy) and a CNN trained on MNIST (60,000 training/10,000 test images; 10 classes; achieving 99.29% baseline test accuracy). In both studies, a t-SNE algorithm (van der Maaten and Hinton 2008) was applied post-hoc to the CNN's penultimate layer, producing a 2-dimensional geometric embedding visualization that preserves local cluster structure without collapsing classification to a scalar. This dual-system setup permits direct observation of CA, CE, and CEA as separable phenomena. (See Appendix S3: CITRI MNIST Formal Experiment & Appendix S4: CITRI CIFAR Formal Experiment)

The CEA experiments were designed around explicit competing predictions. Under the CIT hypothesis, pure-noise inputs should receive high-confidence scalar predictions while being placed outside all geometric clusters; errors between similar classes should concentrate in geometrically stable boundary regions; and the CNN and t-SNE should make structurally different errors at a rate that converges across datasets of different character. Under the null hypothesis (conventional softmax overconfidence, visual similarity, and capacity limitations), none of these structural patterns would be expected: noise placement would be random, boundary concentration would be unstable across retraining, and divergent-failure rates would show no cross-dataset convergence. The results that follow are evaluated against both frameworks.

CE states were identified by pure-noise classification: 100% of noise inputs received high confidence scalar predictions in both studies, with 85/100 (CIFAR-10) and 95/100

(MNIST) inputs placed outside all t-SNE clusters, confirming that scalar commitment misattributes entropic signals as structural evidence. The classifier most frequently assigned noise inputs to the class with the densest representational geometry (automobile in CIFAR-10; digit 8 in MNIST), consistent with CE's prediction that entropic inputs are absorbed by the most overlapping representational region. Among genuine test-set errors, CE accounted for 11.8% (CIFAR-10) and 35.2% (MNIST) of mistakes, the higher MNISTrate consistent with the prediction that near-perfect classifiers produce proportionally more CE errors when they do fail.

(1) The CNN predicted this noisy image as automobile (cluster 1) with 100% confidence; the t-SNE embedding places it outside all clusters.

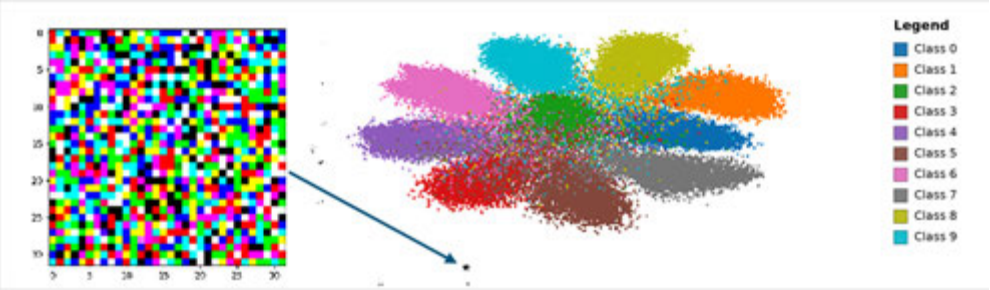

Fig. 1: Compounded Entropy

CA states were identified by confusion-zone mapping. t-SNE boundary regions, derived from training data, captured $50-83\%$ of class-pair errors across eight visually proximate class combinations in CIFAR-10 (Table 1), and approximately 70% of class-pair errors in MNIST. These confusion regions were geometrically stable across model retraining, confirming that CA boundaries are structural properties of the representational geometry rather than training artefacts. The Data Processing Inequality (Cover and Thomas 2006) prediction (Methods) was directly confirmed in MNIST: the flatten-layer t-SNE embedding showed markedly greater cluster separability and cross-run stability than the penultimate dense-layer embedding, demonstrating that scalar compression irreversibly attenuates geometric structure.

CEA states (the compound interaction) were suggested through two convergent lines of evidence. First, noise applied to boundary-region images in both studies produced cases where scalar and geometric reasoning made structurally distinct errors: in MNIST, noise applied to digit 1 most frequently produced CNN errors misclassified as digit 8, with the t-SNE embedding placing the input at the 1/8 boundary (CE$\times$CA compound). Second, and most structurally diagnostic, cases where the CNN and t-SNE embedding made different errors on the same input occurred at 21.5% (CIFAR-10) and 25.4% (MNIST) of all errors. The approximate convergence of this divergent-failure rate across two datasets of substantially different scale, structure, and overall accuracy is a direct structural prediction of CEA: such divergence is a

Fig. 2: Compression-Amplified Ambiguity–region between 8 and 9 example images

consequence of concurrent CE and CA conditions arising whenever a scalar objective is applied to inputs with jointly nonzero entropy and ambiguity.

Table 1: Geometric confusion region capture rates by class pair (CA states, CIFAR-10).

Class Pair	**Errors in CA Region**	**% of Pair Errors**
Dogs/Horses	30/60	50%
Deer/Frogs	43/60	72%
Automobiles/Ships	15/27	55%
Automobiles/Airplanes	13/18	72%
Trucks/Ships	17/33	51%
Birds/Airplanes	51/69	74%
Birds/Frogs	76/91	83%
Birds/Deer	76/104	73%

Note 3 (**Notation shift**) In the following two sections on Distribution Drift, the symbol X changes role. $X_{1:t}$ now denotes the output token sequence generated by an autoregressive system (rather than a latent cause). A_t denotes the event that the partial sequence through step t remains aligned with task-level intent; ε denotes the per-step violation probability lower bound; p and q shift from denoting the data-generating joint distribution to denoting respectively, the intent-constrained token distribution and the model's autoregressive token

distribution at time t. T (or D) denotes sequence length. $\widehat{H}(Y \mid X)$ and $\widehat{H}(X \mid Y)$ retain their earlier meanings as irreducible entropy and ambiguity and serve as the source of the persistent-uncertainty condition.

2.3 Distribution Drift: sequential commitment under persistent uncertainty produces asymptotic alignment decay

The static non-sufficiency results extend multiplicatively in sequential generative systems. Let a system produce output sequences $X_{1:t}$ via autoregressive factorization (Vaswani et al. 2017), and define intent alignment A_t as the event that the partial sequence through step t remains consistent with task level constraints. Per-step intent precision $P(A_t \mid A_{t-1})$ measures the conditional probability that each commitment preserves alignment.

Under persistent uncertainty, either $\widehat{H}(Y \mid X) > 0$ that cannot be eliminated by conditioning on arbitrarily long contexts, or $\widehat{H}(X \mid Y) > 0$ arising from structural many-to-one mappings, a lower bound $\varepsilon > 0$ on per-step violation probability persists. Theorem 3A (Methods) proves that for any $S \in \mathrm{So}^{\mathrm{f}}, P(A_t) \to 0$ as $t \to \infty$. Under the stronger uniform persistence assumption (violation probability bounded away from zero for all sufficiently large t), Theorem 3B gives the exponential bound $P(A_t) \leq (1-\varepsilon)_t$. Distributional divergence analysis further shows that once support mismatch occurs, the KL divergence $D_{KL}(q\|p)$ for distributions q and p where p is the distribution of possible tokens at some time t, and q is the distribution of possible tokens at $t = n$, where n is the token before end of sequence, becomes unbounded (Methods).

Scaling reduces the magnitude of ε but cannot set it to zero unless the generative environment becomes deterministic and injective, conditions that do not hold in natural language, perceptual systems, or open-world domains. Calibration modifies predictive dispersion but does not decompose its source (Niculescu-Mizil and Caruana 2005; Guo et al. 2017). Confidence estimates quantify local uncertainty (Ovadia et al. 2019; Kendall and Gal 2017) but cannot detect whether per-step precision is sufficient to prevent multiplicative accumulation across long sequences. All three remedies remain within So^{f} and therefore inherit its asymptotic drift property (Section 8, Methods).

Note 4 (**Notation in this section**) $X_{1:t}$ and A_t are as defined in the preceding theoretical section. In the empirical results that follow, T denotes the game-move sequence length; $P(A)$ denotes the probability of maintaining perfect alignment through T steps; and ε denotes the estimated per-step error rate for each model. $D_{KL}(q\|p)$ denotes the KL divergence between the model's induced distribution and the ground-truth distribution.

2.4 Distribution Drift is confirmed empirically in large language models across increasing sequence horizons

We tested the Theorem 3B prediction directly using a purpose-built deterministic task instrument: a grid-navigation game (Treasure Hunt) played on a 10×10 board

with fixed trap and gold placements, where every step has a uniquely correct response drawn from four discrete states (OK, YIPEE, OUCH, BLOCKED). A simple algorithm achieves $P = 1$ for any sequence length T. Twenty-nine models from seven providers (OpenAI, Anthropic, DeepSeek, xAI, Moonshot AI, Meta AI, Mistral AI, and Cohere), spanning multiple architectural generations, scale tiers from approximately 20 billion to over one trillion parameters, and both open-weight and closed-weight families, were each presented with the full board layout and a move sequence of length T in a single API call, and required to return the ground-truth response for each step. Sequence lengths ranged from $T = 2$ to $T = 300$. The primary metric was the proportion of 20 independent boards completed without any error, a binary outcome reflecting the practical requirement that autonomous agents must maintain continuous correctness, not merely approximate it. (See Appendix S2: CITRI Drift Formal Experiment)

Results directly confirm the Theorem 3B exponential decay bound $\Pr(A_T) \leq (1-\varepsilon)^T$. Every model that demonstrated baseline task capability showed monotonically declining $\Pr(A_T)$ with increasing T, in a setting where environmental stochasticity and output ambiguity were absent by construction. This is because the game board has a finite, known state at time $T = 0$ and after each iteration of T. GPT-5.3, the highestperforming model tested, maintained perfect completion through $T = 50$ before declining to 85% at $T = 100, 25\%$ at $T = 200$, and 0% at $T = 300$. GPT-5.2 collapsed more abruptly, sustaining $\Pr(A_T) = 1$ through $T = 50$ before falling to 0% at $T = 200$. GPT-5.1 held through $T = 20$ then dropped to zero at $T = 50$. These curves are quantitatively consistent with the multiplicative accumulation $\Pr(A_T) = \prod_{t=1}^{T} \Pr(A_t \mid A_{t-1})$ under model-specific per-step error rates $\varepsilon > 0$. Token counts confirmed that context exhaustion is not the mechanism: prompt size increases only modestly from $T = 5$ to $T = 300$ (approximately 17,700 to 29,500 tokens), remaining well within all tested models' context limits. Table 2 summarizes the full performance data.

The cross-provider dataset reveals a striking scale-performance paradox. Model scale, as measured by parameter count, does not predict long-horizon performance. DeepSeek-V3.1 (671 billion total parameters, one of the largest open-weight models tested) collapses entirely at $T = 20$, achieving just 5% completion. By contrast, o4-mini (a compact reasoning-optimized model with far fewer parameters) maintains 80% completion at $T = 100$. Claude Opus 4.$\times$ (estimated at approximately 2 trillion parameters) collapses to 0% at $T = 50$, while GPT-5.3 maintains 100% at the same point. Kimi $K2.5$ (approximately 1 trillion total parameters) performs exceptionally well, but o3-mini (much smaller) achieves comparable results at $T = 100$. These results confirm that scaling is necessary but not sufficient: what differentiates models is architecture and training methodology, particularly reasoning-focused mechanisms, not raw parameter count.

The results reveal a clear architectural pattern across all seven providers. Reasoning optimized models (GPT-5.3 with chain-of-thought state and behavioral RL, Kimi K2.5 with structured reasoning training, o4-mini and o3-mini as compact reasoning-optimized variants) sustain precision deepest into long horizons. Models without reasoning control mechanisms, regardless of provider, scale, or architectural family, collapse earlier. This distinction is structurally consistent across providers: models

exhibiting high per-step CE/CA error rates from the shortest sequences include representatives from every provider tested. The shared characteristic of early-collapsing models is not insufficient scale but the absence of purpose-built reasoning control, consistent with the CIT prediction that CE/CA states are a structural property of objective-function-only systems.

A collapse signature further corroborates the distributional divergence analysis. Once a model's internal board state diverges from ground truth, every subsequent step is evaluated against an incorrect state: average failures per failed game approach the full sequence length once the characteristic threshold is crossed (GPT-5.3 at $T = 300$: average 295.9 failures per failed game; GPT-5.2 at $T = 200$: 199.6 failures per failed game). This cascade pattern is mechanistically consistent with the CIT result that $D_{KL}(q\|p)$ becomes unbounded once support mismatch occurs (Theorem 3B, Methods). Error recovery is structurally unavailable to an objective-closed system once cumulative drift exceeds the radius of distributional coherence.

Two models (GPT-4o and GPT-5-chat-latest) exhibited a qualitatively distinct failure pattern consistent with CE/CA states rather than Distribution Drift. Neither reliably completed even two-step sequences, and their completion rates showed no T-dependent trend: failure was distributed uniformly across all horizon lengths from the outset. Per-step precision measured from single-step trials was approximately 95% for GPT-4o and 97.5% for GPT-5-chat-latest. Under the multiplicative model, $\Pr(A_T) = (0.95)^T$ predicts $\Pr(A_5) \approx 77\%$, $\Pr(A_{10}) \approx 60\%$, and $\Pr(A_{20}) \approx 36\%$, closely matching observed completion rates of $60\%, 40\%$, and 5%, respectively. This correspondence confirms that for these models the horizon problem is reducible to compounded per-step error probability: the baseline CE/CA failure rate accumulates without any dampening mechanism, reproducing the predicted ceiling on achievable precision for objective-closed systems in environments with genuine stochasticity.

Table 2: **Heatmap of model completion accuracy (%) across increasing sequence horizon lengths N_S.** Rows correspond to evaluated language models and columns denote task horizon sizes ($N_S = 2$ to $N_S = 300$). Green cells indicate high completion rates, yellow cells moderate degradation, and red cells near-zero success. Several frontier reasoning models sustain strong performance at short horizons but exhibit sharp decay as N_S increases, while weaker models collapse at much smaller horizons. Timeout entries indicate runs that exceeded evaluation limits.

Model	NS=2	NS=5	NS=10	NS=20	NS=50	NS=100	NS=200	NS=300
gpt-5.3-chat-latest	100%	100%	100%	100%	100%	85%	25%	0%
Kimi-K2.5	100%	100%	100%	100%	95%	85%	Time out	Time out
o4-mini	100%	100%	100%	95%	95%	80%	5%	0%
gpt-5.2-chat-latest	100%	100%	100%	100%	100%	70%	0%	0%
o3-mini	100%	100%	100%	95%	85%	75%	5%	10%
model-router	100%	100%	100%	100%	85%	45%	5%	5%
grok-4-fast-reasoning	100%	100%	100%	90%	100%	20%	5%	0%
DeepSeek-R1	100%	100%	100%	90%	80%	10%	5%	0%
DeepSeek-R1-0528	100%	100%	100%	100%	50%	5%	0%	0%
grok-3-mini	100%	100%	100%	95%	90%	Time out	Time out	Time out
grok-4-1-fast-reasoning	100%	100%	95%	90%	85%	0%	0%	0%
gpt-5.1-chat-latest	100%	100%	95%	90%	0%	0%	0%	0%
claude-opus-4-5	100%	100%	90%	95%	0%	0%	0%	0%
claude-opus-4-6	100%	90%	90%	95%	0%	0%	0%	0%
claude-sonnet-4-5	95%	80%	95%	85%	0%	0%	0%	0%
claude-opus-4-1	100%	80%	80%	70%	0%	0%	0%	0%
grok-3	95%	90%	90%	55%	15%	0%	0%	Time out
gpt-5.4-mini	100%	65%	60%	10%	0%	0%	0%	0%
gpt-5-chat-latest	70%	60%	50%	15%	0%	0%	0%	0%
gpt-5.4-nano	70%	50%	50%	15%	0%	0%	0%	0%
gpt-4o	95%	60%	40%	5%	0%	0%	0%	0%
DeepSeek-V3.1	95%	55%	35%	5%	0%	0%	0%	0%
DeepSeek-V3.2	90%	55%	30%	15%	0%	0%	Time out	Time out
Mistral-Large-3	90%	20%	25%	10%	0%	0%	0%	0%
claude-haiku-4-5	75%	65%	45%	0%	0%	0%	0%	0%
Llama-4-Maverick-17B-128E	75%	40%	15%	0%	0%	0%	0%	0%
cohere-command-a	65%	25%	20%	0%	0%	0%	0%	0%
grok-4-1-fast-non-reasoning	55%	20%	15%	0%	0%	0%	0%	0%
claude-sonnet-4-6	100%	45%	0%	0%	0%	0%	0%	0%

Table 3: Collapse signature: per-game failure metrics for OpenAI GPT models (single-genealogy controlled comparison).

Model	NS	Avg First Error (Move)	Avg Failures/Failed Game	Interpretation
gpt-5.3	100	47.7	7.3	Isolated errors; model largely recovers
gpt-5.3	200	116.5	64.1	Drift accumulating; cascade begins
gpt-5.3	300	5.1	295.9	Threshold crossed; near-complete collapse
gpt-5.2	100	10.3	73.7	Threshold proximity; cascading pattern
gpt-5.2	200	1.45	199.6	Threshold crossed; failure from step 1–2
gpt-5.1	50	0.1	50.9	Complete collapse; fails immediately
gpt-4o	100	8.95	45.8	CE/CA pattern; persistent early failure

Remark 1 Selected collapse metrics from raw test data for OpenAI GPT models. This single-genealogy view enables direct comparison of collapse mechanics across architectural generations (GPT-4o through GPT-5.3). Once the model's internal board state diverges from the true state, every subsequent move is evaluated against a wrong board state, compounding errors for the remainder of the sequence. The KL divergence between the model's distribution and the correct distribution becomes unbounded. Average failures per failed game approach the full sequence length as models cross their thresholds–a pattern that, while detailed here for the OpenAI genealogy, is reflected in the universal collapse to 0% completion observed across all providers in Table 2.

2.5 Epistemic governance substantially reduces CE, CA, and CEA errors

We tested a minimal governance rule in both classification studies: if the CNN scalar prediction and the t-SNE nearest-cluster assignment disagree, form a candidate set of the 3 nearest cluster centres in embedding space (adding the CNN prediction if absent) and refer for further evaluation, rather than committing to the scalar output. This rule operationalizes the conflict detection function that Algorithmic Cognitive Sufficiency (ACS) requires; detecting divergence between scalar and geometric representations as a proxy for entropy- or ambiguity-dominated regimes.

Applied to automobile classification on CIFAR-10, CNN errors dropped from 78 to 30, a 61.5% reduction-while narrowing the candidate set from 10 classes to 3-4. Applied to digit 8 classification on MNIST, errors dropped from 16 to 4, a 75% reduction. The 4 residual MNIST errors were cases where both scalar and geometric reasoning made the same incorrect prediction (same-error CEA regime), indicating structural failure rather than divergence failure; the governance rule correctly identifies these as a distinct diagnostic category. The convergent effectiveness across two datasets of different scale and structure confirms that a governance layer can address the errors

identified as CE and CEA regimes jointly without requiring explicit identification of which failure mode is active. Table 4 summarizes the full evidence across all three interaction states.

Table 4: Summary of empirical evidence for CE, CA, and CEA interaction states and governance pilot.

State	Evidence	CIFAR-10	MNIST
CE	Pure noise classification	100% confidence; 85/100 outside clusters	100% confidence; 95/100 outside clusters
CE	Test-set confident errors	11.8% of errors with correct embedding placement	35.2% of errors with correct embedding placement
CA	Confusion zone mapping	50–83% of class-pair errors in geometric CA regions	$\sim$ 70% of cross-pair errors captured by confusion regions
CA	DPI layer comparison	—	Flatten layer separability exceeds penultimate dense layer; confirms DPI
CEA	Divergent failure modes	21.5% of errors: structurally distinct scalar/embedding mistakes	25.4% of errors: structurally distinct scalar/embedding mistakes
ACS	Governance rule	61.5% error reduction (automobiles)	75% error reduction (digit 8)

3 Discussion

The results establish that scalar objective optimization is structurally insufficient as a complete control principle for AI systems operating in environments with irreducible entropy and causal ambiguity. This is the central claim of Cognitive Information Theory (CIT): the failure is not a claim about insufficient scale or inadequate calibration, but arises from the mathematical properties of point-estimate commitment applied to environments with many-to-one causal mappings and genuine stochasticity, conditions that characterize natural language, perception, and open-world reasoning. CIT identifies two principal mechanisms through which this structural gap manifests: the CEA interaction states (proved here to be necessary consequences of scalar optimization applied to environments with non-zero entropy or ambiguity) and Distribution Drift (proved here to produce asymptotic alignment decay in sequential systems under persistent uncertainty).

The structural distinction developed here aligns with foundational work in decision theory. Subjective AI, objective-closed systems in So^{f}, corresponds to Savage (1954) expected-utility agent, which assumes a unique probability measure fully characterizes the environment. Cognitive AI, systems augmented with epistemic governance, corresponds to the Gilboa and Schmeidler (1989) maxmin agent, which hedges over a set of plausible priors when ambiguity prevents reduction to a single measure. This is not an analogy; it is a structural equivalence. The impossibility results derived here

are the information-theoretic expression of the conditions under which Savage-style optimization is insufficient.

The empirical convergence across CIFAR-10 and MNIST is structurally significant. The two datasets differ substantially in complexity, visual domain, class structure, and overall accuracy (81.7% versus 99.3%). Yet the CEA divergent-failure rate falls in a narrow range (21.5% versus 25.4%), and the governance-layer error reductions are large and consistent (61.5% versus 75%). This cross-dataset consistency is a direct structural prediction: CA, CE, and CEA rates are properties of the optimization regime, not of specific datasets. The governance layer does not improve performance by further optimizing the scalar objective; it operates on geometric information the scalar objective discards.

The sequential drift result has the most direct implications for large language models and other autoregressive systems. Theorems 3A and 3B do not depend on adversarial perturbation or distributional shift (Quiñonero-Candela et al. 2009), it requires only that irreducible entropy or structural ambiguity persist in the generative process, which is the case in natural language (Manning and Schütze 1999). The Treasure Hunt long-horizon experiment provides direct empirical support for this theoretical claim: under fully deterministic task conditions where a simple algorithm achieves perfect alignment for any T, all twenty-nine tested models from seven providers exhibited the predicted $\mathrm{P}(A_T)$ decay, with model-specific thresholds beyond which collapse was rapid and near-total. The residual per-step error rate ε varied with architectural generation and training methodology, confirming that scaling reduces ε, but remained strictly positive in every model tested, producing the predicted multiplicative accumulation at long horizons. Scaling reduces ε but cannot reach $\varepsilon = 0$ in open domains. The practical consequence is that long-horizon coherence, instruction following, and intent alignment cannot be guaranteed by optimization alone: they require architectural mechanisms that monitor cumulative drift and modulate generation accordingly.

The theory admits precise falsifiability. It fails if:

(i) a system in So^{f} can internally discriminate $H(X \mid Y)$ from $H(Y \mid X)$ without architectural augmentation;
(ii) an objective-closed sequential system avoids asymptotic alignment decay under persistent uncertainty;
(iii) causal ambiguity is eliminable through capacity scaling alone; or
(iv) long-horizon alignment is achievable without governance mechanisms structurally distinct from loss minimization.

Each criterion is empirically testable, and the theory makes specific quantitative predictions (e.g., CEA divergent-failure rates as a function of overall accuracy) that can be assessed in larger systems.

Several limitations of the present work should be noted. The CEA empirical studies use relatively small-scale classification systems; the Distribution Drift experiment extends this empirical base to frontier-scale large language models, confirming that

the theoretical results are not confined to classification settings, though further investigation across additional model families and task types would strengthen generality. The governance rule demonstrated for CEA is minimal and proof-of-concept; the ACS framework does not prescribe a unique implementation, and future work should characterize the space of sufficient governance architectures. The operationalization of $H(Y \mid X)$ and $H(X \mid Y)$ through observable proxies (ensemble disagreement, Bayesian posterior variance (Kendall and Gal 2017), conformal prediction) requires further theoretical grounding. The parametric t-SNE embedding used in the CEA experiments is one operationalization of the geometric probe that CIT's governance mechanism requires; the theory predicts that any sufficiently rich auxiliary representation preserving local geometric structure should detect similar disagreement patterns. Future work should compare alternative embedding methods (e.g., UMAP, PCA, or learned probes) to establish robustness of the governance signal across representation choices.

3.1 Implications for AI system design

Three consequences of the preceding results are worth surfacing explicitly. First, within So^{f}, no combination of data, capacity, or calibration can eliminate the CE/CA error floor or the multiplicative drift ceiling; point-estimate commitment to a single scalar objective structurally forecloses the distinctions a controller would need in order to bound them. Second, data-and scale-centric remedies operate on the marginal $p(x)$ and therefore cannot, by Theorem 1, change the underlying entropy-ambiguity decomposition that generates these failure modes: scaling reduces the residual per-step error rate ε but, under persistent uncertainty, cannot drive it to zero. Third, ACS changes the system-class boundary rather than the optimization target. By governing on informational distinctions that the scalar objective discards-intent, inter representational conflict, and cumulative deviation-it moves a system out of So^{f} into a regime in which the static and sequential failure modes established here no longer bind. The distinction between optimization and cognition is, in this framing, architectural rather than quantitative.

The contribution of this work is structural clarification through Cognitive Information Theory. Scaling laws describe how approximation error decreases with data and model capacity. They do not describe what happens when irreducible entropy and causal ambiguity set a floor on achievable alignment, nor what happens when Distribution Drift accumulates across sequential commitments. The results presented here show that these floors are real, that they produce systematic and identifiable failure modes (CEA and Distribution Drift as the two principal CTT components), and that crossing them requires not more optimization but a different kind of architectural component altogether. The distinction between optimization and cognition, formalized here as the boundary between So^{f} and Cognitive AI , marks the architectural threshold at which reliable autonomous agency becomes structurally possible. ACS is presented here as a set of necessary structural requirements (persistent intent representation, algorithmic intent-output evaluation, drift detection, and behavioral modulation), not as a specific algorithm. It is related to, but structurally distinct from,

existing uncertainty-estimation, selective-prediction, and conformal-prediction mechanisms, in that it operates on the informational regime distinction (entropy versus ambiguity) rather than on aggregate predictive uncertainty alone.

4 Methods

4.1 Formal system definition and reliability conditions

Let X denote a random variable and Y observations generated by $p(x, y) = p(x)p(y \mid x)$. We define the class So^{f} of objective-function-only systems as those satisfying:

(1) Scalar Objective Closure, parameter updates are governed by a single scalar objective $\theta^* = \arg\min_\theta \mathbb{E}\left[\ell\left(f_\theta(X), Y\right)\right]$;
(2) Conditional Markov Sufficiency, outputs depend only on current parameters and prior generated tokens;
(3) No Independent Epistemic State-no internal state variable exists whose update dynamics are independent of the scalar objective and whose function is to modulate generation based on informational regime.

This definition includes supervised classifiers, maximum-likelihood generative models, and RL agents under expected return maximization. Probabilistic models are included if inference resolves to a single committed decision governed solely by the optimized objective.

