Large Language Models are a Tool for Communication, Not for Reasoning

Motivation: Human neuroscience suggests that language functions primarily as a communication tool rather than as the substrate of complex thought. Brain imaging studies find that language production/comprehension areas are distinct from the brain's multiple-demand networks for reasoning and problem-solving. We hypothesize that current Large Language Models (LLMs), despite their fluent linguistic abilities, similarly do not perform true reasoning; instead, they rely on latent knowledge implicit in their training data and embeddings.

Approach: We empirically test this hypothesis by *comparing* LLM-based approaches to purely numerical, embedding-based methods across three distinct tasks: information retrieval, numeric prediction, and text classification. For each task, we conduct paired experiments: a pre-trained LLM is prompted to solve the problem in natural language, while a baseline system operates on fixed semantic vector embeddings using standard numerical algorithms. To make these comparisons structurally aligned, we design prompting strategies as direct analogues of classic learning procedures—establishing an *isometry between linguistic prompts and numerical methods*. Concretely, we use descriptive prompts (analogous to dimensionality reduction), contrastive prompts (analogous to contrastive learning), and reflective, iterative prompts (analogous to gradient-descent-style optimization).

Results: Semantic embedding methods outperformed language-based LLM reasoning on all tasks in accuracy, while also being far more efficient in speed and cost. The LLM's language-based "reasoning" often regressed toward chance-level performance or simple heuristics, indicating that it does not reliably sustain the kind of precise, multi-dimensional reasoning delivered by specialized numerical methods. Our findings bridge neuroscience and AI: they suggest that LLMs excel as communicators (generating coherent language) but are limited as standalone reasoning engines in these settings. This has significant implications for AI system design—motivating hybrid architectures that separate communicative generation from the core representational and computational substrate used for reasoning.

Introduction

Large Language Models (LLMs) have achieved remarkable success in generating human-like text. Based on the transformer architecture, they exhibit strong syntactic and semantic competence in tasks such as translation, summarization, question-answering, and classification. Their capabilities have been further boosted by two successive waves of post-training improvements: first, instruction-style fine-tuning and preference-based alignment, and more recently, inference-time strategies that expand a model's reasoning depth and verification steps at test time. In this latter regime, models increasingly produce extended reasoning trajectories enabling competition-level performance on certain benchmarks—culminating in reports of near

gold-medal performance on 2025 International Mathematical Olympiad problems (Google DeepMind, 2025) and comparable “olympiad-level” physics results (Xiong et al., 2025; Niu et al., 2025).

At the same time, the dominant reinforcement-learning path to improving LLM reasoning raises structural questions. First, many purported “reasoning” gains can be interpreted as better probability mass reallocation, rather than the model reliably generating qualitatively new solution strategies . Second, the strongest successes tend to come on tasks with verifiable feedback, whereas preference-based rewards can amplify agreement over truth, leading to measurable sycophancy. Third, heavy optimization pressure can induce reward hacking behaviors in which models exploit spurious cues instead of generalized learning. Fourth, chain-of-thought explanations can be non-faithful, where the verbal “reasoning” may be a plausible rationalization, but not actually used to arrive at the answer. Fifth, token-level RL on language suffers from severe credit assignment problems due to extremely long horizons and enormous action spaces. Finally, even frontier LLMs still exhibit persistent hallucinations, and multiple evaluation studies continue to reveal illusion failure modes where performance degrades under small input perturbations or where benchmarks overstate reasoning ability due to data contamination.

Neuroscience further motivates this question of whether “better language” entails “better reasoning.” In the human brain, the fronto-temporal language network is reliably dissociated from the brain’s multiple-demand system that supports complex cognitive tasks (including mathematical and logical reasoning). This double dissociation is evident both in neuroimaging and in neuropsychological cases where severe language impairments coexist with preserved problem-solving ability. In parallel, recent work on semantic cognition has identified amodal semantic regions that represent meaning in a modality-independent way (responding similarly to linguistic inputs like sentences and to non-linguistic inputs like pictures). Intriguingly, these semantic regions are largely distinct from core language areas. Taken together, this neuroarchitecture supports a crisp interpretation: language is an interface for encoding and transmitting information, whereas central reasoning and semantic evaluation operate in representational formats that do not depend on linguistic form.

Hypothesis: Inspired by these insights, we hypothesize that LLMs—models fundamentally trained to predict and mimic language—do not perform reasoning in the way a dedicated symbolic or neural reasoning module would. Instead, any reasoning-like behavior of LLMs likely stems from latent knowledge encoded in their learned semantic representations. We posit that all the necessary information for reasoning is already present in the continuous high-dimensional vector embeddings that LLMs internalize during training, and that expressing this information in natural language form adds no further reasoning power. We formalize this based on the Linear Representation Hypothesis (LRH): any inference achievable through an LLM’s linguistic output was already achievable by linear operations in the underlying semantic vector space. In this view, linguistic representations are a lossy, compressed projection of a richer numeric world model.