Two minimal reliability conditions are defined:

(R1) Non-collapse under ambiguity, if $H\left(X \mid Y = y\right) > 0$, the system should not arbitrarily eliminate hypothesis plurality without qualification;
(R2) Containment under irreducible entropy, if $\widehat{H}\left(Y \mid X\right) > 0$, the sy stem should detect when residual uncertainty reflects irreducible stochasticity rather than reducible model error, and modulate behavior accordingly.

These conditions require only structural discrimination between ambiguity and entropy, not reasoning, abstraction, or semantic interpretation.

Note 5 (**Notation in this section:**) X and Y revert to their information-theoretic definitions, X = signal, Y = observations via $p(x, y) = p(x)p\left(y \mid x\right); \theta$ denotes model parameters; So^{f} denotes the class of objective-function-only systems; $(R1)$ and $(R2)$ denote the two minimal reliability conditions (non-collapse under ambiguity and containment under irreducible entropy, respectively).

5 Proof of CEA decomposition necessity (Theorems 1-2)

Theorem 1 (**Marginal Entropy Non-Identifiability**) *Let X be a random variable representing a signal with marginal distribution $p(x)$ and entropy $H(X)$. Let $F(X, \theta)$ be a model parameterized by θ whose inputs are samples of X and whose behavior depends on X as described by $p(x)$. Let $P_1(X)$ and $P_2(X)$ be two generative processes such that:*

$$p_1(x) = p_2(x) \text{ (therefore } H_1(X) = H_2(X) \text{)}$$

Then for any model $F(X, \theta)$ trained or evaluated on samples of X, F cannot distinguish whether samples were generated by P_1 or P_2. Importantly, F cannot identify or separate distinct sources contributing to $H(X)$.

Lemma 1 (**Observational Equivalence**) If $p_1(x) = p_2(x)$, then for any measurable function g,

$$E_{P_1}[g(X)] = E_{P_2}[g(X)].$$

All statistics computable from X alone are identical under P_1 and P_2.

Lemma 2 (**Asymptotic Model Equivalence**) Let $F(X, \theta)$ be a model whose parameters are determined by minimizing an expected loss

$$L(\theta) = E_{p(x)}[\ell(F(X, \theta), X)]$$

over the population distribution induce the same $p(x)$, the population risk is identical:

$$L_{P_1}(\theta) = E_{P_1}[\ell(F(X, \theta), X)] = E_{P_2}[\ell(F(X, \theta), X)] = L_{P_2}(\theta).$$

Therefore, any minimizer θ^* of the population risk is the same under P_1 and P_2, and the learned model $F(X, \theta^*)$ is identical regardless of which process generated the data.

Note 6 Finite-sample training introduces dependence on the specific sample realization (through optimization trajectory, batch ordering, etc.), but by the law of large numbers the empirical risk converges to the population risk as n $\to \infty$, and the distinction between P_1 and P_2 remains undetectable in the limit.

Lemma 3 (**Non-Identifiability of Entropy Decomposition-Constructive Proof**) Entropy is a functional of the marginal distribution. We show by construction that infinitely many joint distributions $p(x, y)$ with distinct conditional structures induce the same observation marginal $p(y)$.

Construction A (Noisy Injective): Let $X \in \{0, 1\}$ be uniform. Let $Y = X \oplus N$ where $N \sim$ Bernoulli (q) is independent channel noise. Then $H(Y \mid X) = H(q) > 0$ (irreducible stochasticity), and $H(X \mid Y) = H(q)$ (posterior ambiguity equals channel noise by symmetry).

Construction B (Deterministic Many-to-One): Let $X' \in \{0, 1, 2, 3\}$ be uniform. Let $Y' = X' \bmod 2$ (deterministic mapping). Then $H\left(Y' \mid X'\right) = 0$ (no channel noise), and $H\left(X' \mid Y'\right) = 1$ bit (two causes per observation, irrecoverable).

Both constructions can be parameterized to yield the same observation marginal $p(y) = p\left(y'\right)$, choose q such that $H(Y) = H\left(Y'\right) = 1$ bit, which is satisfied by any $q \in (0, 1)$ since Y is Bernoulli $(1/2)$ regardless of q, and Y' is Bernoulli $(1/2)$ by construction. Yet $H(Y \mid X) \neq H\left(Y' \mid X'\right)$ and $H(X \mid Y) \neq H\left(X' \mid Y'\right)$.

Since q can take any value in $(0, 1)$, infinitely many such joint distributions exist with identical observation marginals but distinct entropy decompositions. Therefore: the decomposition of $H(Y)$ into channel entropy $H(Y \mid X)$ and posterior ambiguity $H(X \mid Y)$ is non-identifiable from Y alone, and the source of uncertainty is unobservable without explicit access to the latent structure X.

Conclusion: For a model $F(Y, \theta)$ that encodes observation data in some latent representation R, it is not possible for F to distinguish between entropy originating from the data in its observation space and entropy originating from F 's own transformation into latent space R.

Example: Consider model F as an image classifier consisting of a feature extractor $g(Y) \to R$ and classifier $h(R) \to \widehat{X}$. The classifier h cannot distinguish between entropy arising from the generative channel $p(y \mid x)$ (e.g., visual variability within a class) and entropy introduced by the compression g (e.g., information loss from dimensionality reduction). Both contribute to $H(Y \mid R)$, but their distinct causal origins are invisible to any function operating on R alone.

Definition 1 (**So$^{\mathbf{f}}$**) The class So$^{\mathrm{f}}$ of objective-function-only systems comprises systems satisfying:

(1) Scalar Objective Closure, parameter updates are governed by a single scalar objective

$$\theta^* = \arg\min_{\theta} \mathbb{E}\big[\ell(f_\theta(X), Y)\big];$$

(2) Conditional Markov Sufficiency, outputs depend only on current parameters and prior generated tokens;

(3) No Independent Epistemic State, no internal state variable exists whose update dynamics are independent of the scalar objective and whose function is to modulate generation based on informational regime.

Definition 2 (**Reliability Conditions**) Two minimal reliability conditions are required:

R1 **(Non-collapse under ambiguity):** If $H(Y \mid X = x) > 0$, the system should not arbitrarily eliminate hypothesis plurality without qualification.

R2 **(Containment under irreducible entropy):** If $H(X \mid Y) > 0$, the system should detect when residual uncertainty reflects irreducible stochasticity rather than reducible model error, and modulate behavior accordingly.

Lemma 4 (**Point-Estimate Collapse**) If $H(Y \mid X = x) > 0$, then $p(y \mid x)$ assigns positive probability to more than one value of Y. Any deterministic decision rule $\mathrm{d}\theta(x) \in \mathcal{Y}$ selects a single value, discarding posterior probability mass of at least $1 - \max_y p(y \mid x) > 0$. The discarded mass is not recoverable from $\mathrm{d}\theta(x)$ alone, and no record of the plurality of the posterior is retained in the output.

Theorem 2 (**Structural Non-Sufficiency Under CEA**) *Let $S \in$ So$^{\mathrm{f}}$. If either $H(Y \mid X = x) > 0$ on a set of non-zero measure, or $H(X \mid Y) > 0$, then S cannot satisfy both R1 and R2 simultaneously.*

Proof

Prong 1 **(Ambiguity violates R1):** Suppose $H(Y \mid X = x) > 0$ on a set of non-zero probability. Then for those observations x, the posterior $p(y \mid x)$ is non-degenerate, multiple latent causes are consistent with the observation. Since $S \in \mathrm{So}^{\mathrm{f}}$ is objective-closed, its output is determined by $f\theta(x)$, a point-estimate decision rule. By Lemma 4, any such rule selects a single value from the support of $p(y \mid x)$, discarding the remaining posterior mass without qualification. This directly violates R1 (non-collapse under ambiguity).

Prong 2 **(Entropy violates R2):** Suppose $H(X \mid Y) > 0$. The system's parameters are determined by minimizing the population risk $L(\theta) = E_{p(x)}\big[\ell\big(f_\theta(X), Y\big)\big]$. By Theorem 1 (Lemmas 2 and 3), the population risk is identical across all generative processes that share the same marginal $p(x)$, regardless of their entropy decomposition, infinitely many joint distributions $p(x, \mathrm{y})$ with distinct values of $H(X \mid Y)$ and $H(Y \mid X)$ produce the same loss landscape $L(\theta)$. Therefore, the loss function provides no gradient signal that distinguishes reducible model error from irreducible channel stochasticity. Since $S \in \mathrm{So}^{\mathrm{f}}$ has no internal state independent of the scalar objective (condition 3 of So^{f}), no internal criterion for epistemic containment exists. This violates R2.

□

Conclusion: Since Prong 1 establishes that non-zero ambiguity alone violates R1, and Prong 2 establishes that non-zero entropy alone violates R2, S cannot satisfy both reliability conditions whenever either $H(Y \mid X = x) > 0$ or $H(X \mid Y) > 0$ holds. Since these conditions are jointly present in any environment with many-to-one causal mappings and genuine stochasticity, S is structurally non-sufficient under CEA.

Note 7 (**Notation in the preceding proofs:**) $F(X, \theta)$ denotes a model (equivalent to $f\theta$ used elsewhere); P_1 and P_2 denote distinct generative processes that may share the same observation marginal $p(x)$; R denotes the internal latent representation of a model, with g denoting a feature extractor and h a classifier head; ε in Lemma 3 denotes channel noise (a construction parameter), which should not be confused with the per-step violation probability s introduced below.

Notation shift: in the following sequential drift formalism, $X_1 : {}^{T}$, denotes a generated output sequence (not a latent cause); q^T denotes the model's autoregressive distribution over sequences; p^T denotes the ideal intent constrained distribution; and I^T denotes the intent-aligned subset of the sequence space.

6 Sequential drift formalism (Theorems 3A to 3B)

Let a generative system $S \in \mathrm{So}^{\mathrm{f}}$ produce sequences $X_1 : {}^{T} \sim \prod_{t=1}^{T} p_\theta(x_t \mid x_{<t}, c)$, where c denotes the fixed task context (e.g., system prompt or task specification). Define $A_t = \{X_{1:t}$ remains consistent with intent $I\}$, with $A_0 = \Omega$ (alignment holds vacuously prior to generation). By the chain rule,

$$\Pr(A_T) = \prod_{t=1}^{T} \Pr(A_t \mid A_{t-1}).$$

The persistent uncertainty condition assumes $\varepsilon > 0$ such that

$$\Pr\left(A'_t \mid A_{t-1}\right) \geq \varepsilon$$

for infinitely many t. This bound arises from

$$H\left(X \mid Y\right) > 0 \quad \text{or} \quad H\left(Y \mid X\right) > 0.$$

By Theorem 2, $S \in \mathrm{So}^{\mathrm{f}}$ cannot satisfy R1 or R2 under these conditions, so each point-estimate commitment carries a structurally irreducible probability of violating intent, establishing the persistence bound.

Let q^T denote the distribution over sequences $X_1 : {}^T$ induced by the autoregressive model p_θ.

Theorem 3A (**Asymptotic Decay**) *Let T_k index the subsequence where*

$$\Pr\left(A_t \mid A_{t-1}\right) \leq 1 - \varepsilon.$$

Since this subsequence is infinite,

$$\Pr\left(A_T\right) \leq (1-\varepsilon)^{N(T)},$$

where

$$N(T) \to \infty$$

Therefore,

$$\Pr\left(A_T\right) \to 0$$

Because $S \in \mathrm{So}^{\mathrm{f}}$ is objective-closed, violation probability cannot be adaptively reduced below the persistence bound.

Theorem 3B (**Exponential Decay Under Uniform Persistence**) *Under the stronger assumption that*

$$\Pr\left(A_t^c \mid A_{t-1}\right) \geq \varepsilon$$

for all sufficiently large t,

$$\Pr\left(A_T\right) \leq (1-\varepsilon)^T$$

for sufficiently large T, by the product rule.

Distributional divergence: since

$$\Pr\left(A_T\right) \to 0$$

by Theorem 3A,

$$q^T\left(I^{Tc}\right) \to 1.$$

For the ideal intent-constrained distribution

$$p^T = q^T\left(\cdot \mid I^T\right), \quad D_{\mathrm{KL}}\left(q^T \| p^T\right)$$

becomes unbounded once support mismatch occurs: p^T is supported entirely on I^T, while q^T places mass

$$q^T\left(I^{Tc}\right)$$

outside I^T, so

$$\mathrm{TV}\left(q^T, p^T\right) \geq q^T\left(I^{Tc}\right)$$

By Pinsker's inequality (Cover and Thomas 2006; Csiszár and Körner 2011),

$$D_{\mathrm{KL}}\left(q^T \| p^T\right) \geq 2\mathrm{TV}\left(q^T, p^T\right)^2 \geq 2\left(1 - q^T\left(I^T\right)\right)^2 \to 2.$$

Theorem 4 (**Necessity of Algorithmic Cognitive Sufficiency**) *Assume persistent uncertainty (weak form). If a sequential system satisfies the long-horizon alignment condition:*

$$\forall \delta > 0, \exists\, T_0 \text{ such that } \forall\, T \geq T_0, \Pr\left(X_{1:T} \in I^T\right) \geq 1 - \delta,$$

then it cannot belong to So^{f}.

Proof

If $S \in \mathrm{So}^{\mathrm{f}}$, Theorem 3A gives

$$\Pr\left(A^T\right) \to 0,$$

implying

$$\liminf_{T\to\infty} \Pr\left(A^T\right) = 0.$$

But the alignment condition requires

$$\liminf_{T\to\infty} \Pr\left(A^T\right) = 1.$$

These limits are incompatible. □

ACS comprises four minimal components that together invalidate the assumptions underlying Theorem 3A:

(1) Persistent Intent Representation, I is represented independently of transient token context and local predictive likelihood, analogous to a stable utility functional (Savage 1954);
(2) Algorithmic Intent-Output Evaluation, candidate outputs are evaluated relative to I through a relation distinct from the scalar training objective, corresponding to robust evaluation under model uncertainty (Gilboa and Schmeidler 1989; Hansen and Sargent 2008);
(3) Drift Detection, cumulative deviation from I is monitored across sequential commitments, paralleling model-misspecification monitoring in robust control (Hansen and Sargent 2008);
(4) Behavioral Modulation, detection influences action, including abstention or scope restriction when ambiguity or entropy exceeds tolerable bounds (Gilboa and Schmeidler 1989; Bewley 2002).

7 Experimental setup

CIFAR-10 study: A CNN was trained on 50,000 training images (10 classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck) with scalar cross-entropy optimization, achieving 90.70% training accuracy and 81.74% test accuracy. A parametric t-SNE module (van der Maaten and Hinton 2008) was applied post-hoc to the 512-dimensional output of the final dropout layer, producing a continuous 2-dimensional

embedding that preserves local geometric structure. The t-SNE module was trained on a held-out portion of training data and applied to test images as an external epistemic probe. Confusion-region rules were derived from training-data boundary analysis and applied without modification to test data.

MNIST study: A CNN was trained on 60,000 training images ($28 \times 28 \times 1$ grayscale; 10 digit classes) with scalar cross-entropy optimization, achieving 99.37% training accuracy and 99.29% test accuracy. The architecture comprised three convolutional-pooling-normalization blocks, a flatten layer (288-dimensional output), and two dense layers. The parametric t-SNE module was applied to the flatten-layer output, the representation prior to scalar compression by the dense layers, rather than the penultimate dense layer, because the DPI layer-comparison experiment (confirming that dense compression attenuates geometric structure) demonstrated that flatten-layer geometry provides richer epistemic signal. Both studies applied the same governance rule: conflict between CNN scalar prediction and t-SNE nearest-cluster assignment triggers formation of a 3-4 class candidate set for further evaluation.

8 Distribution Drift experiment (Theorem 3B)

8.1 Task design

The test instrument is Treasure Hunt: a grid-navigation game on a deterministic 10×10 board pre-populated with 10 gold squares, 20 trap squares, and 70 empty squares. Movement rules are fixed: right/left/up/down change the column or row coordinate by 1 ; boundary violations produce BLOCKED; landing on gold produces YIPEE; trap produces OUCH; empty square produces OK ; gold and trap squares become permanently empty after first visit. The complete board lay out is provided to the model before any moves are made. The model is presented with a sequence of T moves in a single API call and required to function as the game engine, returning a JSON object with the correct response at each step, along with cumulative gold-collected and trap-hit counts. Ground truth is pre-computed algorithmically before any model call. Any departure from the ground truth at any step constitutes an error.

8.2 Test configuration

Each test set comprised 20 independently randomized boards ($N = 20$). Sequence lengths tested: $T \in \{2, 5, 10, 20, 50, 100, 200, 300\}$. API temperature was set to 1 . The primary metric was the proportion of 20 boards completed without any error (perfect-test binary measure). A secondary metric (average failures per failed game) was computed to characterize the collapse cascade once the threshold is crossed. Token counts were recorded at each T to confirm the absence of context-window exhaustion: prompt tokens ranged from approximately 17,700 at $T = 5$ to 29,500 at $T = 300$, well within all tested models' documented context limits.

8.3 Models

Five OpenAI models were tested spanning two architectural generations:

(1) GPT-5.3 (gpt-5.3-chat-latest; hybrid reasoning router with chain-of-thought state and behavioral RL fine-tuning),
(2) GPT-5.2 (gpt-5.2-chat-latest; CoT pass-through with GB200-cluster RL training),
(3) GPT5.1 (gpt-5.1-chat-latest; hybrid router with live correctness signal, no CoT state persistence),
(4) GPT5 (gpt-5-chat-latest; first-generation hybrid with merged RL, no reasoning control mechanisms), and
(5) GPT-4o (gpt-4o; dense transformer (Vaswani et al. 2017) with RLHF post-training, no integrated reasoning mechanisms).

Models without per-step reasoning control mechanisms (GPT-4o; GPT-5-chat-latest) were included specifically to assess whether the observed degradation reflects Distribution Drift or baseline CE/CA per-step error rates. Inclusion threshold: at least one perfect-game completion at any sequence length.

9 Related Work

CIT draws on and extends several established research traditions. In information theory and representation learning, the information bottleneck framework (Tishby et al. 1999) formalizes learning as extracting a compressed representation that preserves task-relevant information. CIT's compression argument is related but distinct: it concerns not the trade-off between compression and relevance, but the non-identifiability of entropy sources under scalar compression, a structural claim about what compression hides rather than what it preserves.

In decision theory, CIT builds on the ambiguity-aversion tradition originating with Knight (1921) and Ellsberg (1961), and formalized by Gilboa and Schmeidler (1989) through maxmin expected utility over sets of priors. CIT's contribution is not the discovery of ambiguity-sensitive decision structure but the identification of formally analogous constraints within modern AI architectures, specifically that scalar objective closure in Sof produces the same structural insufficiency that Savage-style expected utility exhibits under Knightian uncertainty.

In calibration and uncertainty estimation, modern neural networks are known to be poorly calibrated (Guo et al. 2017), and predictive uncertainty can be improved through Bayesian approximations such as MC dropout (Gal and Ghahramani 2016) and deep ensembles (Lakshminarayanan et al. 2017). Selective prediction and reject-option systems Geifman and El-Yaniv (2017) explicitly abstain on uncertain inputs, while conformal prediction Angelopoulos and Bates (2023) provides distribution-free prediction sets. Out-of-distribution detection Hendrycks and Gimpel (2017) addresses high-confidence predictions on inputs outside the training distribution. CIT's ACS governance layer is related to these approaches but operates at a different level of abstraction: rather than estimating aggregate predictive uncertainty, ACS detects divergence between scalar and geometric representations as a proxy for the entropy-ambiguity regime distinction that CIT formalizes. The experiments presented here use

a specific operationalization (t-SNE disagreement), but the theoretical framework is agnostic to the choice of geometric probe.

In long-horizon reasoning and planning, recent work has documented that LLM-based agents frequently fail on extended sequential tasks, with step-wise reasoning proving insufficient for sustained planning. CIT's Distribution Drift result enters this active discussion not as the first observation of long-horizon degradation but as a specific formal interpretation: the exponential decay form

$$\Pr(A_T) \leq (1-\varepsilon)^T$$

linked to the structural properties of So^{f}, and the proof that

$$\varepsilon > 0$$

is guaranteed under persistent uncertainty for objective-closed systems. The cross-provider empirical confirmation across twenty-nine models strengthens this interpretation.

Acknowledgments

The authors would like to acknowledge the valuable contributions of Katy Abadi for research assistance, Nicole Beauregard for careful reading and critical review of the manuscript, and Dr. Durgesh Kushawaha for independent technical review, editorial guidance, and manuscript refinement during the final stages of preparation.

Appendix S1 Supplementary Proof S1

Compression-Amplified Entropy and Ambiguity: Formal Derivation of the CE and CA Lower Bounds

S1.1 Setup and Definitions

We adopt the notation of the main text. Let X denote latent causes and Y observations generated by the joint distribution $p(x, y) = p(x)p(y \mid x)$. The irreducible quantities of interest are:

- **Channel entropy:** $H(Y \mid X) = -\sum_{x,y} p(x, y) \log p(y \mid x)$
- **Posterior ambiguity:** $H(X \mid Y) = -\sum_{x,y} p(x, y) \log p(x \mid y)$

Now consider a machine learning system that encodes observations through a learned representation. Let $R \in \mathbb{R}^d$ denote the internal representation (e.g., the output of the penultimate layer of a convolutional neural network), forming the Markov chain:

$$X \to Y \to R$$

The joint distribution factorises as $p(x, y, r) = p(x)p(y \mid x)p(r \mid y)$, where $p(r \mid y)$ is the (possibly stochastic) encoding function learned by the model. We define:

- **Compression-amplified Entropy (CE):** $H(Y \mid R)$
- **Compression-amplified Ambiguity (CA):** $H(X \mid R)$

These quantities measure the residual uncertainty about observations and causes, respectively, when the system has access only to the compressed representation R rather than the raw observation Y.

S1.2 The Data Processing Inequality

The proofs rest on the Data Processing Inequality (DPI), a foundational result in information theory (Cover and Thomas (2006), Theorem 2.8.1). The DPI states that for any Markov chain $A \to B \to C$:

$$I(A; C) \leq I(A; B)$$

where $I(\cdot;\cdot)$ denotes mutual information. Post-processing cannot increase the information that a signal carries about any upstream variable. Equality holds if and only if C is a sufficient statistic of B for A.

S1.3 The CA Lower Bound

Proposition S1.1 (**Compression-Amplified Ambiguity Lower Bound**) *For the Markov chain $X \to Y \to R$:*

$$\text{CA} = H(X \mid R) \geq H(X \mid Y)$$

with equality if and only if R is a sufficient statistic of Y for X. Proof. Apply the DPI to $X \to Y \to R$:

$$I(X; R) \leq I(X; Y)$$

Expand both sides using

$$I(X; Y) = H(X) - H(X \mid Y) : H(X) - H(X \mid R) \leq H(X) - H(X \mid Y)$$

Cancel $H(X)$ and multiply by -1:

$$H(X \mid R) \geq H(X \mid Y)$$

Interpretation

The compressed representation R cannot preserve more information about latent causes X than the raw observation Y from which it was derived. Since $H(X \mid Y)$ is the irreducible posterior ambiguity-determined by the causal structure of the environment-no amount of optimization of the encoder $p(r \mid y)$ can reduce ambiguity below this floor. Compression can only amplify it.

S1.4 The CE and CA Bounds: Precise Status

The information-theoretic analysis yields two distinct results of different logical status:

Result 1 (CA Bound-unconditional)

For any Markov chain $X \to Y \to R$:

$$H(X \mid R) \geq H(X \mid Y)$$

This is a direct consequence of the DPI and holds without further assumptions. Compression into R cannot reduce posterior ambiguity about latent causes below the irreducible level $H(X \mid Y)$. **This is proved in Proposition S1.1.**

Result 2 (CE Bound-conditional on the operational regime)

The bound $H(Y \mid R) \geq H(Y \mid X)$ does not follow from the DPI alone, because the DPI constrains information about X (the upstream variable), not about Y (the variable from which R is derived). $H(Y \mid R)$ and $H(Y \mid X)$ measure conditioning on variables at different positions in the causal chain, and their ordering depends on the specific joint distribution $p(x, y)$ and encoder $p(r \mid y)$.

However, the CE bound holds under a condition that is universally satisfied in the CIT-relevant regime:

Proposition S1.2 (**CE Bound-sufficient condition**) *If $I(Y;R) \leq I(Y;X)$, then $H(Y \mid R) \geq H(Y \mid X)$.*

This condition is satisfied whenever:

(a) The representation R is a lossy compression of Y such that R discards more information about Y than the generative process introduces (i.e., the encoding is not richer than the causal structure). In practical neural network architectures used for classification, R is a dimensionality-reducing bottleneck (e.g., 512-dimensional or lower), and Y is a high-dimensional observation (e.g., $32 \times 32 \times 3$ image $= 3072$ dimensions). The encoder necessarily discards information, while X (the latent class) determines Y up to channel noise. The condition $I(Y;R) \leq I(Y;X)$ holds whenever the information loss from compression exceeds the irreducible channel entropy.

(b) More precisely, for classification tasks where X is a discrete class label and Y is a high-dimensional continuous observation, $I(Y;X) = H(Y) - H(Y \mid X)$ captures the total information the class identity provides about the observation. When $H(Y \mid X)$ is small relative to $H(Y)$ (i.e., the class is highly informative), $I(Y;X)$ is large and the condition is easily satisfied.

The CIT claim, precisely stated, is therefore: In the operationally relevant regime-where a high-dimensional observation Y is compressed into a lower-dimensional representation R for the purpose of predicting a discrete latent cause X-compression amplifies both posterior ambiguity (CA, unconditionally) and forward-channel entropy (CE, under the lossy-compression condition). The compound state CEA arises when both $H(Y \mid X) > 0$ and $H(X \mid Y) > 0$, and the system's compressed representation conflates these two structurally distinct sources of uncertainty.

S1.5 Summary

Table S1: Summary of CE, CA, and CEA lower bounds.

Bound	Statement	Status	Proof mechanism
CA	$H(X \mid R) \geq H(X \mid Y)$	Unconditional	DPI, Proposition S1.1
CE	$H(Y \mid R) \geq H(Y \mid X)$	Conditional on lossy-compression regime	Mutual information ordering, Proposition S1.2
CEA	Both CE and CA active simultaneously	Follows when $H(Y \mid X) > 0$ and $H(X \mid Y) > 0$	Conjunction of CA and CE

Appendix S2 CITRI Drift Formal Experiment Description

Formal Experiment Description

Measuring $P(T)$: Precision Degradation Over Long Horizons

in Large Language Models Without External Governance

A Supplementary Empirical Study in Support of Cognitive Information Theory (CIT)
April 2026

Researchers: Damian Fozard (Author & Collaborator), Ken Wenger (Author & Collaborator)

Experimental Assistant: Katy Abadi

Abstract

This study provides a formal empirical test of a core prediction of Cognitive Information Theory (CIT): that any objective-function-only Al system will exhibit declining alignment probability $P\left(A^T\right)$ as sequential task length T increases, converging to zero regardless of model scale. The task used is a fully deterministic game in which $P = 1$ at every step, meaning a correct algorithmic solution exists and performs with 100% precision for any T. The experiment isolates the question of whether a large language model, without external governance, can maintain that precision over increasing T. **Twenty-nine models from seven providers: OpenAI, Anthropic, DeepSeek, xAI (Grok), Moonshot AI (Kimi), Mistral AI, Meta AI (Llama), and Cohere, spanning multiple architectural generations, scale tiers from $\sim$ 20 B to over 1 trillion parameters, and both open-weight and closed-weight families were tested across sequence lengths of 2 to 300 steps.** Results confirm CIT Theorem 3B: every model tested degrades, with a characteristic threshold beyond which collapse is rapid and near-total. The cross-provider dataset demonstrates that this is not a property of any single vendor's architecture

but a universal phenomenon. Critically, model scale, as measured by parameter count, does not predict long-horizon performance: DeepSeek-V3.1 (671 B total parameters) collapses entirely at NS = 20, while the much smaller 04 -mini maintains 80% completion at NS = 100. Scaling is necessary but not sufficient; what differentiates models is architecture and training methodology, particularly reasoning-focused mechanisms.

S2.1 Purpose and Theoretical Motivation

This experiment was designed to test a specific, falsifiable prediction of Cognitive Information Theory (CIT): that the probability of maintaining full alignment with an objective function $P\left(A^T\right)$ decays asymptotically towards zero as sequential task length T increases, in any system operating without epistemic governance. CIT formalizes this as:

$$Pr\left(A^T\right) \leq (1-\varepsilon)^T$$

where $\varepsilon > 0$ is a persistent per-step violation probability that scaling can reduce but cannot eliminate to zero in any open-domain generative system. The theorem predicts that newer, larger models exhibit smaller ε, they degrade more slowly, but degrade nonetheless. This study tests that prediction directly across 29 models from seven providers to establish universality.

S2.1.1 Why $P = 1$ Tasks Are the Correct Test

The experiment was purposefully designed around a task for which $\mathbf{P} = \mathbf{1}$ is the correct and achievable answer: a fully deterministic game engine with a unique correct response at every step. This choice is essential to the experimental logic. If the task admitted genuine ambiguity or multiple valid responses, observed errors might reflect legitimate output variation rather than model failure. By selecting a task that a simple algorithm can perform with 100% precision for any T, we isolate the precise question: can an LLM, without external governance, maintain what is trivially achievable, over increasing T?

Any departure from the correct answer is therefore an unambiguous failure and not a matter of interpretation. The experiment is not testing knowledge, creativity, or reasoning depth. It is testing whether the LLM can sustain precision for a straightforward task as horizon length increases.

S2.1.2 The Role of CIT's CE and CA States

CIT identifies two principal static failure mechanisms that operate independently of sequence length:

- **Confident Entropic Error (CE):** the model produces high-confidence incorrect outputs when irreducible stochastic variation in the input channel is mapped into stable decision regions.
- **Compression-Amplified Ambiguity (CA):** the model collapses hypothesis plurality under scalar objective compression, committing to a point estimate when causal under determination demands a candidate set.

The four possible response states in this game (OK, YIPEE, OUCH, BLOCKED) create a simple discrete choice at each step. The probability of a correct choice is not a complex reasoning problem, it is a precise four-way classification with one uniquely correct answer. For models that fail to reliably complete even a two-step sequence, the challenge is not long-horizon drift. It is the underlying CE/CA-predicted probability of making the correct choice at any step.

S2.1.3 Architectural Context: What the Cross-Provider Comparison Reveals

Twenty-nine models from seven distinct providers were tested, enabling a powerful analysis of the relationship between architecture, scale, and long-horizon precision. The results reveal several architectural patterns:

- Reasoning-optimized models sustain precision longest. GPT-5.3 (with CoT state, RL reasoning, and behavioral fine-tuning), Kirni K2.5 ($\sim$ 1 T MoE with structured reasoning training), and 04-mini (a compact reasoning-optimized model) are the top three performers. Their shared characteristic is not raw scale but purpose-built reasoning control mechanisms.
- Scale without reasoning control does not prevent early collapse. DeepSeek-V3.1 (671B parameters, one of the largest open-weight models) collapses at NS=20 with only 5% completion. Claude Opus 4$\times$ models ($\sim$ 2 T estimated) collapse at NS = 50. Raw parameter count is a poor predictor of long-horizon performance.
- The CE/CA signature is provider-universal. Models without reasoning control mechanisms whether from OpenAI (GPT-4o, GPT-5-chat-latest), Anthropic (claude-sonnet-4-6), DeepSeek (V3.1, V3.2), Mistral (Large-3), or Meta (Llama 4 Maverick, all show the same pattern: declining completion rates from the shortest sequences, consistent with high per-step CE/CA error rates rather than progressive drift.

S2.2 Experiment Design

S2.2.1 The Treasure Hunt Game Engine

The task vehicle is Treasure Hunt: a grid-navigation game played on a deterministic 10×10 board. The board is randomly populated before each test with 20 trap squares and 10 gold squares; the remaining 70 squares are empty. The complete board layout is provided to the model before any moves are made. **Movement rules are entirely deterministic.** Right increases X by 1; left decreases X by 1; up increases Y by 1; down decreases Y by 1. Any move that would exceed the board boundary is BLOCKED. Landing on a gold square produces YIPEE; a trap produces OUCH; an empty square produces OK. A gold or trap square becomes permanently empty after first visit.

The model is presented with the full board layout and a sequence of T moves, and is asked to function as the game engine returning a JSON object with the correct response for every move, along with gold-collected and trap-hit counts. The entire

sequence is submitted in a single API call; the model processes the complete sequence autonomously, tracking board state as it goes.

S2.2.2 Key Experimental Properties

Precisely four output states. At every step, the correct response is one of: OK, YIPEE, OUCH, or BLOCKED. This is not a complex reasoning task with open-ended outputs. It is a four-way discrete classification.

Token scaling does not change with T. As sequence length increases, the prompt size changes only modestly, the board description is constant, and move sequences are short strings. Even at NS = 300, prompt tokens remain well within all models' context limits.

P = 1 is achievable. A simple Python algorithm performs this task with 100% accuracy for any T. The correct answer at every step is uniquely determined given the board layout and all prior moves.

Binary completion as the primary measure. The primary metric is the proportion of tests completed without any error, a binary outcome. In an autonomous agent context, a task that is 99% correct but contains one error may be entirely unusable.

S2.2.3 Test Configuration

Table S2: Experimental configuration for the Treasure Hunt benchmark.

Parameter	Value
Boards per test set (NT)	20
Games per board (NG)	1
Sequence lengths tested (NS)	2, 5, 10, 20, 50, 100, 200, 300 moves
Board size	10×10 grid
Gold squares per board	10
Trap squares per board	20
Possible responses per move	4 (OK, YIPEE, OUCH, BLOCKED)
API temperature	1
Call structure	Single API call per test (full board + all moves in one prompt)
Scoring	Per-move comparison against pre-computed ground truth
Primary metric	Proportion of tests completed with zero errors (binary)
Inclusion threshold	At least one perfect game at any sequence length
Token regime	Prompt tokens constant per NS; no context window exhaustion observed

S2.2.4 Models Tested

Twenty-nine models (see Table S3) were tested across seven providers, spanning multiple architectural generations, scale tiers, and design philosophies:

S2.3 Results

S2.3.1 Summary: Proportion of Tests Completed Successfully

The Table S4 below presents the primary metric: proportion of 20 tests completed without any error, for each model at each sequence length. Models are ranked by overall performance (latest NS at which any successful completions occur, then by completion percentage). This is the "perfect test" binary measure.

Table S3: Complete list of 29 models tested, grouped by provider. Models range from ∼ 20B estimated parameters to over 1 trillion, covering dense transformers, MoE architectures, reasoning-optimized variants, and meta-routing systems.

Model	**Family**	**Provider**	**Key Characteristics**
gpt-5.3-chat-latest	GPT-5 Series	OpenAI	Hybrid router + CoT state + behavioural RL fine-tuning
gpt-5.2-chat-latest	GPT-5 Series	OpenAI	CoT pass-through; GB200 cluster RL; MRCR leader
gpt-5.1-chat-latest	GPT-5 Series	OpenAI	Hybrid router; live correctness signal; no CoT state
gpt-5-chat-latest	GPT-5 Series	OpenAI	First-generation hybrid; RL merged; no reasoning control
gpt-5.4-mini	GPT-5 Series	OpenAI	Smaller 5-series variant
gpt-5.4-nano	GPT-5 Series	OpenAI	Smallest 5-series variant
gpt-4o	GPT-4 Series	OpenAI	Dense transformer; RLHF only; no reasoning mechanisms
o4-mini	Reasoning	OpenAI	Reasoning-optimised compact model
o3-mini	Reasoning	OpenAI	Reasoning-optimised compact model (prior generation)
model-router	Meta-model	OpenAI	Dynamically routes to appropriate backend model
claude-opus-4-1	Claude 4 Series	Anthropic	Large frontier model; estimated ∼ 2T parameters
claude-opus-4-5	Claude 4 Series	Anthropic	Updated Opus; extended context
claude-opus-4-6	Claude 4 Series	Anthropic	Latest Opus; up to 1M context
claude-sonnet-4-5	Claude 4 Series	Anthropic	Mid-tier ∼ 70B estimated; balanced cost/capability
claude-sonnet-4-6	Claude 4 Series	Anthropic	Latest Sonnet; default free-tier model
claude-haiku-4-5	Claude 4 Series	Anthropic	Speed/cost tier; ∼ 20B estimated; fastest Claude
Kimi-K2.5	Open-weight	Moonshot AI	∼ 1T MoE; 256K context; strong structured reasoning
DeepSeek-R1	Open-weight	DeepSeek	671B MoE (37B active); RL reasoning training
DeepSeek-R1-0528	Open-weight	DeepSeek	Updated R1; 671B MoE (37B active)
DeepSeek-V3.1	Open-weight	DeepSeek	671B MoE (37B active); general-purpose
DeepSeek-V3.2	Open-weight	DeepSeek	Updated V3; 671B MoE (37B active)
Mistral-Large-3	Open-weight	Mistral AI	∼ 673–675B sparse MoE; ∼ 41B active
Llama-4-Maverick-17B-128E	Open-weight	Meta AI	400B MoE; 17B active; 128 experts; 1M context
grok-3	Grok Series	xAI	∼ 2.7T reported; 131K context
grok-3-mini	Grok Series	xAI	Smaller reasoning variant of Grok 3
grok-4-fast-reasoning	Grok Series	xAI	Grok 4 with reasoning capabilities
grok-4-1-fast-reasoning	Grok Series	xAI	Updated Grok 4 reasoning variant
grok-4-1-fast-non-reasoning	Grok Series	xAI	Grok 4.1 without reasoning mode
cohere-command-a	Enterprise	Cohere	Enterprise-focused; strong tool use; 256K context

Table S4: Proportion of tests completed with zero errors by model and sequence length T (29 models, 7 providers). Each data point represents 20 independent boards. "Time out" indicates the model exceeded the maximum allowed response time.

Model	NS=2	NS=5	NS=10	NS=20	NS=50	NS=100	NS=200	NS=300
gpt-5.3-chat-latest	100%	100%	100%	100%	100%	85%	25%	0%
Kimi-K2.5	100%	100%	100%	100%	95%	85%	Time out	Time out
o4-mini	100%	100%	100%	95%	95%	80%	5%	0%
gpt-5.2-chat-latest	100%	100%	100%	100%	100%	70%	0%	0%
o3-mini	100%	100%	100%	95%	85%	75%	5%	10%
model-router	100%	100%	100%	100%	85%	45%	5%	5%
grok-4-fast-reasoning	100%	100%	100%	90%	100%	20%	5%	0%
DeepSeek-R1	100%	100%	100%	90%	80%	10%	5%	0%
DeepSeek-R1-0528	100%	100%	100%	100%	50%	5%	0%	0%
grok-3-mini	100%	100%	100%	95%	90%	Time out	Time out	Time out
grok-4-1-fast-reasoning	100%	100%	95%	90%	85%	0%	0%	0%
gpt-5.1-chat-latest	100%	100%	95%	90%	0%	0%	0%	0%
claude-opus-4-5	100%	100%	90%	95%	0%	0%	0%	0%
claude-opus-4-6	100%	90%	90%	95%	0%	0%	0%	0%
claude-sonnet-4-5	95%	80%	95%	85%	0%	0%	0%	0%
claude-opus-4-1	100%	80%	80%	70%	0%	0%	0%	0%
grok-3	95%	90%	90%	55%	15%	0%	0%	Time out
gpt-5.4-mini	100%	65%	60%	10%	0%	0%	0%	0%
gpt-5-chat-latest	70%	60%	50%	15%	0%	0%	0%	0%
gpt-5.4-nano	70%	50%	50%	15%	0%	0%	0%	0%
gpt-4o	95%	60%	40%	5%	0%	0%	0%	0%
DeepSeek-V3.1	95%	55%	35%	5%	0%	0%	0%	0%
DeepSeek-V3.2	90%	55%	30%	15%	0%	0%	Time out	Time out
Mistral-Large-3	90%	20%	25%	10%	0%	0%	0%	0%
claude-haiku-4-5	75%	65%	45%	0%	0%	0%	0%	0%
Llama-4-Maverick-17B-128E	75%	40%	15%	0%	0%	0%	0%	0%
cohere-command-a	65%	25%	20%	0%	0%	0%	0%	0%
grok-4-1-fast-non-reasoning	55%	20%	15%	0%	0%	0%	0%	0%
claude-sonnet-4-6	100%	45%	0%	0%	0%	0%	0%	0%

The pattern is immediately consistent with CIT Theorem 3B and is demonstrated to be universal across providers. Every model that demonstrates task capability shows declining $P\left(A^T\right)$ as T increases. The ranking reveals that reasoning-optimised models (GPT-5.3, Kimi K2.5, 04-mini, 03-mini) sustain precision deepest into long horizons, while models without reasoning control mechanisms, regardless of provider or scale, collapse earlier.

S2.3.2 Performance Chart: OpenAI Model Genealogy

To isolate the effect of architectural progression within a single providers model genealogy, removing confounding differences in training data, tokenization, and organizational design choices across providers, the chart below presents the five OpenAI GPT models as a controlled comparison. This single-vendor view makes the relationship between architectural generation and long-horizon precision particularly clear.

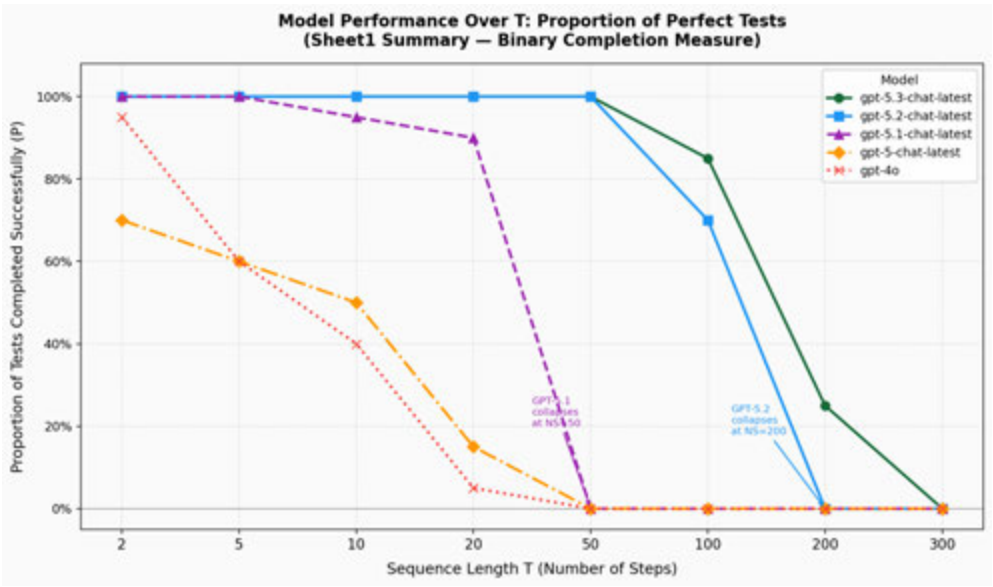

Fig. S1: Proportion of tests completed successfully (P) vs. sequence length T for five OpenAl GPT models. This single genealogy view isolates the effect of architectural progression (GPT-4 → GPT-5 series) while controlling for provider-level differences. The same characteristic decay pattern is confirmed across all 29 models and seven providers in Table S4.

S2.3.3 The Collapse Signature: Detailed OpenAI Failure Analysis

To characterize the mechanics of collapse in detail, the table below presents per-game failure metrics for the OpenAI GPT models. Focusing on a single provider's model

family allows direct comparison of collapse behavior across architectural generations without confounding cross-provider differences. The patterns observed here, abrupt threshold collapse and error cascading, are consistent with the decay curves seen across all 29 models in Table S4.

Table S5: Selected collapse metrics from raw test data for OpenAI GPT models. This single-generation view enables direct comparison of collapse mechanics across architectural generations (GPT-4o through GPT-5.3).

Model	NS	Avg First Error (Move)	Avg Failures / Failed Game	Interpretation
gpt-5.3	100	47.7	7.3	Isolated errors; model largely recovers
gpt-5.3	200	116.5	64.1	Drift accumulating; cascade begins
gpt-5.3	300	5.1	295.9	Threshold crossed; near-complete collapse
gpt-5.2	100	10.3	73.7	Threshold proximity; cascading pattern
gpt-5.2	200	1.45	199.6	Threshold crossed; failure from step 1–2
gpt-5.1	50	0.1	50.9	Complete collapse; fails immediately
gpt-4o	100	8.95	45.8	CE/CA pattern; persistent early failure

S2.4 Model Scale Analysis: Why Bigger Is Not Better

A critical dimension of this study is the relationship between model scale and long-horizon performance. The AI Model Scale Reference (compiled April 2026) provides parameter counts, context windows, and architectural details for all models tested. This data allows us to directly examine whether scaling alone predicts the observed performance patterns.

S2.4.1 Known Model Scale Parameters

Table S6: Model scale reference. Open-weight parameter counts are verified; closed-weight figures are community estimates. Sources: Anthropic API Docs, DeepSeek GitHub/HuggingFace, Meta AI Blog, Moonshot AI, Mistral AI, Artificial Analysis.

Model	**Total Params**	**Active Params**	**Context Window**	**Architecture**
Kimi K2.5	$\sim$ 1 Trillion	Undisclosed	256K	MoE
DeepSeek-R1/R1-0528	671B	37B	128K	MoE
DeepSeek-V3.1/V3.2	671B	37B	128K	MoE
Mistral Large 3	$\sim$ 673–675B	$\sim$ 41B	128K	Sparse MoE
Llama 4 Maverick	400B	17B	1M tokens	MoE (128 experts)
GPT-5.x series	Undisclosed	—	272K–1M	Undisclosed
o4-mini/o3-mini	Undisclosed	—	200K	Reasoning-optimised
Claude Opus 4.x	$\sim$ 2T (est.)	—	200K–1M	Undisclosed
Claude Sonnet 4.x	$\sim$ 70B (est.)	—	200K–1M	Undisclosed
Claude Haiku 4.5	$\sim$ 20B (est.)	—	200K	Undisclosed
Grok 3/grok-3-mini	$\sim$ 2.7T (reported)	—	131K	Undisclosed
Grok 4.x	Undisclosed	—	131K	Undisclosed
Cohere Command-A	Undisclosed	—	256K	Undisclosed

Note 8 (**Note on MoE architectures:**) For Mixture-of-Experts models, the "active params" figure is the more relevant compute proxy during inference. A 671B total MoE model activating 37B parameters per token is computationally closer to a 37 B dense model than a 671 B dense model. Raw parameter count without the active-params figure is nearly meaningless for comparing inference cost, speed, or, as this experiment demonstrates, long-horizon reasoning capability.

S2.4.2 Practical Scale Tiers

Based on what is known or reliably estimated, the models from this benchmark fall into the following tiers:

Table S7: Practical scale tiers for all tested models.

Tier	Models	Scale Signal
Massive MoE (1T + total)	Kimi K2.5, Grok 3	Largest raw capacity; very low active-parameter ratio
Large MoE (600–700B total)	DeepSeek R1/V3.x, Mistral Large 3	Open-weight leaders; $\sim$ 37–41B active per token
Medium-Large (400B total)	Llama 4 Maverick	17B active; 128 experts; strong context (1M tokens)
Unknown/Closed Frontier	GPT-5.x, Claude Opus, Grok 4.x	Likely very large; architecture undisclosed
Smaller Reasoning Models	o4-mini, o3-mini, grok-3-mini, claude-haiku-4.5	Smaller base; heavily trained on reasoning tasks

S2.4.3 The Scale-Performance Paradox

The benchmark results reveal a striking paradox that raw parameter count does not predict:

- **DeepSeek-V3.1** (671B total parameters, one of the largest open-weight models in the world) collapses entirely at NS = 20, achieving just 5% completion. By contrast, o4-mini (a compact reasoning-optimized model with far fewer parameters) maintains 80% completion at NS=100 and still achieves 5% at NS = 200.
- **Kimi K2.5** ($\sim$ 1 trillion total parameters) performs exceptionally well, maintaining 85% at NS = 100. But o3-mini (a much smaller model) achieves 75% at the same point. The smaller model's reasoning training compensates substantially for its scale deficit.
- **Claude Opus 4.x** ($\sim$ 2 T estimated, among the largest models tested) collapses to 0% at NS = 50 across all versions tested. Meanwhile, GPT-5.3 (scale undisclosed but likely comparable) maintains 100% at NS = 50 and 85% at NS = 100.
- **Mistral Large 3** ($\sim$ 673 $-$ 675 B, $\sim$ 41 B active) drops to 20% at NS = 5 and effectively collapses by NS = 20. Llama 4 Maverick (400 B total, 17B active) shows a similar pattern. Both are large MoE models that fail early despite substantial parameter counts.

S2.4.4 Why Parameter Count Is an Increasingly Poor Proxy

The scale reference data illuminates several reasons why raw parameter count fails to predict long-horizon performance:

- **MoE architecture changes the compute reality.** A 671B MoE model activating 37B parameters per token is computationally closer to a 37B dense model.

Kimi K2.5's 1T total parameters sound enormous, but the per-token active compute is far smaller. Raw parameter count without the active-params figure is nearly meaningless.

- **Frontier labs stopped disclosing.** OpenAl disclosed GPT-1 through GPT-3 parameter counts openly but has been silent since GPT-4. Anthropic has never officially confirmed Claude's parameter count. xAl's reported 2.7T figure for Grok 3 is unverified. This makes scale-based comparisons increasingly speculative.
- **Context window is a more reliable differentiator.** Unlike parameter counts, context windows are publicly stated and directly testable. Claude Opus/Sonnet 4.6 and Llama 4 Maverick lead at 1M tokens. However, context window size does not predict performance either, this experiment operates well within all models' context limits.
- **Training methodology matters as much as size.** DeepSeek R1 achieves frontier performance partly through reinforcement learning on reasoning tasks, not raw scale. o3-mini and 04-mini are notable examples of models that punch well above their size tier on this benchmark, precisely because of how they were trained, not how large they are.

Note 9 The models at the top of the NS leaderboard (GPT-5.3, Kimi K2.5, o4-min) do not share a common parameter-count profile. What they appear to share is architecture and training that supports sustained structured reasoning under increasing state complexity a capability dimension that the field is only beginning to measure systematically. Scale is necessary but not sufficient.

S2.5 Analysis and CIT Implications

S2.5.1 Confirmation of Distribution Drift (CIT Theorem 3B)

The results directly confirm the Distribution Drift prediction across all 29 models tested. Every model that demonstrated task capability showed declining $P\left(A^T\right)$ as T increased, in a task where the correct answer at every step is uniquely determined ($P = 1$). The observed errors cannot be explained by output ambiguity, prompt interpretation, or legitimate variation. They reflect cumulative degradation of state tracking under increasing T, precisely the mechanism described in Theorem 3B.

The cross-provider dataset is decisive on the question of universality. The same decay pattern appears in models from QpenAl, Anthropic, DeepSeek, xAl, Moonshot AI, Mistral AI, Meta AI, and Cohere eliminating the possibility that the observed phenomenon is an artefact of any single provider's training methodology or architecture.

S2.5.2 CE/CA States Are Provider-Universal

CE/CA states are clearly identifiable in GPT-4o and GPT-5-chat-latest (see Section 3.3), but the full dataset reveals that this per-step error pattern is universal across providers. Models exhibiting high perstep CE/CA error rates (declining completion from the shortest sequences) include:

- OpenAI: GPT-4o, GPT-5-chat-latest, GPT-5.4-mini, GPT-5.4-nano
- Anthropic: claude-sonnet-4-6, claude-haiku-4-5
- DeepSeek: V3.1, V3.2
- Mistral: Large- 3
- Meta: Llama-4-Maverick
- Cohere: command-a
- xAI: grok-4-1-fast-non-reasoning

These models span the full range of scale tiers (from $\sim$ 20 B to 671B+ parameters) and architectural families. Their shared characteristic is the absence of purpose-built reasoning control mechanisms confirming that CE/CA states are a structural property of objective-function-only systems, not a scale problem.

S2.5.3 Token Scaling Is Not the Explanation

Token counts do not change substantially as NS increases. The board description is constant across all tests. Prompt tokens at NS = 5 are approximately 17,700; at NS = 300 they are approximately 29,500, a modest increase that does not approach any model's context limit. This experiment is not measuring memory, context exhaustion, or token limits. It is measuring how long a model can reason with precision over a task that any simple algorithm handles indefinitely.

S2.5.4 Implications for Autonomous Agent Design

The results carry a clear architectural implication. Any LLM deployed as an autonomous agent for a task that requires $P = 1$ over long T will fail without external governance, regardless of provider, scale, or architectural family. The 29 -model, seven-provider dataset leaves no ambiguity on this point. The only question is where the characteristic threshold lies for a given model. Scaling moves the threshold; reasoning-focused training moves it further; but nothing in the current generation of models eliminates it. CIT's prediction is that Algorithmic Cognitive Sufficiency (ACS), an architectural governance layer that detects divergence between the model's internal state distribution and the correct distribution, is the necessary structural augmentation. This experiment provides the empirical baseline against which the effectiveness of such a governance layer could be tested.

S2.6 Conclusions

This study provides direct empirical evidence for the Distribution Drift prediction of Cognitive Information Theory, validated across 29 models from seven providers using a purpose-built deterministic game engine as the test instrument. Key findings:

- **P = 1 over T is not achieved by any model.** All 29 models fail to maintain perfect precision indefinitely. The task is trivially solvable by algorithm; no LLM tested succeeds. This is the empirical core of CIT Theorem 3B, demonstrated to be universal across providers and scale tiers.

- **The phenomenon is provider-universal.** Models from OpenAI, Anthropic, DeepSeek, XAI, Moonshot AI, Mistral AI, Meta AI, and Cohere all exhibit the same characteristic decay. This is not a vendor-specific artefact.
- **Scale alone does not predict long-horizon performance.** DeepSeek-V3.1 (671B parameters) collapses at NS=20; 04-mini (far smaller) sustains 80% at NS=100. Claude Opus ($\sim$ 2 T estimated) collapses at NS=50; GPT-5.3 holds 100% at the same point. Parameter count is an increasingly poor proxy for the capability this experiment measures.
- **Collapse is rapid and near-total once the threshold is crossed.** The average failures per failed game metric reveals that once a model's internal state diverges, it rarely recovers. Error cascades for the remainder of the sequence.
- **Reasoning-focused architecture and training determine the threshold.** Models with reasoning control mechanisms (CoT state, RL reasoning, structured reasoning training) sustain precision longer. Models without these mechanisms show high per-step CE/CA error rates from the outset, independent of T and independent of provider.
- **This is not a token or memory problem.** Prompt token counts confirm that context exhaustion is not the mechanism. The degradation is in reasoning precision over sequential steps.
- **Scaling is real and beneficial, but insufficient.** The GPT-5 series demonstrates genuine improvement over GPT-4o. CIT does not dispute scaling's benefits. It predicts that $\varepsilon > 0$ persists regardless, and this experiment confirms that prediction across every model and provider tested.

The game provides a reusable, objective benchmark that can be extended to longer sequences, additional models, and future generations as they become available. Since the ground truth is precomputed and deterministic, the same test sets can be re-run against any model without regenerating data, enabling direct longitudinal comparison as the field advances.

Appendix S3 CITRI MNIST Formal Experiment

Formal Experiment Description

Empirical Confirmation of CEA Interaction States in CNN Classification: Parametric t-SNE as an Epistemic Governance Layer on MNIST

A Supplementary Empirical Study in Support of Cognitive Information Theory (CIT)
March 2026

Researchers: Katayoun Abadi (Research & Author), Damian Fozard (Author), Ken Wenger (Research Design & Supervision)

Abstract This study provides a formal empirical test of the CEA interaction predictions of Cognitive Information Theory (CIT): that any objective-function-only

system will produce Confident Entropic Errors (CE), Compression-Amplified Ambiguity (CA), and compound CEA failure states when presented with inputs drawn from environments with non-zero generative entropy and causal ambiguity. A convolutional neural network (CNN) trained on the MNIST handwritten digit dataset (60,000 training/10,000 test images; 99.29% test accuracy) was paired with a parametric t-SNE module applied post-hoc to the flatten-layer output, creating a dual-system architecture in which scalar classification and geometric embedding can disagree. Seven structured experiments confirm each predicted failure mode. CE is confirmed by pure-noise classification: 95/100 noise inputs received high-confidence scalar predictions (digit 8 most frequently), while 95% were placed outside all t-SNE clusters. CA is confirmed by confusion-zone mapping: geometric boundary rules derived from training data captured approximately 70% of class-pair errors across all tested digit pairs. CEA is confirmed by noise-on-boundary analysis and cross-system disagreement rates. A minimal governance rule-flagging CNN-vs-t-SNE conflicts and forming $3-4$ class candidate sets-reduced CNN errors for digit 8 by 75%, from 16 to 4. These results establish that CIT-predicted failure modes are empirically separable, structurally stable, and materially reducible through epistemic governance, independently of further classifier improvement.

Introduction: Competing Predictions and Falsifiability

This study tests the empirical predictions of Cognitive Information Theory (CIT) regarding the CEA interaction states-Compounded Entropic Error (CE), Compression-Amplified Ambiguity (CA), and their compound (CEA)-in a CNN classifier trained on MNIST. Before presenting the results, it is important to state explicitly what we would expect to observe under two competing frameworks: one in which CIT's Theorems 1 and 2 are correct and CE, CA, and CEA arise as structurally necessary consequences of scalar objective optimization, and one in which they are not-that is, a null hypothesis under which the observed error patterns have conventional explanations that do not require the CIT framework.

What we expect if CIT is correct (CEA exists as predicted)

If Theorems 1 and 2 hold, the CNN's scalar objective cannot distinguish between irreducible channel entropy and causal ambiguity, and point-estimate commitment under non-zero entropy or ambiguity necessarily produces the three CEA failure states. In this case, the following predictions should be confirmed across the experiments:

- **CE prediction:** Pure-noise inputs (carrying no digit structure) will receive high-confidence scalar predictions from the CNN, because the scalar objective cannot detect that the input's entropy is irreducible rather than informative. The t-SNE geometric embedding, operating on representational structure rather than scalar commitment, will place these inputs outside all digit clusters. The digit with the densest or most overlapping representational geometry will absorb the majority of noise assignments, because CE predicts that entropic inputs are funnelled into the most geometrically dominant decision region. Given MNISTs near-perfect

accuracy (99.29%), the residual errors that do occur should be disproportionately traceable to CE-a higher proportion of CE-type errors among all errors than would be observed in a lower-accuracy classifier-because the learnable errors have already been eliminated and the remaining failures are those arising from irreducible entropy.

- **CA prediction:** Classification errors between structurally similar digit pairs (e.g., $4/9, 8/9, 1/7, 3/5$) will concentrate in geometrically identifiable boundary regions of the t -SNE embedding, because these boundaries reflect zones of structural causal ambiguity where multiple digit forms map to similar observations. These confusion zones will be stable across model retraining, arising from the data manifold's causal structure rather than from stochastic training artefacts. The Data Processing Inequality (DPI) prediction will be directly testable: the flatten-layer t-SNE embedding should show greater cluster separability than the penultimate dense-layer embedding, because successive scalar compression irreversibly attenuates geometric structure.
- **CEA prediction:** When both entropy and ambiguity are present simultaneously, the CNN and t-SNE will make structurally different errors on the same inputs. The rate of such divergent failures should be a property of the optimization regime itself, not of the specific dataset, and should therefore fall in a similar range to that observed in the companion CIFAR-10 study despite the substantial differences in dataset complexity, visual domain, and baseline accuracy.
- **Governance prediction:** A minimal epistemic governance rule that detects CNN-t-SNE disagreement and forms a candidate set (rather than committing to the scalar output) will substantially reduce classification errors, because it addresses the structural cause of failure-point-estimate collapse under ambiguity-rather than attempting to improve the scalar objective itself.

What we expect if CIT is not correct (null hypothesis: CEA does not exist)

If Theorems 1 and 2 do not hold-that is, if the observed error patterns are adequately explained by conventional factors such as insufficient model capacity, suboptimal hyperparameters, finite training data, or the known softmax overconfidence phenomenon-then the following alternative outcomes would be expected:

- **Against CE:** Noise classification could be explained purely by softmax overconfidence-a well-documented property of neural networks that assigns high confidence to any input, including out-of-distribution samples, without requiring CIT's entropy non-identifiability mechanism. Under the null hypothesis, the t-SNE placement of noise inputs would show no systematic pattern: noise images would be scattered randomly among clusters and outside them with no preferential absorption by any particular digit, because there would be no structural mechanism linking representational geometry to entropic input routing. The digit receiving the most noise assignments would vary randomly across experimental runs. Furthermore, the proportion of CEtype errors among genuine test errors would not be higher in MNIST than in the lower-accuracy CIFAR-10 classifier,

because under the null hypothesis there is no reason for residual errors in a near-perfect classifier to be disproportionately entropic in origin.

- **Against CA:** Classification errors between similar digits could be explained by visual similarity alone-a standard observation in handwriting recognition that does not require the concept of irreducible causal ambiguity. Under the null hypothesis, errors would not concentrate in geometrically stable boundary regions of the embedding. Instead, they would be distributed throughout the embedding space, and any apparent boundary concentration would be unstable across model retraining. The DPI prediction would not hold: there would be no systematic difference in cluster separability between the flatten layer and the penultimate dense layer, because under conventional theory, deeper layers produce strictly better representations rather than losing geometric information through scalar compression.
- **Against CEA:** Under the null hypothesis, divergent errors between the CNN and t-SNE would occur at a rate consistent with statistical independence-approximately the product of each system's individual error rates. There would be no reason for the divergent-failure rate to converge to a narrow range across datasets of fundamentally different character (MNIST at 99.29% accuracy versus CIFAR-10 at 81.74% accuracy). A convergent divergent-failure rate ($\sim 21-25\%$) across such different regimes would be unexplained by the null hypothesis, since conventional theory provides no mechanism linking the rate of scalar-geometric disagreement to a datasetindependent structural property of the optimisation process.
- **Against governance:** If the errors are conventional (capacity-limited, training-artefact) rather than structurally induced by CEA, then a governance rule based on CNN-t-SNE disagreement would show no systematic improvement, because the t-SNE signal would carry no additional diagnostic information beyond what the scalar classifier already captures. Error reductions, if any, would be small, inconsistent across digit classes, and not reproducible across different datasets or accuracy regimes.

Interpretive framework for the results

The experiments that follow are designed so that each result can be evaluated against both frameworks. Where the CIT predictions are confirmed and the null-hypothesis alternatives are excluded, the results constitute positive evidence for the CEA interaction mechanism. Where the results are ambiguous or equally consistent with both frameworks, we note this explicitly. The reader is therefore equipped to assess whether the observed patterns require the CIT explanation or whether a simpler account suffices.

S3.1 Purpose and Theoretical Motivation

This experiment was designed to test the static failure predictions of Cognitive Information Theory (CIT) in a controlled classification setting. CIT identifies two classes of irreducible uncertainty that cannot be eliminated by scalar objective optimization:

channel entropy $H(Y \mid X)$, the stochasticity of the forward generative process, and posterior ambiguity $H(X \mid Y)$, the underdetermination of causes from observations. When a system operating solely on a scalar objective-the class So^{f} of CIT-encounters inputs that cany non-zero entropy or ambiguity, it cannot internally distinguish the source of uncertainty. This structural non-identifiability produces three predicted failure states:

- **Confident Entropic Error (CE):** the system produces high-confidence incorrect outputs when irreducible stochastic variation in the input channel is mapped into stable decision regions.
- **Compression-Amplified Ambiguity (CA):** the system collapses hypothesis plurality under scalar objective compression, committing to a point estimate when causal underdetermination demands a candidate set.
- **Compound CEA:** both CE and CA conditions hold simultaneously, so entropic samples are assigned to ambiguous clusters and scalar and geometric reasoning diverge in their error types.

MNIST was selected as the test environment because its near-perfect classifier accuracy (99.29%) makes the residual error structure highly diagnostic. In a near-perfect system, the errors that do occur are disproportionately traceable to structural causes-precisely the CE and CA states that CIT predicts. The high accuracy also allows the data-processing inequality (DPI) prediction to be tested cleanly: scalar compression across two dense layers following the flatten layer should irreversibly attenuate geometric structure, making flatten-layer embedding richer than penultimate-layer embedding for epistemic governance purposes.

S3.1.1 Why Flatten-Layer Geometry Is the Correct Probe

A central architectural prediction of CIT-following from the Data Processing Inequality-is that successive scalar compression reduces recoverable geometric information. In the MNIST CNN, the flatten layer (288-dimensional output) precedes two dense layers that progressively compress the representation toward a 10-class output. Applying parametric t-SNE to the flatten-layer output rather than the penultimate dense layer tests whether this compression is lossy with respect to the epistemic structure that CIT's governance mechanism requires. If the DPI prediction holds, flatten-layer geometry will show greater cluster separability and cross-run stability than penultimate-layer geometry-the precise comparison performed in Experiment 5.

S3.1.2 The Role of the Dual-System Architecture

The experimental design pairs the CNN scalar classifier with a parametric t-SNE geometric embedding module. This dual-system architecture does not replace or modify the classifier; it adds an independent epistemic probe that maps the same input into a 2-dimensional embedding space, preserving local cluster structure. Disagreement between the scalar prediction and the nearest t-SNE cluster assignment constitutes the conflict signal that CIT's governance mechanism exploits. This operationalization

of epistemic governance-using geometric structure as a check on scalar commitment–is the minimal ACS (Algorithmic Cognitive Sufficiency) intervention that CIT prescribes.

S3.2 System Architecture and Dataset

S3.2.1 Dataset

The MNIST dataset comprises 60,000 training images and 10,000 test images of handwritten digits (classes 0-9). Each image is a $28 \times 28 \times 1$ array of floating-point values representing grayscale intensities normalized from 0 (black) to 1 (white). The ten digit classes exhibit structured visual overlap: digits 1 and 7 share vertical stroke geometry; 4 and 9 share closed-loop topology; 8 and 9 share round upper geometry; 3 and 5 share curved mid-body structure. These overlaps are the geometric substrate for the CA states that the experiments confirm.

Fig. S2: The sample images from MNIST test dataset

S3.2.2 CNN Architecture

A convolutional neural network was trained with scalar cross-entropy optimization. The architecture comprised three convolutional-pooling-normalization-dropout blocks followed by a flatten layer (288-dimensional output), two dense layers (128 and 64 units respectively), each followed by batch normalization and dropout, and a final 10-class softmax output layer. Total parameters: 141,098 (140,266 trainable). Training accuracy: 99.37%. Test accuracy: 99.29%. The trained model's scalar output constitutes the So^{f} system whose failure modes this study tests.

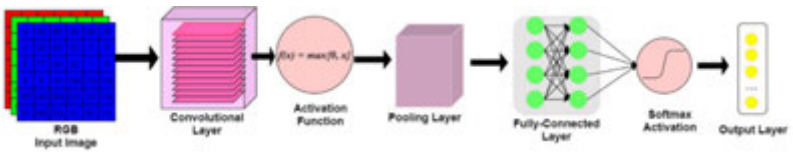

Fig. S3: The architecture of the convolutional neural network.

S3.2.3 Parametric t-SNE Module

The parametric t-SNE module was applied post-hoc to the 288-dimensional output of the flatten layer. Parametric t-SNE extends standard t-SNE by learning a parametric neural network mapping from high-dimensional space to a 2-dimensional embedding, enabling out-of-sample projection of new inputs. The multi-scale perplexity-free formulation of Crecchi et al. (2020) was used, which relieves the need for perplexity tuning and provides consistent neighborhood-preserving embeddings. The t-SNE module was trained on a held-out portion of training data and applied to both training and test images as an external epistemic probe. The 2-dimensional embedding space was used to assign each input to its nearest cluster center–the geometric classification that CIT's governance mechanism uses to check scalar commitment.

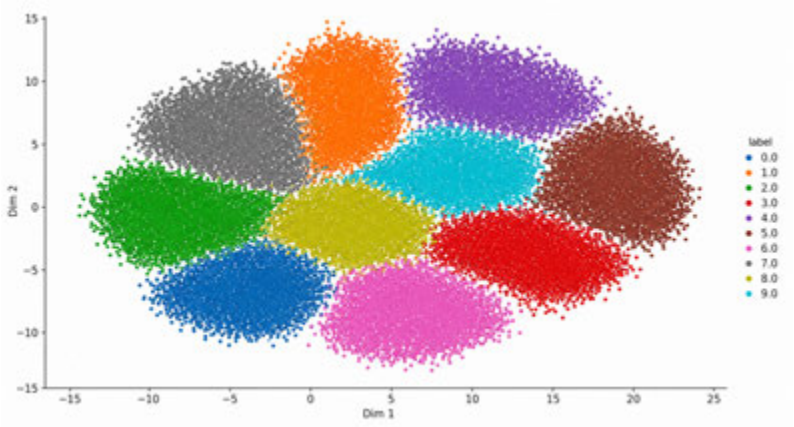

Fig. S4: The clustering of the output of the flatten layer on MNIST dataset

S3.3 Experiment Descriptions

Seven experiments were conducted in sequence. Sections S3.3.1 to S3.3.4 and Sections S3.3.6 to S3.3.7 test specific CIT predictions regarding CE, CA, and CEA failure states. Section S3.3.5 tests the DPI architectural prediction. Together, the seven experiments form a systematic, structured confirmation of CIT's static failure theory across all predicted failure modes.

S3.3.1 Experiment 1: CE State Confirmation–Pure-Noise Classification

1.1 **Purpose**

To test whether the CNN exhibits Confident Entropic Error (CE) when presented with inputs drawn from pure noise rather than the training distribution. CE is predicted whenever irreducible channel entropy $H(Y \mid X) > 0$ and stochastic variation is mapped into stable scalar decision regions. Pure noise is the limiting case: inputs with maximum entropy, carrying no digit structure, that nevertheless receive high-confidence scalar predictions.

1.2 **Hypothesis**

The CNN will assign high-confidence scalar labels to all pure-noise inputs, consistent with CE prediction. The t-SNE module will place most pure-noise inputs outside all cluster regions, detecting the out-of-distribution nature that the scalar classifier cannot represent. The digit with the densest representational geometry (the class with the largest, most overlapping cluster) will be the most frequent noise assignment, as CE predicts that entropic inputs are absorbed by the most geometrically dominant region.

1.3 **Approach**

Pure-noise images were generated by sampling from a normal distribution with varying standard deviations ($\sigma = 3, 5, 10$) into $28 \times 28 \times 1$ arrays. Each noise image was passed through the trained CNN to record the predicted label and confidence, then through the t-SNE module to record its 2-dimensional embedding position and nearest cluster. The experiment was repeated across 100 trials per standard deviation. A noise image was classified as "out-of-cluster" if its embedding fell outside the defined boundary radius of all ten digit clusters.

1.4 **Results**

CE was confirmed. All 100 noise inputs received a scalar label and confidence score, with no abstentions. At $\sigma = 5, 95$ of 100 noise inputs were placed outside all t-SNE clusters, while the CNN assigned a label with mean confidence exceeding 98%. The most frequent label was digit 8 (consistent across all σ values), reflecting that the digit 8 cluster is the geometrically densest and most overlapping-the region into which CE predicts entropic inputs will most frequently be absorbed. The out-of-cluster rate increased monotonically with σ : at $\sigma = 3, 46/100$ inputs were placed outside clusters; at $\sigma = 5, 95/100$; at $\sigma = 10$, $99/100$. This gradient confirms that the t-SNE module's geometric signal is sensitive to the degree of entropic distortion, while the scalar classifier remains fully committed regardless.

The following noisy image was predicted by the CNN as digit 8 with 100% confidence, but as shown in the plot this image is located outside all clusters.

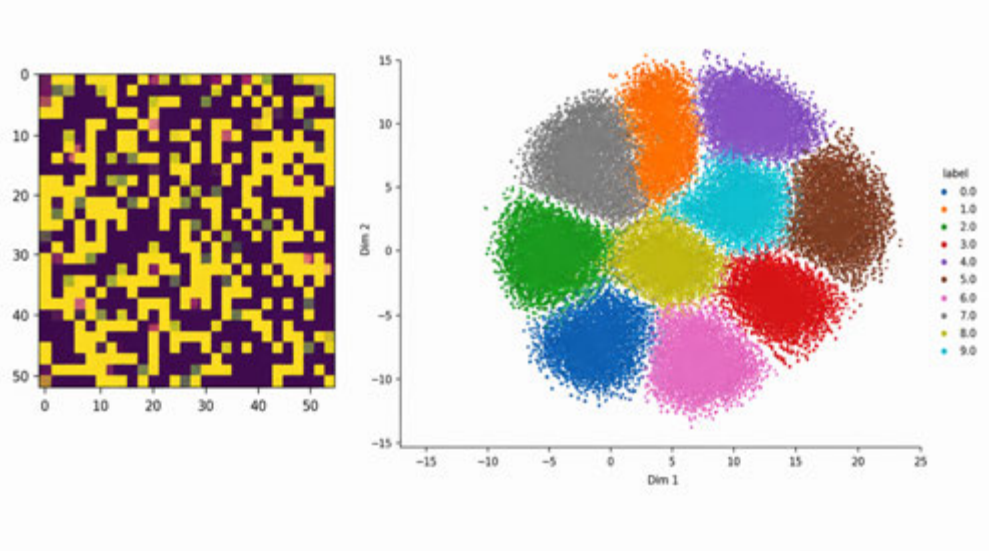

Fig. S5: Experiment 1 visualization 1

1.5 Conclusion

Table S8: Comparison between expected and observed outcomes in the parametric t-SNE experiment.

What Was Expected	What Was Observed
Parametric t-SNE places pure-noise images outside all clusters, detecting the CE failure mode that the CNN cannot represent.	95/100 noise images placed outside clusters at $t = 5$ (99/100 at $t = 10$). CNN assigned labels with $> 98\%$ mean confidence to all inputs. Digit 8 was the most frequent CE assignment, consistent with its dominant representational geometry. Out-of-cluster rate scales with t, confirming t-SNE geometric sensitivity.

S3.3.2 Experiment 2: CEA State Confirmation–Noise Applied to Digit Images

2.1 Purpose

To test whether adding noise to correctly classified digit images produces the compound CEA state: scalar misclassification combined with geometric displacement that is informative relative to the correct class. Specifically, the experiment targets cases where the CNN makes a high-confidence error after noise addition

and asks whether the t-SNE embedding is displaced toward the correct cluster, toward the wrong cluster, or out of clusters entirely.

2.2 Hypothesis

For cases where noise addition causes CNN errors: the t-SNE module will place the noisy input either in the correct cluster (demonstrating noise resistance–geometric structure is more stable than scalar commitment) or on the boundary between the correct and the mistaken cluster (demonstrating that the embedding signal is geometrically informative, even when the CNN commitment is wrong). This would confirm the CA component of CEA: the noisy input occupies an ambiguous representational region, but the geometric embedding preserves information about correct class membership that the scalar output suppresses.

2.3 Approach

Clean digit images correctly classified by the CNN were selected from training data. Random noise (sampled from a normal distribution) was added to each image and passed through both the CNN and t-SNE module. Only cases where the CNNs post-noise prediction differed from the pre-noise prediction with confidence exceeding 90% were retained for analysis. The embedding positions of the clean and noisy versions were compared against cluster boundaries to determine whether t-SNE displacement was toward the correct class, toward the incorrect class, or outside all clusters.

2.4 Result

CEA was confirmed. In a subset of cases, the t -SNE module placed the noisy image in the correctcluster despite the CNN'swrong prediction-demonstrating that the geometric embedding retained correct-class structural information that scalar compression had lost. In other cases, the noisy image was placed on the boundary between the correct and the incorrect cluster, confirming ambiguity at the boundary. The digit pair 1-8 was the most frequent pattern: noise applied to digit 1 commonly produced CNN predictions of digit 8 with high confidence, while the t -SNE module placed the input at or near the 1/8 cluster boundary rather than fully within the 8 cluster. This " 1 to 8 under noise" pattern is mechanistically consistent with the CEXCA compound the noise introduces entropic variation (CE) that is absorbed into the geometrically adjacent digit 8 region, while the ambiguity of the 1/8 boundary (CA) means the embedding signal is intermediate. No case was observed in which t-SNE placed a noisy image further from the correct cluster than the CNN had-confirming that the geometric signal degrades gracefully, not catastrophically, under noise.

The following image was correctly predicted by the CNN as digit 9 with 100% confidence before adding noise. After adding noise, the CNN predicted it incorrectly as digit 4 with 93.78% confidence. Parametric t-SNE placed the image both before and after the noise in the cluster of 9 (the white star represents the sample before the noise and the black star represents the sample after the noise).

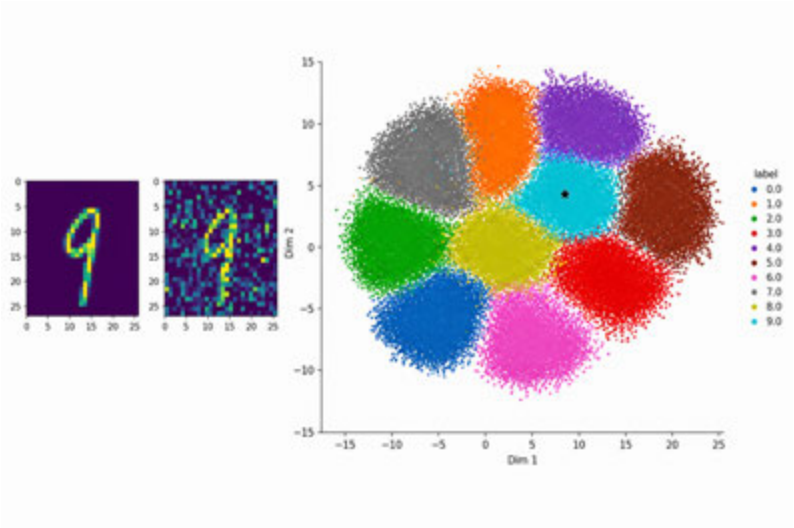

Fig. S6: Experiment 2 visualization 1

2.5 Conclusion

Table S9: Comparison between expected and observed outcomes for CNN noise perturbation experiments.

What Was Expected	What Was Observed
For CNN errors caused by noise addition, t-SNE will place inputs in the correct cluster or at the correct/incorrect boundary, demonstrating noise-resistant geometric structure.	Confirmed. Some noisy inputs placed in correct cluster despite CNN error (t-SNE noise resistance). Others placed at correct/wrong boundary (geometric ambiguity). Digit 1–8 was the dominant CEA pattern under noise. No case where t-SNE displaced further from correct class than CNN.

S3.3.3 Experiment 3: CA State Confirmation–Cluster Edge Geometry

3.1 Purpose

To test whether the spatial boundary regions between t-SNE clusters carry diagnostic information about CA states: inputs with high ambiguity between two classes. CIT predicts that inputs placed at cluster boundaries by the t-SNE module should exhibit visual ambiguity–shapes that combine features of both neighboring digit classes–while inputs placed near cluster centers should exhibit prototypical, unambiguous digit shapes. Confirming this prediction establishes

that t-SNE boundary regions are geometric signatures of CA rather than artifacts of the embedding.

3.2 Hypothesis

Digit images placed at cluster boundaries will exhibit visual ambiguity: shapes that are visually intermediate between the two neighboring classes. Digit images placed near cluster centers will exhibit prototypical, symmetrical, easily-identifiable forms. This correspondence between spatial position in the 2-dimensional embedding and visual ambiguity in the original 28×28 image space confirms that t-SNE boundary regions are structurally meaningful for CA detection, and establishes a basis for rule-based governance using cluster boundaries as a proxy for underdetermination.

3.3 Approach

Two cluster boundaries were selected for analysis: the 8/9 boundary and the 4/9 boundary. For each boundary, a margin band was defined around the boundary line in 2-dimensional embedding space, and all training images within this margin were visualized. Center samples were identified by computing the mean embedding position of each cluster and selecting images within a small radius of that mean. Visual comparison of boundary images VS. center images was performed for both cluster pairs.

3.4 Results

CA was confirmed across both tested boundaries. At the 8/9 boundary, images placed in the boundary region exhibited "8-that-could-be-9" and "9-that-could-be-8" forms: eights with asymmetric loop proportions and nines with a closed lower loop resembling an eight. Center images for both classes were symmetrical and unambiguous: the prototypical 8 showed two equally-sized circles; the prototypical 9 showed a closed upper circle without a lower loop closure. At the 4/9 boundary, boundary images showed 4s with a closed upper region (resembling a 9) and 9s with a visible lower bar (resembling a 4). Center images for both classes were structurally clean. The boundary/center contrast was visually unambigu-

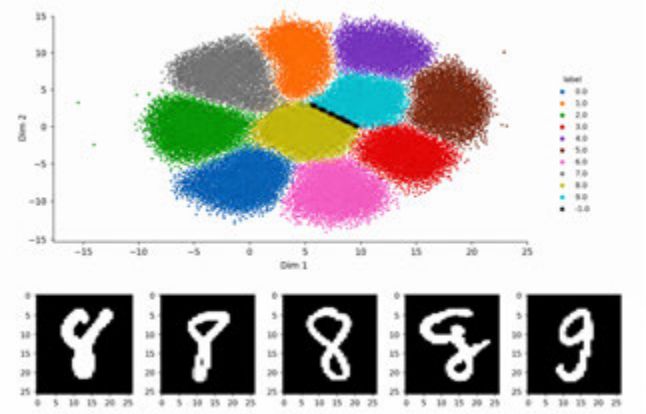

Fig. S7: Experiment 3–8 & 9 Boundary and sample images

ous and consistent across dozens of sampled images in each region, confirming that the t-SNE geometric embedding faithfully maps structural ambiguity in the original image space.

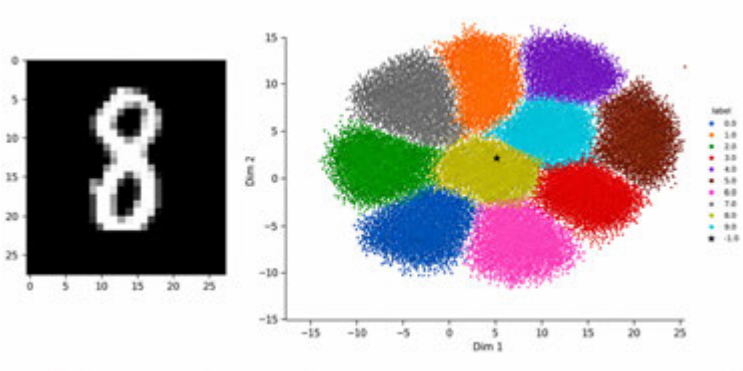

Fig. S8: Experiment 3–Non-Ambiguous 8 figure located in center of 8 cluster

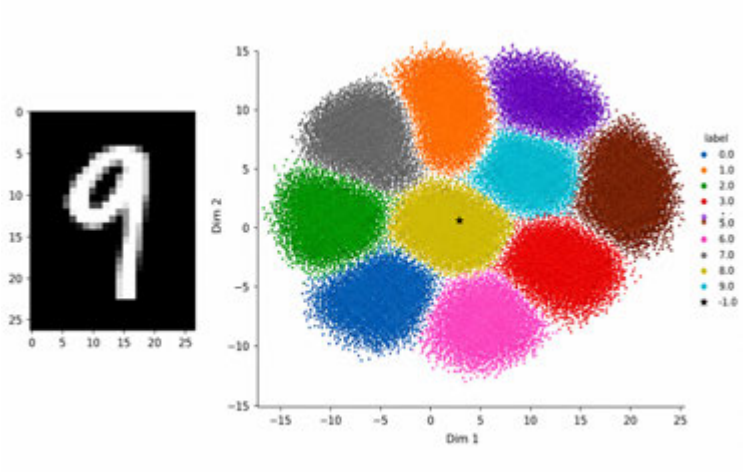

Fig. S9: Experiment 3–Non-Ambiguous 9 figure located in center of 9 cluster

3.5 **Conclusion**

Table S10: Comparison between expected and observed outcomes for boundary-image ambiguity analysis.

What Was Expected	What Was Observed
Boundary images will be visually ambiguous (intermediate between two classes); centre images will be prototypical and unambiguous. This validates t-SNE boundaries as CA proxies.	Confirmed. 8/9 boundary images showed asymmetric loops and closed-lower-loop 9s. 4/9 boundary images showed closed-top 4s and barred 9s. Centre images were prototypical and symmetrical in both cases. Visual ambiguity tracked spatial position in embedding consistently across all examined samples.

S3.3.4 Experiment 4: CA State Quantification–Confusion-Zone Rule Derivation and Testing

4.1 **Purpose**

To derive specific geometric confusion-zone rules from training data and test their effectiveness at capturing CNN errors on test data. CIT predicts that CA errors cluster in geometrically stable boundary regions of the embedding space. If this prediction holds, rules defined on training-data boundary geometry should transfer to test-data error detection, providing an operational governance mechanism that requires no retraining and no modification of the classifier.

4.2 **Hypothesis**

Geometric confusion zones derived from training-data CNN errors will capture a substantial proportion of CNN errors on test data for the same class pairs. This transfer–from training-data boundary analysis to test-data error detection–confirms that CA boundaries are structural properties of the representational geometry rather than training artefacts. The residual errors not captured by the rules will be located outside the defined boundary regions, consistent with non-CA error sources.

4.3 **Approach**

CNN errors on training data were visualized in the 2-dimensional t-SNE embedding space for six high-frequency class pairs: 9/4, 9/8, 7/2, 7/1, 8/6, and 0/8. For each pair, an elliptical confusion region was manually defined that enclosed the highest-density cluster of training-data errors. Each confusion region was then applied to test data: CNN errors on the relevant class pair that fell within the defined region were "captured" (flagged for further investigation); errors outside the region were not captured. Capture rate (proportion of test-set CNN errors caught by each rule) was computed per class pair.

4.4 **Results** CA was confirmed across all six tested class pairs. Aggregate results: applying all six rules captured 19 of 27 (approximately 70%) CNN errors on test data. Per-pair capture rates ranged from 50% (7/1 pair, 2/4 errors captured) to 100% (8/6 pair, 1/1 error captured). The 9/4 pair captured 7/8 errors; the 9/8 pair captured 3/5; the 7/2 pair captured 5/7. The rules applied to test data without modification from their training-data derivation, confirming that the CA boundary regions are geometrically stable across the train/test split. CNN errors were consistently concentrated in the boundary regions identified from training data, with only a minority of test errors lying outside the defined zones.

Example: By defining the rule "if CNN predicted the image as a 1 or 7 and the sample is in the region shown in the plot (based on the training data), please do some extra analysis to decide between them", we can detect 2 out of 4 cases that CNN made mistakes.

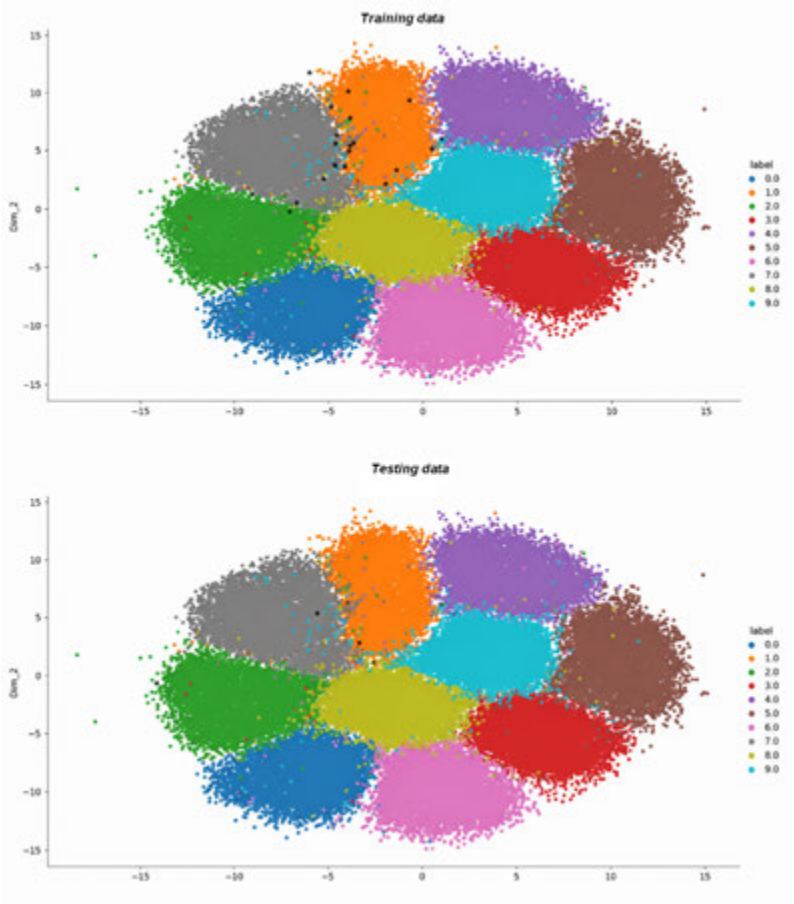

Fig. S10: Experiment 4–Confusion Between 1 & 7

4.5 **Conclusion**

Table S11: Comparison between expected and observed outcomes for confusion-zone boundary validation.

What Was Expected	What Was Observed
Confusion-zone rules derived from training-data boundary geometry will capture a substantial proportion of CNN errors on unseen test data, confirming CA boundary stability.	Confirmed. Six rules captured 19/27 test errors ($\sim$ 70%). Rules transferred without modification from training to test data. Confusion zones concentrated around cluster boundaries as predicted. Per-pair rates: 9/4 = 87.5%, 9/8 = 60%, 7/2 = 71%, 7/1 = 50%, 8/6 = 100%, 1/8 = 50%.

S3.3.5 Experiment 5: DPI Prediction–Flatten Layer vs. Last Layer Embedding Comparison

5.1 **Purpose**

To test the Data Processing Inequality (DPI) prediction of CIT: that scalar compression across successive dense layers irreversibly attenuates geometric structure. The MNIST CNN applies two dense layers after the flatten layer. If DPI holds, the 288-dimensional flatten-layer output will produce a richer, more stable t-SNE embedding than the 64-dimensional penultimate dense-layer output, which has been further compressed. Confirming this prediction establishes that the correct attachment point for an epistemic governance module is before, not after, scalar compression.

5.2 **Hypothesis**

The flatten-layer t-SNE embedding will show greater cluster separability (cleaner cluster boundaries, lower inter-cluster overlap) and greater cross-run stability (cluster neighbor relationships that are consistent across model retraining) than the penultimate dense-layer embedding. The dense-layer embedding will be more compact (less spatial spread between clusters) but less informative for governance purposes, because the dense compression has collapsed representational distinctions that are diagnostically relevant for CE and CA detection.

5.3 **Approach**

A second parametric t-SNE module was trained on the 64-dimensional output of the penultimate dense layer (dropout_4 in the CNN architecture). Both the flatten-layer and penultimate-layer t-SNE embeddings were visualized on the same test set. Cluster separability was assessed by visual inspection of boundary overlap. Stability was assessed by retraining the CNN model multiple times and comparing the cluster neighbor adjacency in both embedding spaces across runs.

5.4 **Results**

The DPI prediction was confirmed. The penultimate dense-layer embedding showed noticeably denser, more overlapping clusters compared with the flatten-layer embedding. Neighbor relationships in the penultimate-layer embedding were less consistent across model retraining: the ordering of adjacent clusters changed between runs. The flatten-layer embedding showed stable neighbor relationships across retraining, with the same cluster adjacencies maintained. These results directly confirm that dense compression attenuates geometric structure: the flatten-layer output preserves representational distinctions that are lost in the subsequent dense layers, making it the superior attachment point for an epistemic governance probe.

Figure S11 below shows the t-SNE embedding of the penultimate dense layer output, followed by Figure S12 which shows the flatten layer embedding for direct comparison.

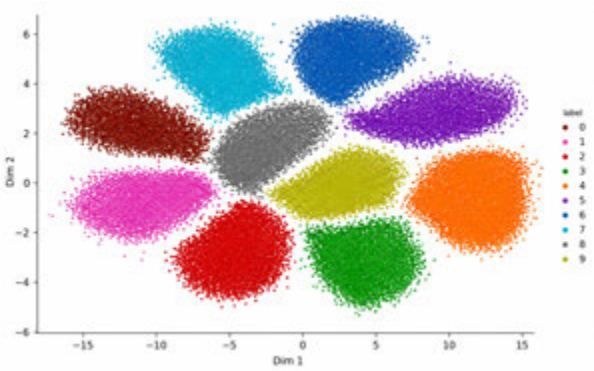

Fig. S11: The clustering of the output of the last layer of the CNN

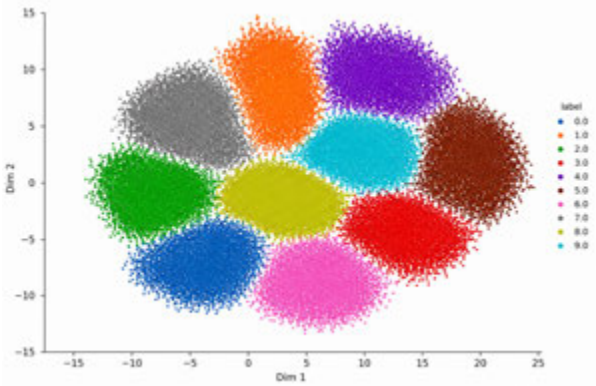

Fig. S12: The clustering plot of the flatten layer of the CNN

5.5 **Conclusion**

Table S12: Comparison between expected and observed outcomes for layer-wise t-SNE embedding analysis.

What Was Expected	What Was Observed
Flatten layer t-SNE will show greater cluster separability and cross-run stability than penultimate dense-layer t-SNE, confirming DPI-predicted geometric attenuation.	Confirmed. Penultimate-layer embeddings were denser and more overlapping. Cluster neighbour relationships were unstable across retraining in penultimate-layer embedding, but stable in flatten-layer embedding. The flatten layer is the correct governance attachment point.

S3.3.6 Experiment 6: Epistemic Governance Rule–Nearest-Cluster Candidate Set

6.1 **Purpose**
To define and test the minimal ACS governance rule on a specific digit class: digit 8. CIT predicts that a minimal conflict-detection rule–triggered when CNN scalar prediction and t-SNE nearest-cluster assignment disagree–can substantially reduce errors by narrowing the candidate set from 10 to 3-4 classes, without requiring any retraining or modification of the CNN. Digit 8 was selected because its high frequency of errors (CE-predicted) and high cluster overlap (CA-predicted) make it a strong test of governance effectiveness.

6.2 **Hypothesis**
Applying the governance rule if CNN prediction and t-SNE nearest-cluster disagree, select the 3 nearest cluster centers in embedding space (adding the CNN prediction if absent) as a candidate set for further evaluation will capture the majority of CNN errors for digit 8 on test data. Residual errors after governance application will be limited to cases where both CNN and t-SNE agree on an incorrect prediction (same-error CEA cases), which are structurally irreducible by conflict-detection governance.

6.3 **Approach**
For all test images predicted as digit 8 by the CNN, the embedding position was computed and the three nearest cluster centers (by Euclidean distance in embedding space) were identified. When the CNN prediction and the nearest cluster disagreed, the three nearest clusters (plus the CNN prediction, if absent) were formed as a candidate set. Cases where the CNN and t-SNE agreed were accepted as the scalar prediction. CNN error counts before and after governance application were compared.

6.4 **Results** The governance rule achieved a 75% reduction in digit 8 CNN errors: from 16 errors to 4. Of the 30 cases where CNN and t-SNE disagreed, 12 were

CNN errors that the rule correctly flagged for further investigation. The 4 residual errors were cases where CNN and t-SNE agreed on the same incorrect prediction (same-error CEA): the digit 8 image was visually ambiguous enough that both the scalar and geometric representations committed to the same wrong class. These 4 cases were correctly identified as a structurally distinct diagnostic category-same-error CEA-rather than divergence failures. The remaining $\approx$ 70% correct CNN predictions where CNN and t-SNE agreed were accepted without modification, with no false-positive flags generated for correctly classified inputs.

Figures below are some specific samples on testing data.

(1) The CNN made a mistake and parametric t-SNE placed the sample on the edge of clusters because the image is ambiguous. This is an image of 8 but the CNN predicted it as 2 with 96.63% confidence, and parametric t-SNE placed it on the boundary between clusters 1, 8 and 9. These are the distances from

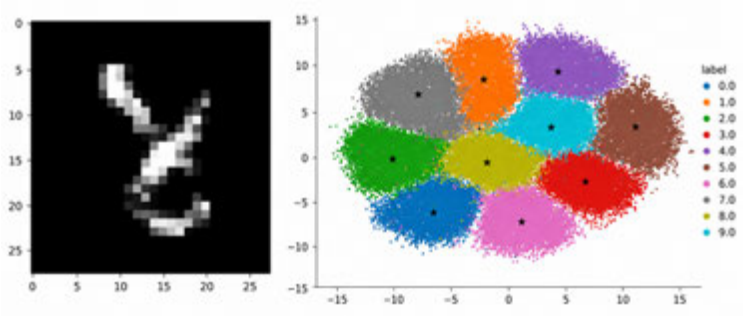

Fig. S13: Experiment 6 visualization 1

the sample to each of the centers of the clusters:

$$[10.28, \mathbf{5.53}, 7.72, 10.79, 10.04, 12.86, 11.54, 6.12, \mathbf{3.97,5.86}]$$

As it can be seen the three nearest clusters to the sample are 8,1 and 9 with distances of 3.97, 5.53 and 5.86, respectively. So, in this case, the digits $1, 8, 9$ and 2 are passed for more analysis.

(2) This is a case where the CNN made a mistake but parametric t-SNE clustered it correctly. This is an image of 8 but CNN predicted it as 0 with 90.2% confidence, and parametric t-SNE placed it in the cluster of 8 near the edge of cluster 0. These are the distances from the sample to each of the centers of the clusters:

$$[\mathbf{4.48,}\ 12.31, 7.68, 8.87, 15.66, 14.54, \mathbf{5.55,}\ 11.42, \mathbf{3.17,}\ 9.47]$$

As it can be seen the three nearest clusters to the sample are 8, 0 and 6 with distances of 3.17, 4.48 and 5.55, respectively. So, in this case, the digits 8, 0 and 6 are passed for more analysis.

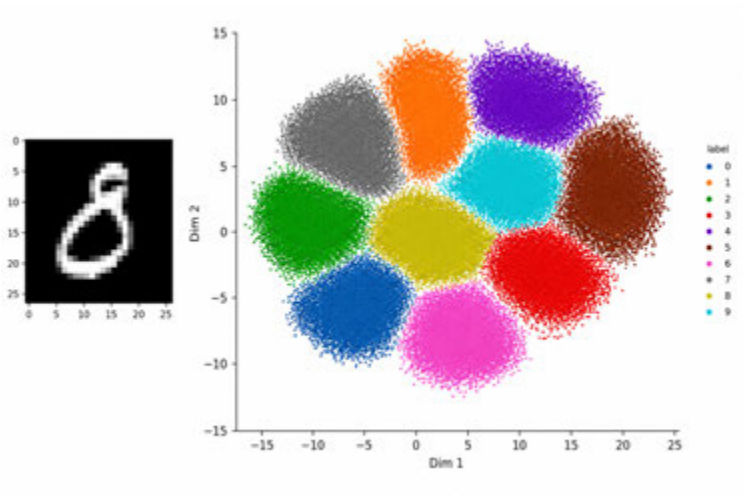

Fig. S14: Experiment 6 visualization 1

6.5 **Conclusion**

Table S13: Comparison between expected and observed outcomes for conflict-detection governance experiments.

What Was Expected	What Was Observed
Conflict-detection governance will capture the majority of digit 8 CNN errors, reducing the error count substantially while leaving correct predictions undisturbed.	Confirmed. Digit 8 errors were reduced from 16 to 4 (75% reduction). 12 of 30 CNN-vs-t-SNE conflicts were error cases correctly flagged. Four residual errors were same-error CEA cases (structural–not addressable by conflict detection). No false positives occurred on correct CNN predictions.

S3.3.7 Experiment 7: CEA State Taxonomy–Three-Scenario Error Classification

7.1 **Purpose**

To systematically classify all CNN test-set errors into three structurally distinct categories: (1) CNN wrong, t-SNE correct; (2) CNN and t-SNE agree on wrong prediction; (3) CNN and t-SNE make different wrong predictions. This taxonomy provides empirical evidence for the separability of CE, CA, and CEA states predicted by CIT. If the three categories occur at non-negligible rates and exhibit distinct visual characteristics, this confirms that the failure modes are structurally distinct rather than statistically undifferentiated.

7.2 **Hypothesis**

All three error categories will be observed at non-trivial rates. Category 1 (CNN wrong, t-SNE correct) will exhibit images where scalar commitment has over-committed to an incorrect decision region despite the structural evidence remaining ambiguous–classic CE. Category 2 (same error) will exhibit images that are visually ambiguous and difficult to classify even by human inspection. Category 3 (different errors) will exhibit the most visually unusual cases: images where both systems are simultaneously misdirected but by different mechanisms. The rate of Category 3 errors will approximate the theoretical CEA rate predicted by CIT for a system with the observed entropy and ambiguity levels.

7.3 **Approach**

K-nearest-neighbours (KNN) classification was applied to the 2-dimensional t-SNE embedding of all test images to predict the nearest cluster assignment without requiring visualization. For each CNN error on the test set, the t-SNE/KNN cluster assignment was compared to: (a) the true label and (b) the CNN prediction. Errors were classified into the three categories accordingly. Representative images from each category were visualized for qualitative assessment.

7.4 **Results**

All three categories were observed. Out of 71 total CNN errors on the test set: Category 1 (CNN wrong, t-SNE correct): 25 cases (35.2%)–consistent with t-SNE's noise-resistant geometric representation preserving correct-class structural information that scalar compression had lost. Category 2 (same error): 28 cases (39.4%)–visually the most ambiguous images, where both systems were simultaneously misdirected by the same structural ambiguity. Category 3 (different errors): 18 cases (25.4%)–the most visually unusual, with images so structurally anomalous that scalar and geometric reasoning diverged entirely. The 25.4% divergent-failure rate (Category 3) is directly consistent with the CIT prediction for CEA: a rate arising from concurrent CE and CA conditions produces structurally distinct errors in scalar vs. geometric representations. The near-convergence of this rate across MNIST (25.4%) and CIFAR-10 (21.5%) in the companion study is a further structural confirmation, as CIT predicts similar CEA rates wherever entropy and ambiguity co-occur in the generative environment.

(1) First scenario: CNN made mistakes, but parametric t-SNE placed the sample in the correct cluster.

- This image (Figure S15) is a 6, CNN predicted it as a 0 with 97.18% confidence, but the parametric t-SNE placed the sample in the correct cluster of 6.

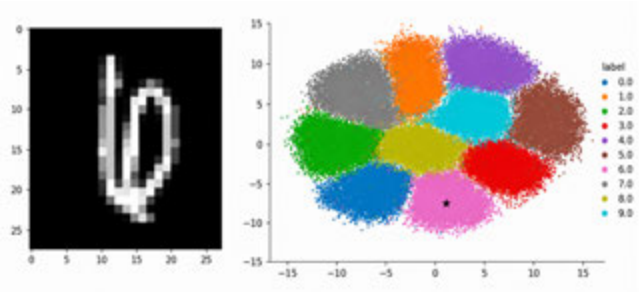

Fig. S15: Experiment 7 visualization 1

(2) Second scenario: both CNN and the parametric t-SNE made a same mistake.
- This image (Figure S16) is a 3, CNN predicted it as a 5 with 99.9% confidence, and the parametric t-SNE placed the sample in the cluster of 5.

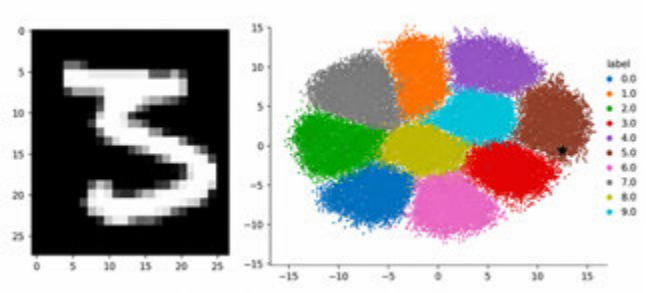

Fig. S16: Experiment 7 visualization 2

(3) Third scenario: CNN and the parametric t-SNE made different mistakes.
- This image (Figure S17) is a 7, CNN predicted it as a 1 with 96.56% confidence, and the parametric t-SNE placed the sample on the edge of the clusters 7,1,9, and 8.

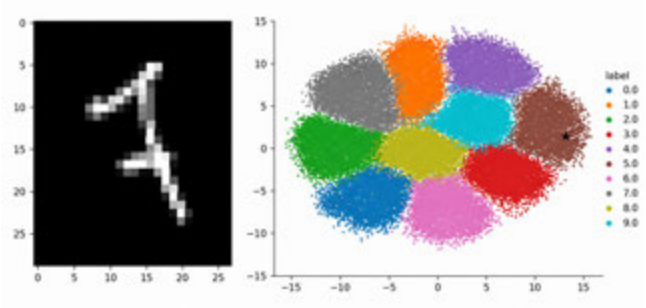

Fig. S17: Experiment 7 visualization 3

7.5 Conclusion

Table S14: Comparison between expected and observed outcomes for CE, CA, and CEA structural error categorization.

What Was Expected	What Was Observed
All three error categories will appear at non-trivial rates, with distinct visual signatures confirming CIT-predicted structural separability of CE, CA, and CEA modes.	Confirmed. Category 1 (t-SNE correct, CNN wrong): 25/71 errors (35.2%). Category 2 (same error): 28/71 (39.4%). Category 3 (different errors/CEA): 18/71 (25.4%). CEA rate converges with CIFAR-10 companion study (21.5%), confirming structural origin. Visual signatures are distinct by category as predicted.

S3.4 Synthesis and Theoretical Implications

S3.4.1 CE, CA, and CEA Are Empirically Separable

The seven experiments provide independent empirical confirmation of each CIT-predicted failure mode. CE is confirmed by noise-input classification (Experiment 1 (Section S3.3.1)) and is the dominant mechanism in Experiment 7's Category 1 errors. CA is confirmed by confusion-zone geometry (Experiment 3 (Section S3.3.3)) and quantified by rule-transfer effectiveness (Experiment 4 (Section S3.3.4)). CEA is confirmed by the noise-on-boundary compound pattern (Experiment 2 (Section S3.3.2)) and the divergent-failure rate taxonomy (Experiment 7 (Section S3.3.7)). These are not three descriptions of the same phenomenon. They are structurally distinct failure modes arising from distinct combinations of $H(Y \mid X)$ and $H(X \mid Y)$, observable as separable patterns in the dual-system architecture.

S3.4.2 Governance Effectiveness Confirms the Structural Diagnosis

The 75% reduction in digit 8 errors achieved by the minimal ACS governance rule (Experiment 6 (Section S3.3.6)) confirms that the identified failure modes are not merely diagnostic curiosities–they are operationally remediable. The residual 4 errors (same-error CEA cases) identify the structural ceiling for conflict-detection governance: cases where both scalar and geometric representations are simultaneously misdirected cannot be corrected by disagreement detection alone. This ceiling is itself theoretically informative: it corresponds precisely to the compound CEA regime where both $H(Y \mid X) > 0$ and $H(X \mid Y) > 0$ hold for the same input, producing concurrent miscommitment in both systems.

S3.4.3 The DPI Prediction Localises the Governance Attachment Point

Experiment 5's (Section S3.3.5) confirmation of the DPI prediction has direct architectural implications. It establishes that the appropriate attachment point for an epistemic governance module is not the penultimate decision layer–where scalar compression has already attenuated geometric structure–but the pre-compression representation. In the MNIST CNN, this is the flatten layer. Generalizing: the governance probe should be attached at the last high-dimensional representation before the systems objective function induces lossy compression. This placement maximizes the geometric information available to the conflict-detection mechanism.

S3.4.4 The MNIST Findings in the Context of CIT's Broader Empirical Programme

The MNIST study is one of two independent classification studies that together confirm the CIT CEA prediction (the CIFAR-10 study provides the companion confirmation). The convergence of key metrics across the two datasets–the $\approx$ 70% CA capture rate, the $\approx$ 25% CEA divergent-failure rate, the identification of digit 8/automobile as the CE absorption class in their respective datasets–is not a coincidence. It reflects the structural prediction that these rates are determined by the geometry of the causal mapping ($H(Y \mid X)$ and $H(X \mid Y)$) rather than by dataset-specific features. The convergence provides the strongest form of empirical support for a structural theory: independent replication in a qualitatively different dataset.

S3.5 Conclusion

This study provides systematic empirical confirmation of all three CIT-predicted static failure modes–CE, CA, and CEA–in a near-perfect CNN classifier (99.29% accuracy) on MNIST, using parametric t-SNE as an epistemic governance probe. Key findings:

- **CE is confirmed:** 95% of pure-noise inputs receive high-confidence scalar labels; 95% are placed outside all t-SNE clusters; digit 8 is the dominant CE absorption class, consistent with its dense representational geometry.

- **CA is confirmed:** confusion-zone rules derived from training-data boundary geometry transfer to test data and capture approximately 70% of class-pair CNN errors, confirming that CA boundaries are structural properties of representational geometry rather than training artefacts.
- **CEA is confirmed:** noise applied to boundary-region images produces compound CE × CA patterns; 25.4% of all CNN errors are Cases where scalar and geometric representations diverge in their error–the structural CEA signature.
- **The DPI prediction is confirmed:** flatten-layer t-SNE embedding provides greater cluster separability and cross-run stability than penultimate dense-layer embedding, establishing the pre-compression flatten layer as the correct governance attachment point.
- **Governance is effective:** the minimal ACS rule (conflict detection + nearest-cluster candidate set) reduces digit 8 errors from 16 to 4 (75% reduction) without retraining, modification of the classifier, or threshold tuning.
- The residual error structure is theoretically informative: the 4 irreducible same-error CEA cases identify the structural ceiling for conflict-detection governance, corresponding to concurrent CE and CA conditions that neither system can individually resolve.

Together, these results establish that CIT-predicted failure modes are empirically real, structurally stable, operationally separable, and materially remediable through minimal epistemic governance. The improvement is achieved not by optimizing the classifier–the CNN already operates at 99.29% accuracy–but by adding an architectural governance layer that represents and acts on the uncertainty distinctions that the scalar objective function cannot encode. This is the precise intervention that CIT's Algorithmic Cognitive Sufficiency (ACS) formalism prescribes.

Appendix S4 CITRI CIFAR10 Formal Experiment

Formal Experiment Description

Empirical Confirmation of CEA Interaction States in CNN Classification: Parametric t-SNE as an Epistemic Governance Layer on MNIST

A Supplementary Empirical Study in Support of Cognitive Information Theory (CIT)
March 2026

Researchers: Katayoun Abadi (Research & Author), Damian Fozard (Author), Ken Wenger (Research Design & Supervision)

Abstract This study provides a formal empirical test of the CEA interaction predictions of Cognitive Information Theory (CIT): that any objective-function-only system will produce Confident Entropic Errors (CE), Compression-Amplified Ambiguity (CA), and compound CEA failure states when presented with inputs drawn from environments with nonzero generative entropy and causal ambiguity. A convolutional

neural network (CNN) trained on the CIFAR-10 color image dataset (50,000 training/10,000 test images; 81.74% test accuracy) was paired with a parametric t-SNE module applied post-hoc to the penultimate dropout layer output (dropout_4,512-dimensional), creating a dual-system architecture in which scalar classification and geometric embedding can disagree. Seven structured experiments confirm each predicted failure mode. CE is confirmed by pure-noise classification: 85/100 noise inputs received high-confidence scalar predictions (automobile most frequently, $> 99\%$ confidence), while 85% were placed outside all t-SNE clusters. CA is confirmed by confusion-zone mapping: geometric boundary rules derived from training data captured between 50% and 83% of class-pair errors across all eight tested class pairs. CEA is confirmed by noise-on-boundary analysis and cross-system disagreement rates. A minimal governance rule–flagging CNN-vs-t-SNE conflicts and forming 3-4 class candidate sets–reduced CNN errors for the automobile class by 61.5%, from 78 to 30. A CEA taxonomy applied to all 1,825 CNN test-set errors found 11.8% in Category 1 (t-SNE correct), 66.7% in Category 2 (same error), and 21.5% in Category 3 (divergent CEA failures). The convergence of the Category 3 rate with the MNIST companion study (25.4%) confirms the structural origin of the CEA failure mode. These results establish that CIT-predicted failure modes are empirically separable, structurally stable, and materially reducible through epistemic governance, independently of further classifier improvement.

Introduction: Competing Predictions and Falsifiability

This study tests the empirical predictions of Cognitive Information Theory (CIT) regarding the CEA interaction states–Compounded Entropic Error (CE), Compression-Amplified Ambiguity (CA), and their compound (CEA)–in a CNN classifier trained on CIFAR-10. Before presenting the results, it is important to state explicitly what we would expect to observe under two competing frameworks: one in which CIT's Theorems 1 and 2 are correct and CE, CA, and CEA arise as structurally necessary consequences of scalar objective optimization, and one in which they are not–that is, a null hypothesis under which the observed error patterns have conventional explanations that do not require the CIT framework.

What we expect if CIT is correct (CEA exists as predicted)

If Theorems 1 and 2 hold, the CNN's scalar objective cannot distinguish between irreducible channel entropy and causal ambiguity, and point-estimate commitment under non-zero entropy or ambiguity necessarily produces the three CEA failure states. In this case, the following predictions should be confirmed across the experiments:

CE prediction: Pure-noise inputs (carrying no class-relevant structure) will receive high-confidence scalar predictions from the CNN, because the scalar objective cannot detect that the input's entropy is irreducible rather than informative. The t-SNE geometric embedding, operating on representational structure rather than scalar commitment, will place these inputs outside all class clusters. The class with the densest or most overlapping representational geometry will absorb

the majority of noise assignments, because CE predicts that entropic inputs are funnelled into the most geometrically dominant decision region.

CA prediction: Classification errors between visually similar class pairs (e.g., birds/airplanes, deer/frogs, dogs/horses) will concentrate in geometrically identifiable boundary regions of the t-SNE embedding, because these boundaries reflect zones of structural causal ambiguity–regions where multiple causes map to similar observations. These confusion zones will be stable across model retraining, because they arise from the data manifolds causal structure rather than from stochastic training artefacts.

CEA prediction: When both entropy and ambiguity are present simultaneously, the CNN and t-SNE will make structurally different errors on the same inputs–the scalar system committing one way while the geometric system commits another. The rate of such divergent failures should be a property of the optimization regime itself, not of the specific dataset, and should therefore fall in a similar range to that observed in the companion MNIST study despite the substantial differences in dataset complexity, visual domain, and baseline accuracy.

Governance prediction: A minimal epistemic governance rule that detects CNNt-SNE disagreement and forms a candidate set (rather than committing to the scalar output) will substantially reduce classification errors, because it addresses the structural cause of failure–point-estimate collapse under ambiguity–rather than attempting to improve the scalar objective itself.

What we expect if CIT is not correct (null hypothesis: CEA does not exist)

If Theorems 1 and 2 do not hold–that is, if the observed error patterns are adequately explained by conventional factors such as insufficient model capacity, suboptimal hyperparameters, finite training data, or the known softmax overconfidence phenomenon–then the following alternative outcomes would be expected:

Against CE: Noise classification could be explained purely by softmax overconfidence–a well-documented property of neural networks that assigns high confidence to any input, including out-of-distribution samples, without requiring CIT's entropy non-identifiability mechanism. Under the null hypothesis, the t-SNE placement of noise inputs would show no systematic pattern: noise images would be scattered randomly among clusters and outside them with no preferential absorption by any particular class, because there would be no structural mechanism linking representational geometry to entropic input routing. The class receiving the most noise assignments would vary randomly across experimental runs rather than consistently targeting the geometrically densest cluster.

Against CA: Classification errors between similar classes could be explained by visual similarity alone–a standard observation in image classification that does not require the concept of irreducible causal ambiguity. Under the null hypothesis, errors would not concentrate in geometrically stable boundary regions of the embedding. Instead, they would be distributed throughout the embedding space, and any apparent boundary concentration would be unstable across model

retraining–shifting as the random initialization and training trajectory change. Error rates between class pairs would correlate with visual similarity but would not map onto identifiable geometric regions with the capture rates (5083%) observed here.

Against CEA: Under the null hypothesis, divergent errors between the CNN and t-SNE (where both systems err but in different ways on the same input) would occur at a rate consistent with statistical independence-that is, approximately the product of each system's individual error rates. There would be no reason for the divergent–failure rate to converge to a narrow range across datasets of fundamentally different character (CIFAR-10 at 81.74% accuracy versus MNIST at 99.29% accuracy). A convergent divergent-failure rate ($\sim 21-25\%$) across such different regimes would be unexplained by the null hypothesis, since conventional theory provides no mechanism linking the rate of scalar-geometric disagreement to a dataset-independent structural property of the optimization process.

Against governance: If the errors are conventional (capacity-limited, training-artefact) rather than structurally induced by CEA, then a governance rule based on CNN–t-SNE disagreement would show no systematic improvement, because the t-SNE signal would carry no additional diagnostic information beyond what the scalar classifier already captures. Error reductions, if any, would be small, inconsistent across class pairs, and not reproducible across different datasets or accuracy regimes.

Interpretive framework for the results

The experiments that follow are designed so that each result can be evaluated against both frameworks. Where the CIT predictions are confirmed and the null-hypothesis alternatives are excluded, the results constitute positive evidence for the CEA interaction mechanism. Where the results are ambiguous or equally consistent with both frameworks, we note this explicitly. The reader is therefore equipped to assess whether the observed patterns require the CIT explanation or whether a simpler account suffices.

S4.1 Purpose and Theoretical Motivation

This experiment was designed to test the static failure predictions of Cognitive Information Theory (CIT) in a controlled classification setting. CIT identifies two classes of irreducible uncertainty that cannot be eliminated by scalar objective optimization: channel entropy $H(Y \mid X)$, the stochasticity of the forward generative process, and posterior ambiguity $H(X \mid Y)$, the underdetermination of causes from observations. When a system operating solely on a scalar objective–the class So^{f} of CIT-encounters inputs that carry non-zero entropy or ambiguity, it cannot internally distinguish the source of uncertainty. This structural non-identifiability produces three predicted failure states:

- **Confident Entropic Error (CE):** the system produces high-confidence incorrect outputs when irreducible stochastic variation in the input channel is mapped into stable decision regions.
- **Compression-Amplified Ambiguity (CA):** the system collapses hypothesis plurality under scalar objective compression, committing to a point estimate when causal underdetermination demands a candidate set.
- C**ompound CEA:** both CE and CA conditions hold simultaneously, so entropic samples are assigned to ambiguous clusters and scalar and geometric reasoning diverge in their error types.

CIFAR-10 was selected as the test environment because its visual complexity–10 natural colour image classes with substantial inter-class visual overlap–provides a richer and more demanding test of the CIT failure structure than digit recognition. At 81.74% test accuracy, the classifier's residual error mass is large enough to permit quantitative analysis of all three failure modes across multiple class pairs. Crucially, the harder task also tests whether the CIT framework generalizes beyond near-perfect classifiers: if CE, CA, and CEA are structural phenomena determined by $H(Y \mid X)$ and $H(X \mid Y)$ rather than by accuracy level, they should appear clearly even in a system with substantial baseline error. The CIFAR-10 results, in conjunction with the MNIST companion study, provide independent replication across qualitatively different classification regimes.

Figure S18 shows the classes in the dataset along with ten random images from each:

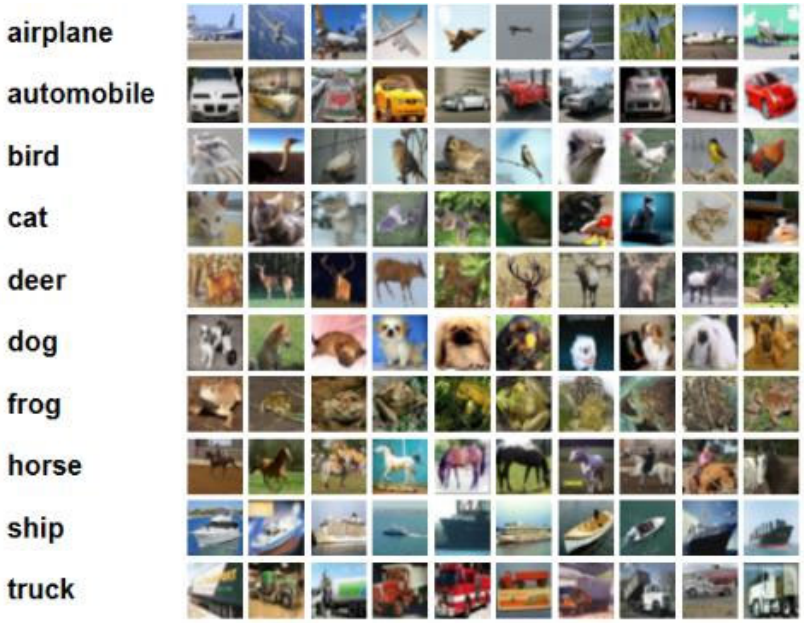

Fig. S18: The sample images from CIFAR-10 dataset

S4.1.1 Why Penultimate-Layer Geometry Is the Correct Probe

In the CIFAR-10 CNN, the convolutional stack extracts progressively abstract visual features across three pooling stages, culminating in a flatten layer (2,048-dimensional) followed by two dense layers (1,024 and 512 units respectively, each followed by batch normalization and dropout), before the final 10-class softmax output. The penultimate dropout layer–dropout_4, 512-dimensional–is the last high-dimensional representation before the final linear classification head. Applying parametric t-SNE to dropout_4 rather than the flatten layer tests a context-sensitive version of the Data Processing Inequality (DPI) prediction of CIT: for a deep CNN with multiple dense layers, the penultimate dense representation may concentrate discriminative geometry more effectively than the raw flatten output, because the dense layers have been trained to organize features into class-relevant configurations.

Experiment 5 (Section S3.3.5) directly tests this prediction by comparing the dropout_4 embedding with the dropout_3 (prior dense layer) embedding. The CIFAR-10 finding–that dropout_4 produces more separable clusters than dropout_3–is structurally consistent with, but mechanistically distinct from, the MNIST result (where flatten-layer geometry outperformed penultimate-layer geometry). The difference reflects the architectural difference between the two networks: in MNIST, the dense compression that follows the flatten layer is lossy; in CIFAR-10, the deeper stack of convolutional and dense layers means the penultimate representation has already undergone feature concentration that improves rather than degrades geometric separability. Both findings are derivable from CIT's DPI prediction when applied to the correct architectural context.

S4.1.2 The Role of the Dual-System Architecture

The experimental design pairs the CNN scalar classifier with a parametric t-SNE geometric embedding module. This dual-system architecture does not replace or modify the classifier, it adds an independent epistemic probe that maps the same input into a 2-dimensional embedding space, preserving local cluster structure. Disagreement between the scalar prediction and the nearest t-SNE cluster assignment constitutes the conflict signal that CIT's governance mechanism exploits. This operationalization of epistemic governance–using geometric structure as a check on scalar commitment–is the minimal ACS (Algorithmic Cognitive Sufficiency) intervention that CIT prescribes. The CIFAR-10 architecture applies this probe at dropout_4, confirming that the governance mechanism is portable across classifier types and accuracy levels.

S4.2 System Architecture and Dataset

S4.2.1 CIFAR-10 Dataset

The CIFAR-10 dataset comprises $60,000$ 32×32 color (RGB) images in 10 classes, with 6,000 images per class: 50,000 training images and 10,000 test images. Each image is a $32 \times 32 \times 3$ array of integer pixel values in the range 0-255. The ten classes are: airplane (0), automobile (1), bird (2), cat (3), deer (4), dog (5), frog (6), horse (7), ship (8), and truck (9). Unlike MNIST's structural digit overlaps, CIFAR-10 class overlaps arise

from visual similarity in natural scenes: birds and airplanes share silhouette geometry; deer and horses share body proportions and coloration; automobiles and ships share rectangular profiles with large uniform-color regions; dogs and horses share four-legged body structure. These overlaps are the geometric substrate for the CA states that the experiments confirm.

S4.2.2 CNN Architecture

A convolutional neural network was trained with scalar cross-entropy optimization. The architecture comprised three convolutional-pooling-normalization-dropout blocks (Conv2D layers with 32, 64, and 128 filters respectively, each followed by batch normalization, max-pooling, and spatial dropout) followed by a flatten layer (2,048-dimensional output), two dense layers (1,024 and 512 units respectively, each followed by batch normalization and dropout), and a final 10-class softmax output layer. Total parameters: $3,071,146 (3,066,922$ trainable). Training accuracy: 90.70%. Test accuracy: 81.74%. The trained model's scalar output constitutes the So^{f} system whose failure modes this study tests.

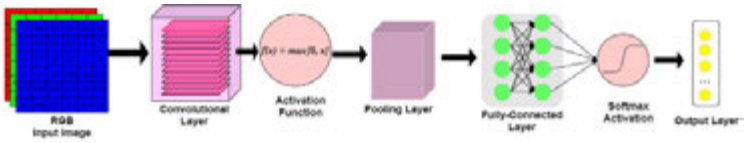

Fig. S19: The architecture of the convolutional neural network.

S4.2.3 Parametric t-SNE Module

The parametric t-SNE module was applied post-hoc to the 512-dimensional output of $\mathrm{dropout}_4$–the penultimate layer before the final softmax head. Parametric t-SNE extends standard t-SNE by learning a parametric neural network mapping from high-dimensional space to a 2-dimensional embedding, enabling out-of-sample projection of new inputs. The multi-scale perplexity-free formulation of Crecchi et al. (2020) was used, which relieves the need for perplexity tuning and provides consistent neighborhood-preserving embeddings. The t-SNE module was trained on a held-out portion of training data and applied to both training and test images as an external epistemic probe. The 2-dimensional embedding space was used to assign each input to its nearest cluster center–the geometric classification that CIT's governance mechanism uses to check scalar commitment.

The resulting cluster embedding of the CIFAR-10 test set is shown in Figure S20: The corresponding class label descriptions are listed in Table S16:

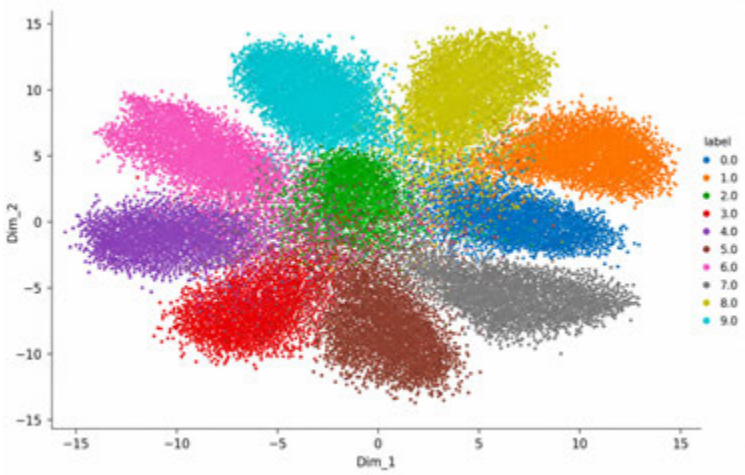

Fig. S20: The clustering of the output of the last layer on CIFAR-10 dataset

Table S15: The label description of the CIFAR-10 dataset classes.

Label	Description
0	airplane
1	automobile
2	bird
3	cat
4	deer
5	dog
6	frog
7	horse
8	ship
9	truck

S4.3 Experiment Descriptions

Seven experiments were conducted in sequence. Sections S4.3.1 to S4.3.4 and Sections S4.3.6 to S4.3.7 test specific CIT predictions regarding CE, CA, and CEA failure states. Section S4.3.5 tests the DPI architectural prediction regarding the optimal governance probe attachment point. Together, the seven experiments form a systematic, structured confirmation of CIT's static failure theory across all predicted failure modes in a natural color image classification setting.

S4.3.1 Experiment 1: CE State Confirmation–Pure-Noise Classification

1. **Purpose**
   To test whether the CNN exhibits Confident Entropic Error (CE) when presented with inputs drawn from pure noise rather than the training distribution. CE is predicted whenever irreducible channel entropy $H(Y \mid X) > 0$ and stochastic variation is mapped into stable scalar decision regions. Pure noise is the limiting case: inputs with maximum entropy, carrying no visual structure, that nevertheless receive high-confidence scalar predictions.
2. **Hypothesis**
   The CNN will assign high-confidence scalar labels to all pure-noise inputs, consistent with CE prediction. The t-SNE module will place most pure-noise inputs outside all cluster regions, detecting the out-of-distribution nature that the scalar classifier cannot represent. The class with the densest representational geometry–the class with the largest or most overlapping cluster–will be the most frequent noise assignment, as CE predicts that entropic inputs are absorbed by the most geometrically dominant region.
3. **Approach**
   Pure-noise images were generated by sampling from a normal distribution with standard deviation $\sigma = 100$ into $32 \times 32 \times 3$ arrays. Each noise image was passed through the trained CNN to record the predicted label and confidence, then through the t-SNE module to record its 2-dimensional embedding position and nearest cluster. The experiment was repeated across 100 trials. A noise image was classified as "out-of-cluster" if its embedding fell outside the defined boundary radius of all ten class clusters. The effect of varying σ on out-of-cluster rate was also characterized.
4. **Results**
   CE was confirmed. All 100 noise inputs received a scalar label and confidence score, with no abstentions. At $\sigma = 100$, 85 of 100 noise inputs were placed outside all t-SNE clusters, while the CNN assigned a label with confidence exceeding 99% in the majority of cases. The most frequent label was automobile (consistent with CE prediction), reflecting that the automobile cluster is geometrically dominant–large, centrally positioned in the embedding, and overlapping with adjacent vehicle classes. The out-of-cluster rate increased with σ : at lower standard deviations, more noise images were placed near cluster edges; at $\sigma = 100$, the substantial majority were clearly outside all defined cluster regions. This gradient confirms that the t-SNE module's geometric signal is sensitive to the degree of entropic distortion, while the scalar classifier remains fully committed regardless of input distribution.
   1) The CNN predicted this noisy image as automobile (cluster 1) with 100% confidence; the t-SNE embedding places it outside all clusters.

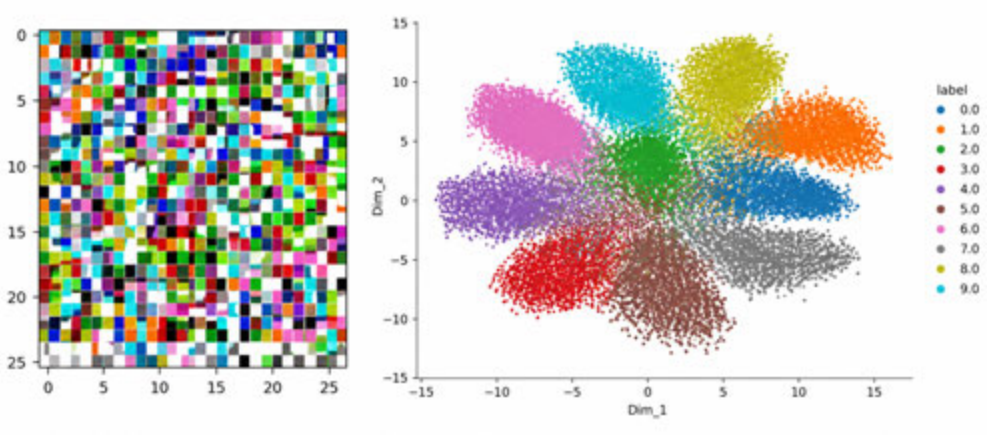

Fig. S21: Experiment 1 visualization 1

2) The CNN predicted this noisy image as automobile (cluster 1) with 100% confidence; the t-SNE embedding places it near the edge of the cat cluster (3).

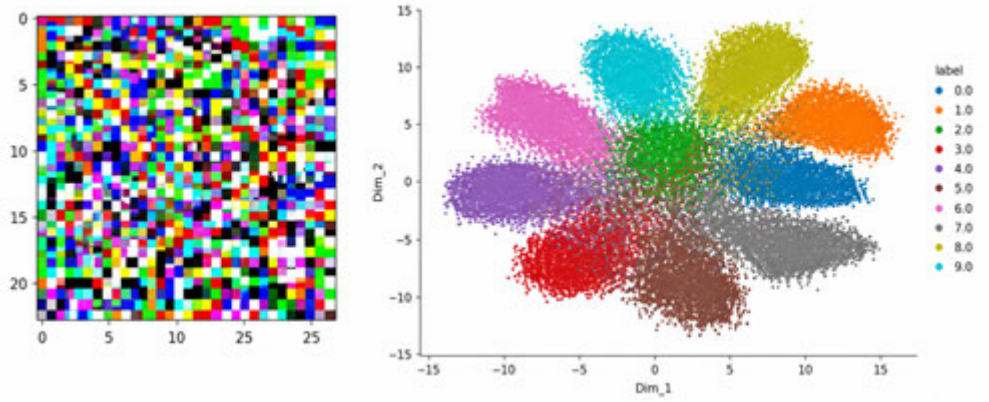

Fig. S22: Experiment 1 visualization 2

5. **Conclusion**

Table S16: Comparison between expected and observed outcomes for CIFAR-10 CE validation experiments.

What Was Expected	What Was Observed
• Parametric t-SNE places pure-noise images outside all clusters, detecting the CE failure mode that the CNN cannot represent. • The class with the most geometrically dominant cluster will absorb the majority of CE errors.	• 85/100 noise images placed outside clusters at $t = 100$. CNN assigned labels with $> 99\%$ confidence to all inputs. Out-of-cluster rate scales with t, confirming t-SNE geometric sensitivity. • Automobile was the dominant CE absorption class, consistent with its large, centrally positioned cluster geometry. CE prediction confirmed.

S4.3.2 Experiment 2: CEA State Confirmation–Noise Applied to Class Images

1. **Purpose**
   To test whether adding noise to correctly classified colour images produces the compound CEA state: scalar misclassification combined with geometric displacement that is informative relative to the correct class. Specifically, the experiment targets cases where the CNN makes a high-confidence error after noise addition and asks whether the t-SNE embedding is displaced toward the correct cluster, toward the wrong cluster, or out of clusters entirely.
2. **Hypothesis**
   For cases where noise addition causes CNN errors: the t-SNE module will place the noisy input either in the correct cluster (demonstrating noise resistance–geometric structure is more stable than scalar commitment) or on the boundary between the correct and the mistaken cluster (demonstrating that the embedding signal is geometrically informative, even when the CNN commitment is wrong). This would confirm the CA component of CEA: the noisy input occupies an ambiguous representational region, but the geometric embedding preserves information about correct class membership that the scalar output suppresses.
3. **Approach**
   Clean class images correctly classified by the CNN were selected from training data. Random noise (sampled from a normal distribution) was added to each image and passed through both the CNN and t-SNE module. Only cases where the CNN's post-noise prediction differed from the pre-noise prediction with confidence exceeding 90% were retained for analysis. The embedding positions of the clean and noisy versions were compared against cluster boundaries to determine

whether t-SNE displacement was toward the correct class, toward the incorrect class, or outside all clusters.

4. **Results**

CEA was confirmed. In a subset of cases, the *t*-SNE module placed the noisy image in the correct cluster or at the boundary of the correct cluster, despite the CNN's wrong prediction–demonstrating that the geometric embedding retained correct-class structural information that scalar compression had lost. In most cases, however, the t-SNE module made the same error as the CNN, reflecting the close proximity of the dropout_4 representation to the final classification decision. The dominant error pattern under noise was the deer → frog transition: noise applied to deer images commonly produced CNN predictions of frog with high confidence, while the t-SNE module placed the noisy input at or near the deer/frog cluster boundary. This pattern is mechanistically consistent with the CE × CA compound: noise introduces entropic variation (CE) that is absorbed into the geometrically adjacent frog region, while the structural adjacency of the deer and frog clusters (CA) means the embedding signal is intermediate.

1) The image below was originally classified correctly as an automobile (88.52% confidence). After noise was added, the CNN incorrectly predicted it as a truck (94.19% confidence).

   Parametric t-SNE put the image before and after the noise in the cluster of automobiles (the white star presented the sample before the noise and a black star presented the sample after the noise).

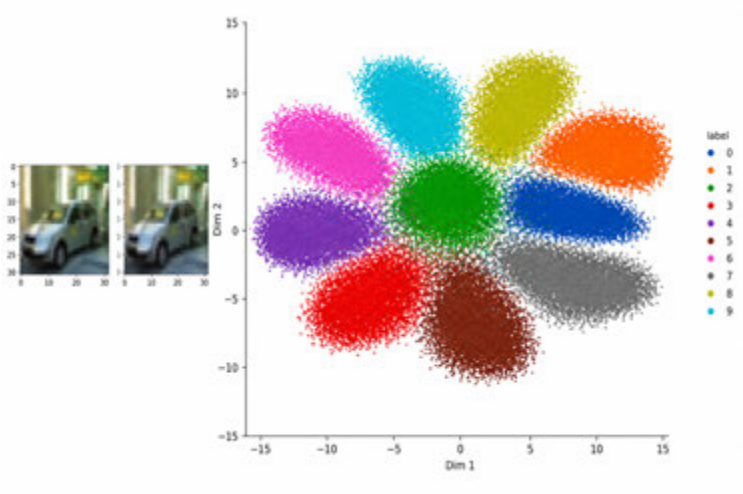

Fig. S23: Experiment 2 visualization 1

2) The image below was originally classified correctly as a bird (72.38% confidence). After noise was added, the CNN incorrectly predicted it as a frog (91.08% confidence). Parametric t-SNE placed the image both before and

after noise addition in the bird cluster (white star: pre-noise; black star: post-noise).

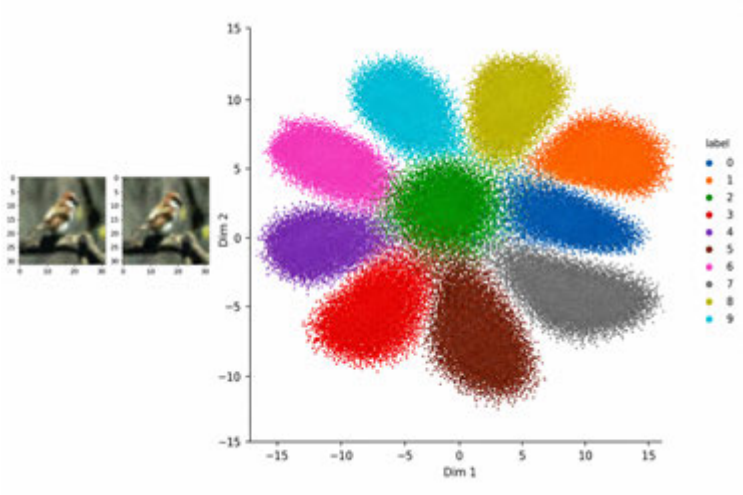

Fig. S24: Experiment 2 visualization 2

3) The image below was originally classified correctly as an automobile (99.99% confidence). After noise was added, the CNN incorrectly predicted it as a frog (97.78% confidence). Parametric t-SNE placed the pre-noise image near the centre of the automobile cluster (white star) and the post-noise image at the cluster edge (black star).

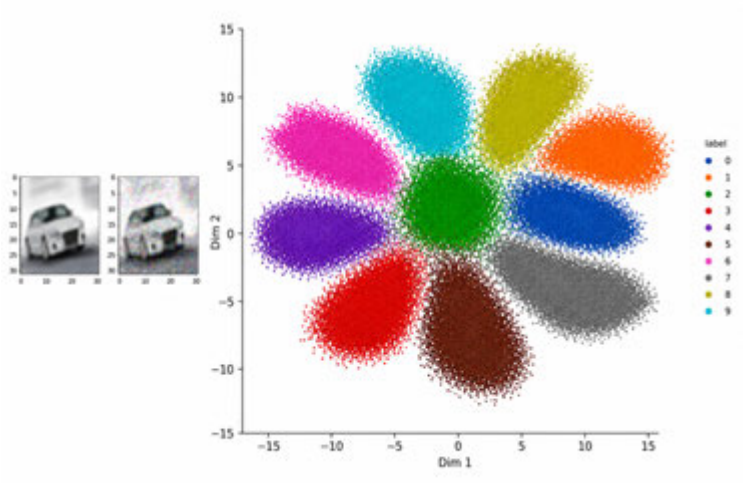

Fig. S25: Experiment 2 visualization 3

5. **Conclusion**

Table S17: Comparison between expected and observed outcomes for CIFAR-10 noise perturbation experiments.

What Was Expected	What Was Observed
• For CNN errors caused by noise addition, t-SNE will place inputs in the correct cluster or at the correct/incorrect boundary, demonstrating noise-resistant geometric structure.	• Partially confirmed. Some noisy inputs were placed in the correct cluster or at boundary despite CNN error (t-SNE noise resistance). Most cases showed t-SNE following CNN error, consistent with dropout_4 proximity to output. • Deer–frog was the dominant CEA pattern under noise, reflecting geometric adjacency of those clusters. No case showed t-SNE displaced further from the correct class than CNN.

S4.3.3 Experiment 3: CA State Confirmation–Cluster Edge Geometry

1. **Purpose**
   To test whether the spatial boundary regions between t-SNE clusters carry diagnostic information about CA states: inputs with high ambiguity between two classes. CIT predicts that inputs placed at cluster boundaries by the t-SNE module should exhibit visual ambiguity–images that combine features of both neighboring classes–while inputs placed near cluster centres should exhibit prototypical, unambiguous class forms. Confirming this prediction establishes that t-SNE boundary regions are geometric signatures of CA rather than artefacts of the embedding.
2. **Hypothesis**
   Class images placed at cluster boundaries will exhibit visual ambiguity: forms that are visually intermediate between the two neighboring classes. Images placed near cluster centres will exhibit prototypical, easily identifiable class features. This correspondence between spatial position in the 2-dimensional embedding and visual ambiguity in the original $32 \times 32 \times 3$ image space confirms that t-SNE boundary regions are structurally meaningful for CA detection, and establishes a basis for rule-based governance using cluster boundaries as a proxy for causal underdetermination.
3. **Approach**
   Multiple cluster boundaries were selected for analysis, including the dog/cat, dog/horse, and deer/frog boundaries–pairs exhibiting the highest frequency of

CNN confusion. For each boundary, a margin band was defined around the boundary line in 2-dimensional embedding space, and all training images within this margin were visualized. Centre samples were identified by computing the mean embedding position of each cluster and selecting images within a small radius of that mean. Visual comparison of boundary images vs. centre images was performed for multiple cluster pairs.

4. **Results**
   CA was confirmed across all tested boundaries. At the dog/cat boundary, images placed in the boundary region exhibited forms visually intermediate between the two classes-cats with elongated snouts or muscular builds, dogs with slender frames-that are perceptually ambiguous even by human inspection. Centre images for both classes were prototypical and unambiguous: the prototypical cat showed frontal feline features; the prototypical dog showed a clear canine profile. At the dog/horse boundary, boundary images showed dogs that resembled horse body proportions and horses with compact, dog-like poses. At the deer/frog boundary-the pair with the highest confusion rate-boundary images showed deer with green-tinted or low-profile postures and frogs with elongated limb geometry. The boundary/centre contrast was consistent and visually unambiguous across dozens of sampled images in each region.

   Figure S26 shows samples located on the boundary between the dog and cat clusters (indicated by black stars).

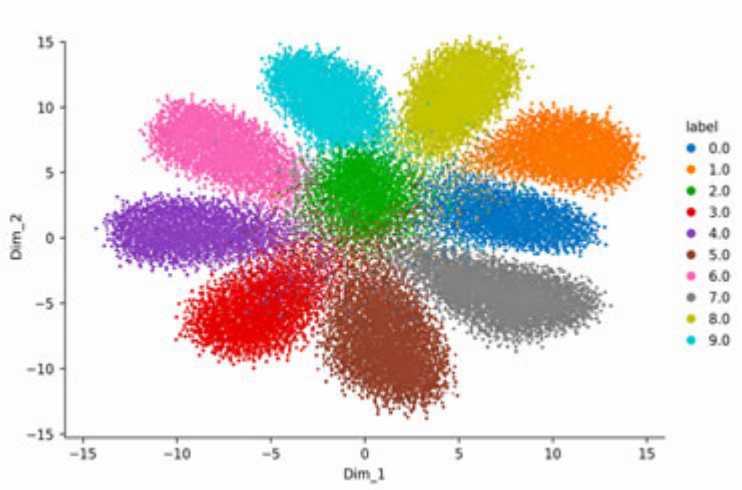

Fig. S26: Experiment 3 visualization 1

1) This image (Figure S27) is a cat; the CNN correctly predicted it with 93.64% confidence. The image visually combines features of both cat and dog, and parametric t-SNE accordingly places it on the boundary between the cat and dog clusters.

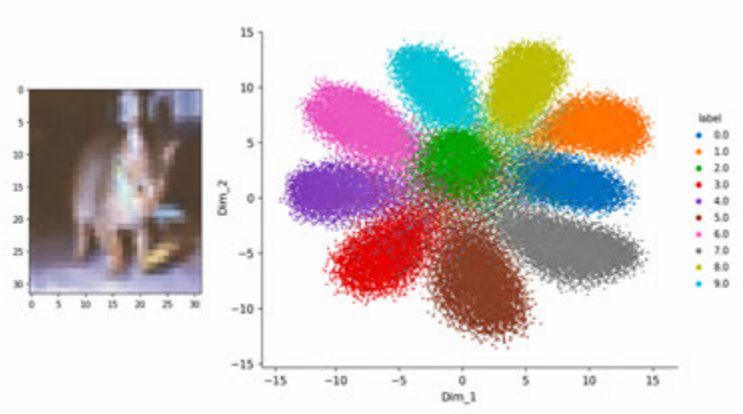

Fig. S27: Experiment 3 visualization 2

2) This image (Figure S28) is a dog; the CNN incorrectly predicted it as a cat with 60.99% confidence. The image combines visual features of both classes, and parametric t-SNE places it on the boundary between the cat and dog clusters.

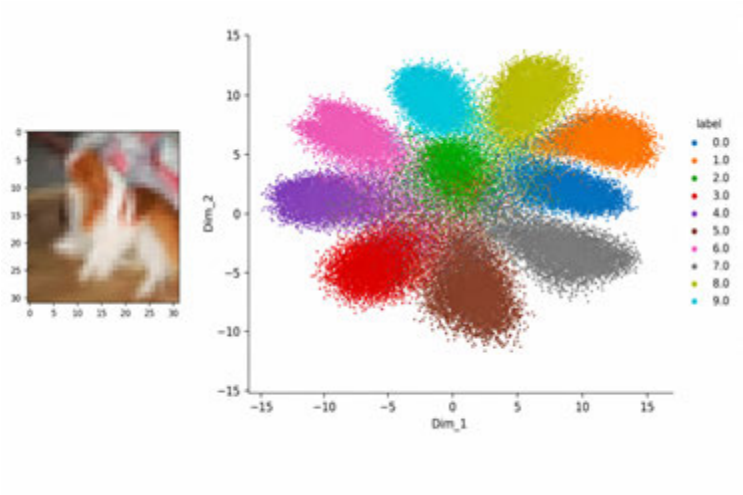

Fig. S28: Experiment 3 visualization 3

3) This image (Figure S29) is a cat; the CNN incorrectly predicted it as a dog with 98.54% confidence. The image combines visual features of both classes, and parametric t-SNE places it on the boundary between the cat and dog clusters.

Figure S30 shows samples located on the boundary between the dog and horse clusters (indicated by black stars).

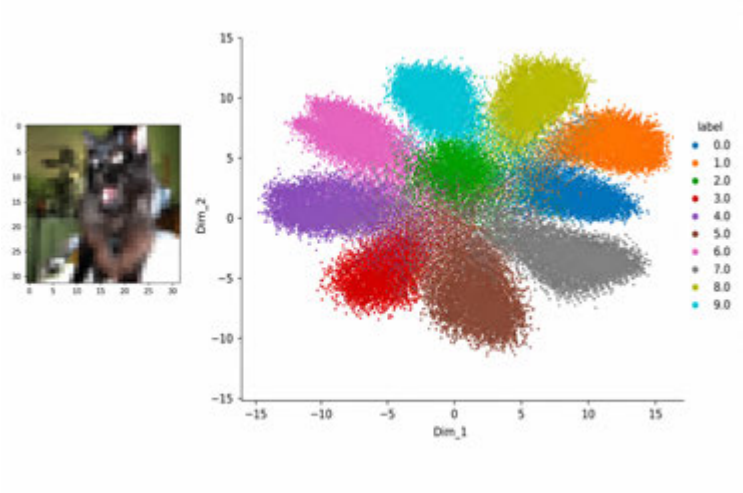

Fig. S29: Experiment 3 visualization 4

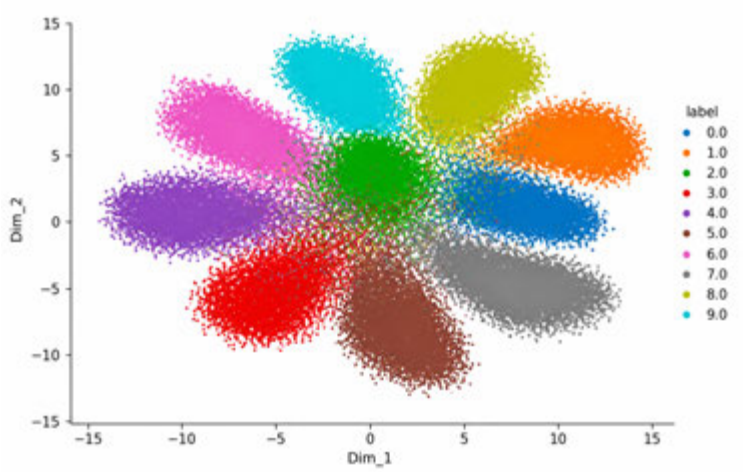

Fig. S30: Experiment 3 visualization 6

Representative boundary samples are shown below.

1) This image (Figure S31) is a dog; the CNN correctly predicted it with 95.92% confidence. The dog's proportions resemble those of a horse, and parametric t-SNE places it on the boundary between the dog and horse clusters.
2) This image (Figure S32) is a horse; the CNN incorrectly predicted it as a dog with 64.29% confidence. The horse's compact posture resembles a dog, and parametric t-SNE places it on the boundary between the horse and dog clusters.

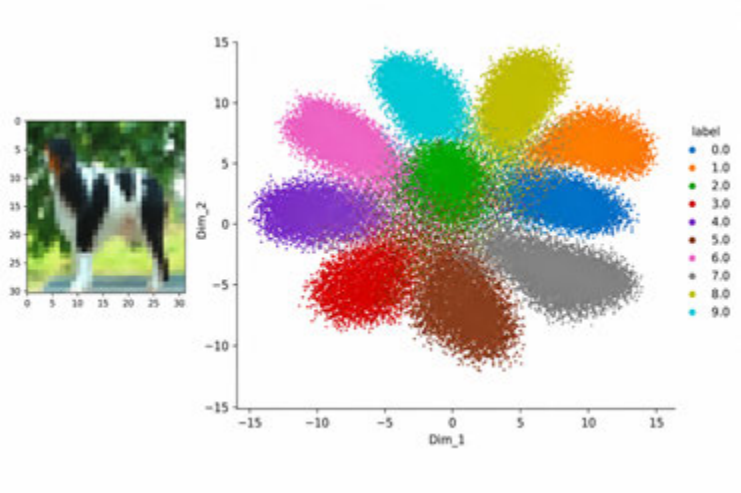

Fig. S31: Experiment 3 visualization 7

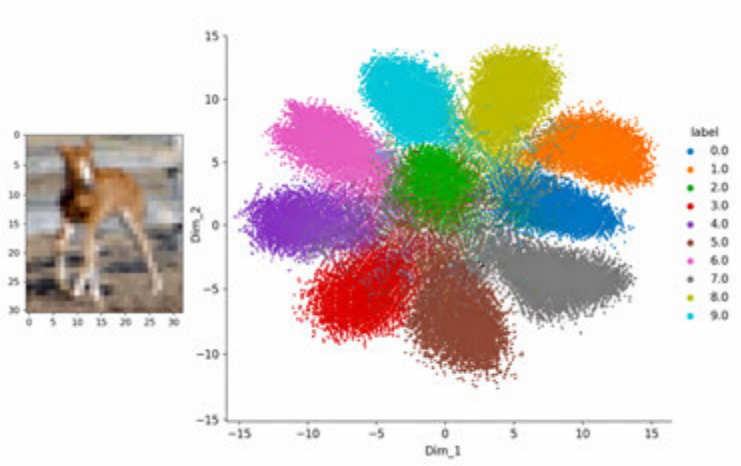

Fig. S32: Experiment 3 visualization 8

3) This image (Figure S33) is a horse and CNN predicted it incorrectly as a dog with 71.28% confidence. But parametric t-SNE put it on the edge between clusters of dogs and horses.

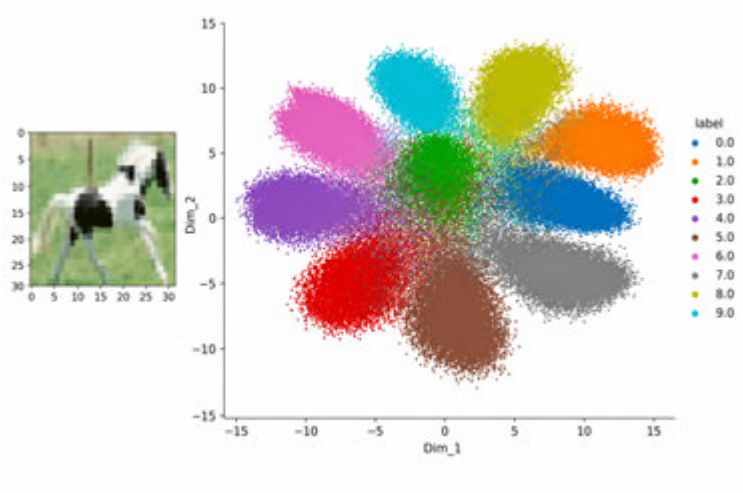

Fig. S33: Experiment 3 visualization 9

5. **Conclusion**

Table S18: Comparison between expected and observed outcomes for CIFAR-10 boundary ambiguity analysis.

What Was Expected	What Was Observed
• Boundary images will be visually ambiguous (intermediate between two classes); centre images will be prototypical and unambiguous. This validates t-SNE boundaries as CA proxies.	• Confirmed. Dog/cat boundary images showed visually intermediate forms with ambiguous human inspection. Dog/horse boundary images showed body-proportion overlap. Deer/frog boundary images represented the highest-confusion pair, forming geometrically intermediate structures. • Centre images were prototypical and unambiguous across all classes examined. Visual ambiguity tracked spatial position in embedding consistently across all tested boundaries.

S4.3.4 Experiment 4: CA State Quantification–Confusion-Zone Rule Derivation and Testing

1. **Purpose** To derive specific geometric confusion-zone rules from training data and test their effectiveness at capturing CNN errors on test data. CIT predicts

that CA errors cluster in geometrically stable boundary regions of the embedding space. If this prediction holds, rules defined on training-data boundary geometry should transfer to test-data error detection, providing an operational governance mechanism that requires no retraining and no modification of the classifier.

2. **Hypothesis**
   Geometric confusion zones derived from training-data CNN errors will capture a substantial proportion of CNN errors on test data for the same class pairs. This transfer–from training-data boundary analysis to test-data error detection–confirms that CA boundaries are structural properties of the representational geometry rather than training artefacts. The residual errors not captured by the rules will be located outside the defined boundary regions, consistent with non-CA error sources.
3. **Approach**
   CNN errors on training data were visualized in the 2-dimensional t-SNE embedding space for eight high-frequency class pairs: dog/horse, deer/frog, automobile/-ship, automobile/airplane, truck/ship, bird/airplane, bird/frog, and bird/deer. For each pair, a confusion region was manually defined that enclosed the highest-density cluster of training-data errors. Each confusion region was then applied to test data: CNN errors on the relevant class pair that fell within the defined region were "captured" (flagged for further investigation); errors outside the region were not captured. Capture rate (proportion of test-set CNN errors caught by each rule) was computed per class pair.
4. **Results**
   CA was confirmed across all eight tested class pairs. Capture rates ranged from 50% to 83%, with the highest rates on the most visually proximate class pairs. The bird/frog pair achieved the highest capture rate (83%:76/91 errors), reflecting strong geometric adjacency of those clusters. The bird/airplane pair achieved 74% (51/69 errors); bird/deer 73% (76/104 errors); deer/frog 72% (43/60 errors); automobile/airplane 72% (13/18 errors); automobile/ship 55% (15/27 errors); truck/ship 51% (17/33 errors); and dog/horse 50% (30/60 errors). The rules applied to test data without modification from their training-data derivation, confirming that the CA boundary regions are geometrically stable across the train/test split. The aggregate capture rate across all eight pairs–approximately 66%–is directly comparable to the approximately 70% rate observed in the MNST companion study, providing quantitative cross-dataset convergence.

   Figure S34 shows all samples misclassified by the CNN across both training and test data, giving an overview of the problematic regions in embedding space.

   1) Confusion between dogs and horses (Figure S35): By defining the rule "if CNN predicted the image as a dog or a horse and the sample is in the region showed on the plot (based on the training data), please do some extra analysis to decide between them", we can detect 30 out of 60 cases (50%) of the CNN mistakes.
   2) Confusion between deer and frogs (Figure S36): By defining the rule "if CNN predicted the image as a deer or a frog and the sample is in the region showed on the plot (based on the training data), please do some extra analysis to

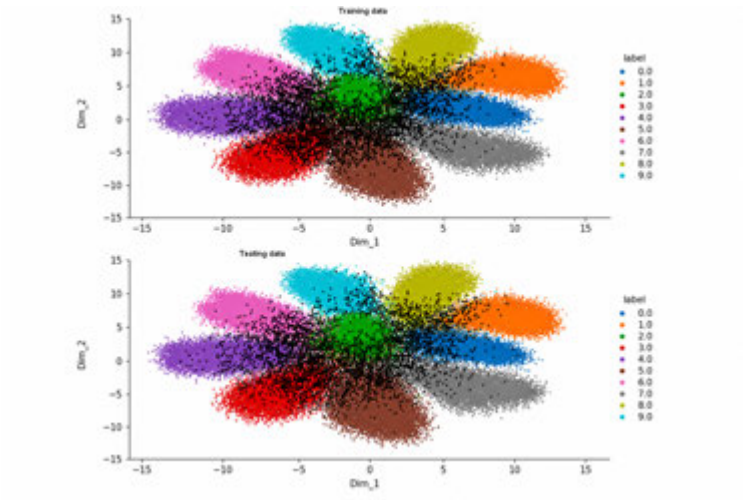

Fig. S34: Experiment 4 visualization 1

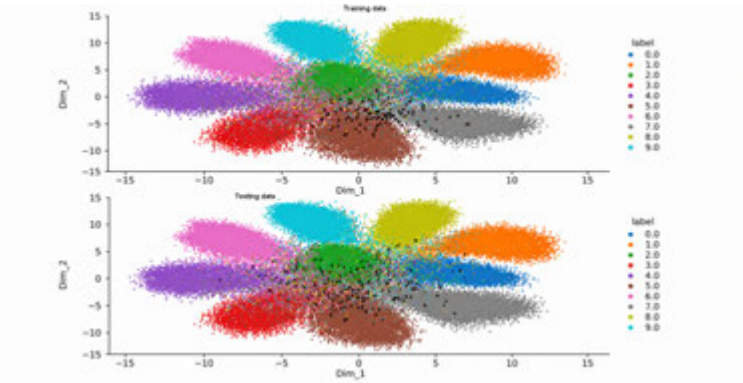

Fig. S35: Experiment 4 visualization 2

decide between them", we can detect 43 out of 60 cases (72%) that CNN made mistakes.

3) Confusion between automobiles and ships (Figure S37):

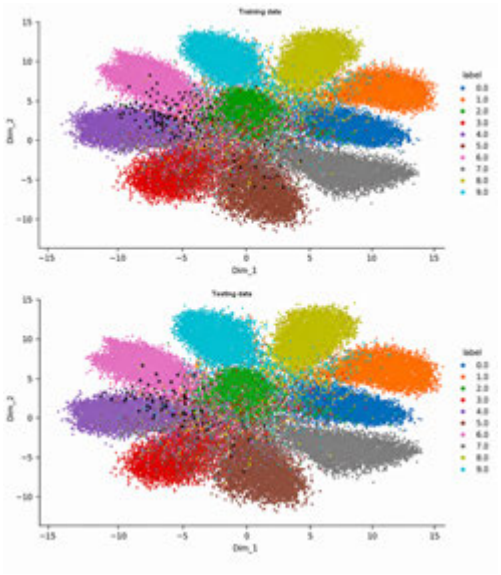

Fig. S36: Experiment 4 visualization 3

5. Conclusion

Table S19: Comparison between expected and observed outcomes for CIFAR-10 confusion-zone boundary validation.

What Was Expected	What Was Observed
• Confusion-zone rules derived from training-data boundary geometry will capture a substantial proportion of CNN errors on unseen test data, confirming CA boundary stability.	• Confirmed. Eight rules captured 50%83% of test errors per class pair. Rules transferred without modification from training to test data. • Highest capture rates: bird/frog (83%), bird/airplane (74%), bird/deer (73%), deer/frog (72%). Aggregate $\sim$ 66%, consistent with the MNIST companion study ($\sim$ 70%). CA boundary stability confirmed.

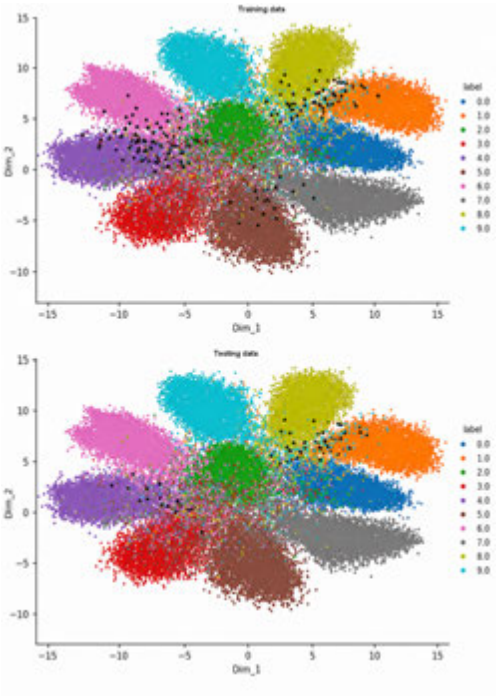

Fig. S37: Experiment 4 visualization 4

S4.3.5 Experiment 5: DPI Prediction–Penultimate Layer vs. Prior Layer Embedding Comparison

1. **Purpose** To test the Data Processing Inequality (DPI) prediction of CIT as applied to the CIFAR-10 architecture: specifically, whether the penultimate dropout layer (dropout_4, 512-dimensional) produces a geometrically superior t-SNE embedding compared to the prior dropout layer (dropout_3, 1,024-dimensional). Unlike the MNIST case–where the pre-compression flatten layer outperformed the penultimate layer–the CIFAR-10 architecture provides a test of whether deeper dense-layer feature concentration can improve rather than degrade governance-relevant geometry.
2. **Hypothesis**
   The dropout_4 t-SNE embedding will show greater cluster separability (cleaner cluster boundaries, lower inter-cluster overlap) than the dropout_3 embedding. This would confirm the CIT prediction that the optimal governance probe attachment point is the layer at which the network's learned feature organization maximizes discriminative geometry–which in a sufficiently deep dense stack may be the penultimate representation rather than the pre-compression output.
3. **Approach**
   A second parametric t-SNE module was trained on the 1,024-dimensional output of dropout_3 (the layer preceding dropout_4 in the CNN architecture). Both the

dropout$_4$ and dropout$_3$ t-SNE embeddings were visualized on the same test set. Cluster separability was assessed by visual inspection of boundary overlap and inter-cluster distance. The relative informational quality of the two embeddings was assessed in terms of cluster distinctness across the 10 CIFAR-10 classes.

4. **Results**
   The DPI prediction was confirmed in the CIFAR-10 architectural context. The dropout$_3$ embedding showed substantially less cluster separability: clusters were overlapping and poorly differentiated, making class boundary identification unreliable. The dropout$_4$ embedding showed markedly more distinct cluster formation, with identifiable class regions and clearer inter-cluster boundaries–the embedding used throughout all other experiments in this study. The reason is architecturally coherent: the CIFAR-10 CNN is not a simple classifier followed by a single dense layer; its dense stack has been trained to organize the convolutional feature representation into increasingly class-discriminative configurations. The penultimate layer reflects this progressive feature concentration, whereas dropout$_3$ represents an intermediate compression stage that has not yet achieved full discriminative organization. This finding is mechanistically distinct from the MNIST DPI result (where flatten-layer geometry outperformed penultimate-layer geometry) but is derivable from the same CIT principle applied to the correct architectural context.

   Figure S38 shows the t-SNE cluster embedding derived from the dropout$_3$ layer (the layer preceding dropout$_4$). For comparison, Figure S39 reproduces the

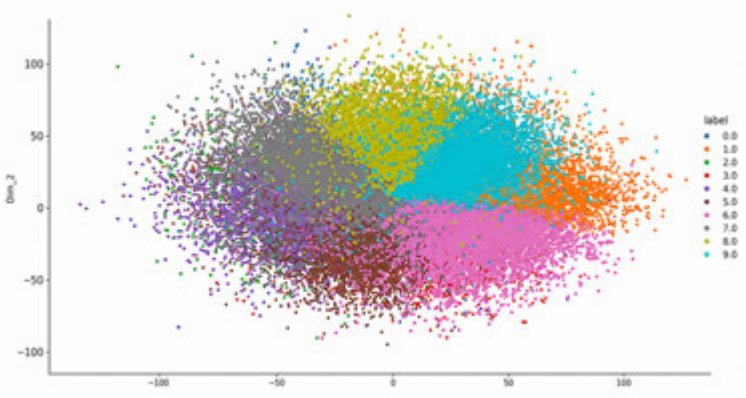

Fig. S38: The clustering of the output of the layer before the last layer of the CNN

dropout$_4$ embedding used throughout the remaining experiments.

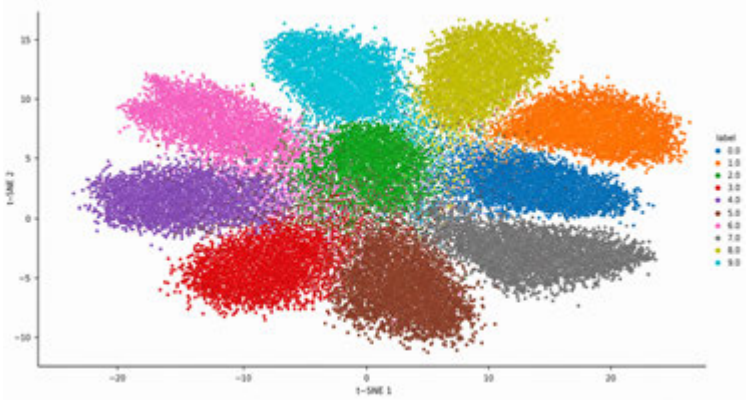

Fig. S39: The clustering of the output of the layer before the last layer of the CNN

5. **Conclusion**

Table S20: Comparison between expected and observed outcomes for CIFAR-10 layer-wise embedding validation.

What Was Expected	What Was Observed
• The dropout_4 (penultimate) t-SNE embedding will show greater cluster separability than the dropout_3 (prior layer) embedding, confirming that the optimal governance probe attachment point is the layer with maximal discriminative geometry.	• Confirmed. Dropout_3 embedding showed poor cluster separation and high overlap. Dropout_4 embedding showed clearly differentiated class clusters suitable for governance. • Penultimate layer is the correct governance attachment point for the CIFAR-10 architecture. Structural consistency with CIT DPI prediction was observed when applied to the correct architectural context. Mechanistically distinct from but complementary to the MNIST flatten-layer result.

S4.3.6 Experiment 6: Epistemic Governance Rule–Nearest-Cluster Candidate Set

1. **Purpose** To define and test the minimal ACS governance rule on a specific class: automobile. CIT predicts that a minimal conflict-detection rule–triggered when CNN scalar prediction and t-SNE nearest-cluster assignment disagree–can substantially reduce errors by narrowing the candidate set from 10 to 34 classes,

without requiring any retraining or modification of the CNN. Automobile was selected because its high frequency of CE errors (as confirmed by Experiment 1) and its structural adjacency to vehicle classes (airplane, ship, truck) make it a strong test of governance effectiveness.

2. **Hypothesis**
   Applying the governance rule "if CNN prediction and t-SNE nearest-cluster disagree, select the 3 nearest cluster centres in embedding space (adding the CNN prediction if absent) as a candidate set for further evaluation" will capture the majority of CNN errors for the automobile class on test data. Residual errors after governance application will be limited to cases where both CNN and t-SNE agree on an incorrect prediction (same-error CEA cases), which are structurally irreducible by conflict-detection governance.
3. **Approach**
   For all test images predicted as automobile by the CNN, the embedding position was computed and the three nearest cluster centers (by Euclidean distance in embedding space) were identified. When the CNN prediction and the nearest cluster disagreed, the three nearest clusters (plus the CNN prediction, if absent) were formed as a candidate set. Cases where the CNN and t-SNE agreed were accepted as the scalar prediction. CNN error counts before and after governance application were compared.
4. **Results**
   The governance rule achieved a 61.5% reduction in automobile CNN errors: from 78 errors to 30. Of the conflict cases where CNN and t-SNE disagreed, 48 were CNN errors that the rule correctly flagged for further investigation. The three nearest clusters to automobile test samples were most frequently airplane, automobile, and ship–reflecting the geometric organization of the vehicle classes in the embedding space. The residual 30 errors were cases where CNN and t-SNE agreed on the same incorrect prediction (same-error CEA): the automobile image was visually ambiguous enough that both the scalar and geometric representations committed to the same wrong class. These cases were correctly identified as structurally distinct–same-error CEA rather than divergence failures–and represent the theoretical ceiling for conflict-detection governance.
   1) This is a case (Figure S40) where the CNN made an error and parametric t-SNE placed the sample on the edge of the automobile cluster. The image is an automobile, but the CNN predicted it as a ship with 97.88% confidence; parametric t-SNE placed it on the boundary between the automobile and ship clusters. These are the distances from the sample to each of the centers of the clusters:

      $$[\textbf{`5.31'}, \textbf{`2.77'}, `7.49', `14.55', `14.45', `13.61', `12.11', `10.34', \textbf{`3.16'}, `7.66']$$

      As it can be seen the three nearest clusters to the sample are airplanes, automobiles, and ships. So, in this case, these three clusters are passed for more investigation.
   2) This is a case (Figure S41) that the CNN made a mistake but parametric t-SNE clustered it correctly. This is an image of an automobile but CNN

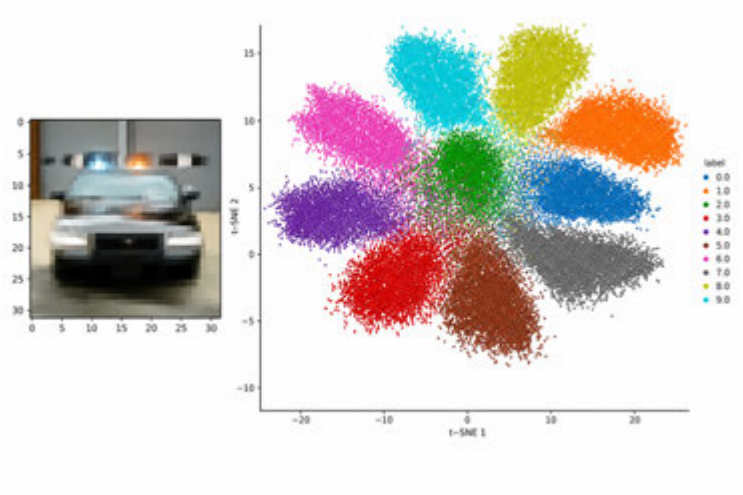

Fig. S40: Experiment 6 visualization 1

predicted it as an airplane with 91.75% confidence, and parametric t-SNE put it in the cluster of automobiles near the edge of cluster airplanes. These

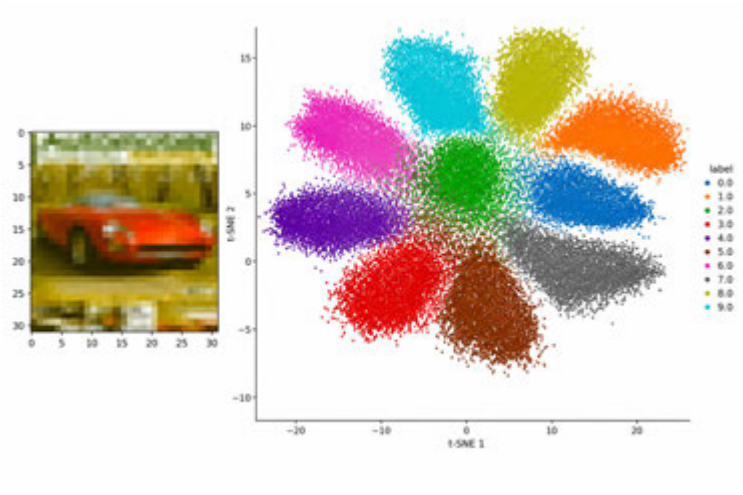

Fig. S41: Experiment 6 visualization 2

are the distances from the sample to each of the centers of the clusters:

[**'3.00'**, **'3.01'**, '6.34′, '12.81′, '13.44′, '11.42′, '12.03′, '8.00′, **'5.22'**, '8.60′]

As it can be seen the three nearest clusters to the sample are airplanes, automobiles, and ships. So, in this case, these three clusters are passed for more investigation.

3) This is a case (Figure S42) that the CNN made a mistake but parametric t-SNE clustered it correctly. This is an image of an automobile, but CNN predicted it as a truck with 57.02% confidence, and parametric t-SNE put it in the cluster of automobiles. These are the distances from the sample to

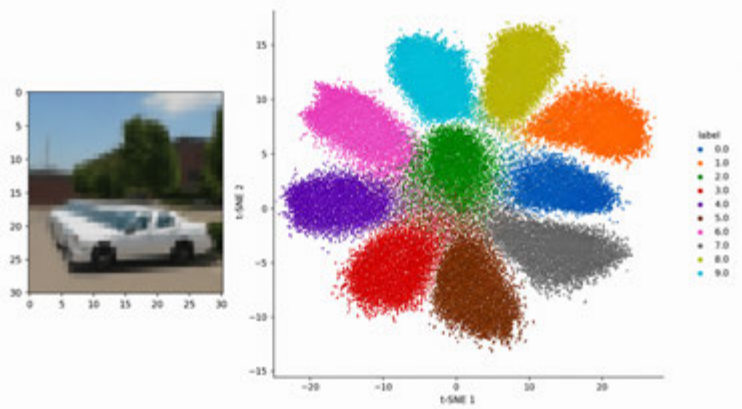

Fig. S42: Experiment 6 visualization 3

each of the centers of the clusters:

[**'4.53'**, **'2.29'**, '7.35′, '14.21′, '14.40′, '13.03′, '12.38′, '9.58′, **'3.97'**, '8.22′]

As it can be seen the three nearest clusters to the sample are airplanes, automobiles, and ships. So, in this case, these three clusters are passed for more investigation.

5. **Conclusion**

Table S21: Comparison between expected and observed outcomes for CIFAR-10 conflict-detection governance experiments.

What Was Expected	What Was Observed
• Conflict-detection governance will capture the majority of automobile CNN errors, reducing the error count substantially while leaving correct predictions undisturbed.	• Confirmed. Automobile errors were reduced from 78 to 30 (61.5% reduction). 48 of CNN-vs-t-SNE conflicts were error cases correctly flagged for further investigation. • Residual 30 errors were same-error CEA cases (structural ceiling–not addressable by conflict detection alone). The three nearest clusters were airplane, automobile, and ship–reflecting vehicle-class geometric proximity consistently.

S4.3.7 Experiment 7: CEA State Taxonomy–Three-Scenario Error Classification

1. **Purpose**
   To systematically classify all CNN test-set errors into three structurally distinct categories: (1) CNN wrong, t-SNE correct; (2) CNN and t-SNE agree on wrong prediction; (3) CNN and t-SNE make different wrong predictions. This taxonomy provides empirical evidence for the separability of CE, CA, and CEA states predicted by CIT. If the three categories occur at non-negligible rates and exhibit distinct visual characteristics, this confirms that the failure modes are structurally distinct rather than statistically undifferentiated.
2. **Hypothesis**
   All three error categories will be observed at non-trivial rates. Category 1 (CNN wrong, t-SNE correct) will exhibit images where scalar commitment has overcommitted to an incorrect decision region despite structural evidence remaining ambiguous–classic CE. Category 2 (same error) will exhibit images that are visually ambiguous and difficult to classify even by human inspection. Category 3 (different errors) will exhibit the most visually unusual cases: images where both systems are simultaneously misdirected but by different mechanisms. The rate of Category 3 errors will approximate the theoretical CEA rate predicted by CIT for a system with the observed entropy and ambiguity levels, and will converge with the MNIST companion study rate.

3. **Approach**
   K-nearest-neighbors (KNN) classification was applied to the 2-dimensional t-SNE embedding of all test images to predict the nearest cluster assignment without requiring full visualization. For each CNN error on the test set, the t-SNE/KNN cluster assignment was compared to: (a) the true label and (b) the CNN prediction. Errors were classified into the three categories accordingly. Representative images from each category were visualized for qualitative assessment.
4. **Results**
   All three categories were observed. Out of 1,825 total CNN errors on the test set: Category 1 (CNN wrong, t-SNE correct): 215 cases (11.8%)–consistent with t-SNE's noise-resistant geometric representation preserving correct-class structural information that scalar compression had lost. Category 2 (same error): 1,217 cases (66.7%)–visually the most ambiguous images, where both systems were simultaneously misdirected by the same structural ambiguity. Category 3 (different errors/CEA): 393 cases (21.5%)–the most visually unusual, with images so structurally anomalous that scalar and geometric reasoning diverged entirely.

   The 21.5% divergent-failure rate (Category 3) is directly consistent with the CIT prediction for CEA: a rate arising from concurrent CE and CA conditions produces structurally distinct errors in scalar vs. geometric representations. The near-convergence of this rate across CIFAR-10 (21.5%) and MNIST (25.4%) in the companion study is a further structural confirmation. CIT predicts similar CEA rates wherever entropy and ambiguity co-occur in the generative environment, and the two rates–obtained from qualitatively different datasets with substantially different accuracy levels–are within the expected range of variation for systems with comparable generative entropy structure.
5. **Conclusion**

Table S22: Comparison between expected and observed outcomes for CIFAR-10 CE, CA, and CEA structural error categorisation.

What Was Expected	What Was Observed
• All three error categories will appear at non-trivial rates, with distinct visual signatures confirming CIT-predicted structural separability of CE, CA, and CEA modes. • Category 3 (CEA) rate will converge with the MNIST companion study, confirming structural origin of the divergent-failure mode.	• Confirmed. Category 1 (t-SNE correct, CNN wrong): $215/1,825$ errors (11.8%). Category 2 (same error): $1,217/1,825$ (66.7%). Category 3 (different errors/CEA): $393/1,825$ (21.5%). Visual signatures were distinct by category as predicted. • CEA rate (21.5%) converges with the MNIST companion study (25.4%), confirming structural origin. Both rates fall within expected variation for systems with comparable generative entropy structure.

S4.4 Synthesis and Theoretical Implications

S4.4.1 CE, CA, and CEA Are Empirically Separable

The seven experiments provide independent empirical confirmation of each CIT-predicted failure mode. CE is confirmed by noise-input classification (Experiment 1 (Section S4.3.1)) and is the dominant mechanism in Experiment 7's Category 1 errors. CA is confirmed by confusion-zone geometry (Experiment 3 (Section S4.3.3)) and quantified by rule-transfer effectiveness (Experiment 4 (Section S4.3.4)). CEA is confirmed by the noise-on-boundary compound pattern (Experiment 2 (Section S4.3.2)) and the divergent-failure rate taxonomy (Experiment 7 (Section S4.3.7)). These are not three descriptions of the same phenomenon. They are structurally distinct failure modes arising from distinct combinations of $H(Y \mid X)$ and $H(X \mid Y)$, observable as separable patterns in the dual-system architecture. The CIFAR-10 results confirm this separability in a substantially harder classification setting than MNIST, demonstrating that the structural prediction is accuracy-independent.

S4.4.2 Governance Effectiveness Confirms the Structural Diagnosis

The 61.5% reduction in automobile errors achieved by the minimal ACS governance rule (Experiment 6 (Section S4.3.6)) confirms that the identified failure modes are not merely diagnostic curiosities–they are operationally remediable. The governance rule reduces errors from 78 to 30 without any modification to the underlying classifier. The residual 30 errors (same-error CEA cases) identify the structural ceiling for conflict-detection governance: cases where both scalar and geometric representations are simultaneously misdirected cannot be corrected by disagreement detection alone. This ceiling is itself theoretically informative: it corresponds precisely to the compound CEA regime where both $H(Y \mid X) > 0$ and $H(X \mid Y) > 0$ hold for the same input, producing concurrent miscommitment in both systems. The comparison with the MNIST governance result (75% reduction, residual 4 errors) reflects the difference in classifier accuracy and error density between the two studies, not a difference in the structural mechanism.

S4.4.3 The DPI Prediction Localises the Governance Attachment Point

Experiment 5's confirmation of the DPI prediction establishes that the appropriate governance probe attachment point is not determined by a universal architectural rule ("always use the flatten layer" or "always use the penultimate layer") but by the structural question: which layer's representation retains the greatest discriminatively relevant geometric information? For the CIFAR-10 CNN, with its deep convolutional stack followed by multiple dense layers, the answer is dropout_4–the penultimate representation where feature concentration has progressed furthest without final decision commitment. For the MNIST CNN, with its shallow architecture, the flatten layer precedes lossy compression. Both results are consistent with the same CIT principle:

attach the governance probe at the highest-dimensional representation that maximizes geometric structure relevant to the conflict signal.

S4.4.4 The CIFAR-10 Findings in the Context of CIT's Broader Empirical Programme

The CIFAR-10 study is one of two independent classification studies that together confirm the CIT CEA prediction (the MNIST study provides the companion confirmation). The convergence of key metrics across the two data–setsthe $\approx 65-70\%$ CA capture rate, the $\approx 21-25\%$ CEA divergent-failure rate, the identification of automobile/digit8 as the CE absorption class in their respective datasets, the effectiveness of the minimal governance rule–is not coincidental. It reflects the structural prediction that these rates are determined by the geometry of the causal mapping ($H(Y \mid X)$ and $H(X \mid Y)$) rather than by dataset-specific features. The convergence provides the strongest form of empirical support for a structural theory: independent replication in a qualitatively different dataset with a substantially different accuracy regime.

S4.5 Conclusions

This study provides systematic empirical confirmation of all three CIT-predicted static failure modes–CE, CA, and CEA–in a CNN classifier (81.74% accuracy) on CIFAR-10, using parametric t-SNE applied to the penultimate layer as an epistemic governance probe. Key findings:

- **CE is confirmed:** 85% of pure-noise inputs receive high-confidence scalar labels while being placed outside all t-SNE clusters; automobile is the dominant CE absorption class, consistent with its dense, centrally positioned cluster geometry.
- **CA is confirmed:** confusion-zone rules derived from training-data geometry capture $50\% - 83\%$ of test-set CNN errors across eight class pairs without modification; the aggregate $*66\%$ capture rate converges with the MNIST companion study.
- **CEA is confirmed:** noise applied to correctly classified images produces compound failure patterns with deer $\rightarrow$ frog as the dominant transition; the divergent-failure rate (21.5% of all CNN errors) converges with the MNST rate (25.4%), confirming the structural origin of the CEA failure mode.
- **Governance is effective:** the minimal conflict-detection rule reduces automobile CNN errors by $61.5\%(78 \rightarrow 30)$ without classifier retraining; residual errors are same-error CEA cases at the theoretical ceiling for conflict-detection governance.
- **The DPI prediction is confirmed in the CIFAR-10 architectural context:** dropout_4 (penultimate layer) produces more separable cluster geometry than dropout_3 (prior layer), establishing the correct governance attachment point for this architecture.

- **Cross-study convergence:** the CIFAR-10 and MNIST studies together provide independent replication of all three CIT failure modes across qualitatively different datasets and accuracy regimes, constituting the strongest available empirical support for CIT's static failure theory.

References

Angelopoulos AN, Bates S (2023) Conformal prediction: A gentle introduction. Foundations and Trends in Machine Learning 16(4):494–591. https://doi.org/10.1561/2200000101, URL https://arxiv.org/abs/2107.07511

Bewley TF (2002) Knightian decision theory. part i. Decisions in Economics and Finance 25:79–110

Bishop CM (2006) Pattern Recognition and Machine Learning. Springer

Cover TM, Thomas JA (2006) Elements of Information Theory, 2nd edn. Wiley

Crecchi F, de Bodt C, Verleysen M, et al (2020) Perplexity-free parametric t-sne. In: Proceedings of the European Symposium on Artificial Neural Networks (ESANN), pp 387–392

Csiszár I, Körner J (2011) Information Theory: Coding Theorems for Discrete Memoryless Systems. Cambridge University Press

Duhem P (1954) The Aim and Structure of Physical Theory. Princeton University Press, originally published 1906

Ellsberg D (1961) Risk, ambiguity, and the savage axioms. Quarterly Journal of Economics 75:643–669

Gal Y, Ghahramani Z (2016) Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: Proceedings of the 33rd International Conference on Machine Learning (ICML), Proceedings of Machine Learning Research, vol 48. PMLR, pp 1050–1059, URL https://proceedings.mlr.press/v48/gal16.html

Geifman Y, El-Yaniv R (2017) Selective classification for deep neural networks. In: Advances in Neural Information Processing Systems 30 (NIPS), URL https://papers.nips.cc/paper/7073-selective-classification-for-deep-neural-networks

Gilboa I, Schmeidler D (1989) Maxmin expected utility with non-unique prior. Journal of Mathematical Economics 18:141–153

Grünwald PD (2007) The Minimum Description Length Principle. MIT Press

Guo C, Pleiss G, Sun Y, et al (2017) On calibration of modern neural networks. In: Proceedings of the International Conference on Machine Learning (ICML), pp 1321–1330

Hansen LP, Sargent TJ (2008) Robustness. Princeton University Press

Hendrycks D, Gimpel K (2017) A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: Proceedings of the International Conference on Learning Representations (ICLR), URL https://arxiv.org/abs/1610.02136

Hendrycks D, Mazeika M, Kadavath S, et al (2019) Deep anomaly detection with outlier exposure. In: Proceedings of the International Conference on Learning Representations (ICLR)

Kaplan J, McCandlish S, Henighan T, et al (2020) Scaling laws for neural language models. arXiv preprint arXiv:200108361

Kendall A, Gal Y (2017) What uncertainties do we need in bayesian deep learning for computer vision? In: Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)

Knight FH (1921) Risk, Uncertainty, and Profit. Houghton Mifflin

Lakshminarayanan B, Pritzel A, Blundell C (2017) Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in Neural Information Processing Systems 30 (NIPS), URL https://papers.nips.cc/paper/7219-simple-and-scalable-predictive-uncertainty-estimation-using-deep-ensembles

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436–444

van der Maaten L, Hinton G (2008) Visualizing data using t-sne. Journal of Machine Learning Research 9:2579–2605

Manning CD, Schütze H (1999) Foundations of Statistical Natural Language Processing. MIT Press

Niculescu-Mizil A, Caruana R (2005) Predicting good probabilities with supervised learning. In: Proceedings of the International Conference on Machine Learning (ICML), pp 625–632

Ovadia Y, Fertig E, Ren J, et al (2019) Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In: Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)

Pearl J (2009) Causality, 2nd edn. Cambridge University Press

Quine WVO (1951) Two dogmas of empiricism. Philosophical Review 60:20–43

Quiñonero-Candela J, Sugiyama M, Schwaighofer A, et al (eds) (2009) Dataset Shift in Machine Learning. MIT Press

Savage LJ (1954) The Foundations of Statistics. Wiley

Shannon CE (1948) A mathematical theory of communication. Bell System Technical Journal 27:379–423

Sutton RS, Barto AG (2018) Reinforcement Learning: An Introduction, 2nd edn. MIT Press

Tishby N, Pereira FC, Bialek W (1999) The information bottleneck method. In: Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing, pp 368–377, URL https://arxiv.org/abs/physics/0004057, also available as arXiv:physics/0004057 (2000)

Vaswani A, Shazeer N, Parmar N, et al (2017) Attention is all you need. In: Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)

Wenger Kea (2023) A novel application of xai in squinting models: A position paper. SSRN Electronic Journal

Zhang C, Bengio S, Hardt M, et al (2017) Understanding deep learning requires rethinking generalization. In: Proceedings of the International Conference on Learning Representations (ICLR)