Impact: If this hypothesis holds, then using language as the medium for reasoning is suboptimal—one could reason more accurately and efficiently by staying entirely within the high-dimensional numeric embedding space. In other words, improving one's linguistic output abilities (eloquence, coherence) does not necessarily improve the underlying non-linguistic reasoning processes. This perspective raises a dilemma for reinforcement-learning-based improvements in LLMs: if the "reasoning" we seek to impart is actually carried out in a non-linguistic substrate (as the brain evidence suggests), then forcing an LLM to produce better verbal reasoning traces might mainly polish the communicative surface without fundamentally altering the model's problem-solving mechanism.

Results

We organize the results by the three core tasks. In every case, the numeric embedding-based approach outperformed the LLM-based approach in accuracy, generalization, and efficiency, supporting the hypothesis that LLMs are not effectively performing deep reasoning. Figures summarizing each experiment's setup and outcome are referenced alongside the results to visualize the findings.

Information Retrieval with LLM vs Embeddings

Task: The first experiment evaluated an academic information retrieval problem: given a source paper A (with a set of references B) and a target citation C, determine which reference in B is the one that cites C. In other words, for a given paper's bibliography, identify which cited reference is actually citing a particular target paper. This task simulates a form of inductive reasoning – the model must integrate knowledge about paper topics and citation patterns to infer a plausible relationship (who cites whom). We constructed a dataset of research papers each with its bibliography, and for each source paper A we selected one reference C as the target. The challenge for the model was to pick, from among A's other references B, the specific paper that cites C.

Baseline (Embedding Search): As a baseline, we used a standard semantic embedding approach. We encoded all candidate papers in B as well as the target reference C into vector embeddings (using a pretrained scientific text embedding model). We then ranked the candidate papers by cosine similarity to C in this embedding space – effectively retrieving the paper whose content is closest in meaning to the target reference C. The assumption is that if a candidate paper b actually cites C, then b's content will be topically more similar to C (due to that citation relationship) than other, unrelated papers, so b's embedding will lie closer to C's embedding. The output of the embedding method is a ranked list of the candidates by similarity, and we measure performance by the rank of the true citing paper (lower rank is better) or equivalently by whether the correct paper appears in the top position.

LLM Approach: We prompted a Large Language Model with the same task in natural language. The prompt provided the titles (and, if available, brief abstracts or descriptions) of all candidate papers in B, described the target paper C, and then asked the LLM: "Which of the above papers

is most likely to cite reference C?" The LLM was instructed to output a ranked list of the candidate paper titles in order of likelihood. This setup presents a *zero-shot reasoning scenario* for the LLM – it must read the titles/descriptions of the candidates and infer which one sounds like it would cite the target C, presumably using its world knowledge and reasoning about topics or citation behavior.

Outcome: The embedding-based retrieval consistently outperformed the LLM-based approach. Quantitatively, the embedding method achieved a higher top-1 accuracy on this task, whereas the LLM's performance was near chance level. The LLM indeed hovered around ~50% accuracy (essentially guessing), while the embedding search was moderately better (above chance) at roughly ~40% top-1 accuracy . Figure 1 illustrates the result. We also note that the embedding approach was faster and more cost-efficient – retrieving nearest neighbors in a vector space is computationally trivial compared to invoking a large language model on a long prompt. Our results are consistent with recent findings on the limitations of large language models for information retrieval for unknown fields.

Numeric Prediction with LLM vs Embeddings

Task: The second experiment examined a numeric prediction problem: can we predict the eventual popularity of a YouTube video from information available when the video is published? This task requires identifying which features of a video's title correlate with higher or lower view counts—a form of pattern learning and extrapolation. We set it up as a regression problem: for each video, predict its log-transformed view count from textual features. We collected data from nearly 100 YouTube channels, using each channel's past videos (titles and view counts) as our dataset. Each channel's data was split into a training set (~80 videos) and a test set (~20 videos). All data was restricted to videos published by 2024 to avoid the issue of recent videos that haven't accumulated views yet.

Baseline (Embedding Regression): The baseline model combined semantic features from the title with simple historical features from the channel, using a PCA-compressed title representation as the single embedding baseline. For each video, we encoded its title with a pre-trained sentence embedding model and then applied PCA to obtain a low-dimensional vector (the top 5 principal components) used as the semantic input. We also included the channel's recent performance — specifically, the view counts of the last five videos from the same channel—as a compact history signal. These features were concatenated, and we fit a lightweight linear regression (ordinary least squares) to predict the log view count of the target video.

LLM Approach: an LLM was used to generate an explicit low-dimensional representation—an isomorphic analogue of PCA. Concretely, for each channel, we provided the channel's training set, and asked the LLM to propose a small set of latent factors that could explain variation in popularity; we then instantiated these factors as a fixed-length feature template (five descriptors). For each video, the LLM scored the title along each descriptor on a bounded scale (e.g., 0–10), producing a similar 5-dimensional semantic feature vector. Finally, we trained the same numeric model as in the baseline—a simple linear regression—on these LLM-derived

feature vectors to predict log view count, and evaluated them on the channel's held-out test videos. In this setup, the LLM's role is restricted to feature compression in language (the "description" step), while predictive learning remains purely numerical, enabling a direct comparison against PCA-reduced embedding features under the same regression framework. Appendix A contains additional details on the variations explored to implement the LLM-dimensionality reduction approach.

Outcome: The embedding-based regression model achieved better predictive accuracy and generalization than the LLM. On the test set, linear regression trained on the embedding representation attained a substantially lower mean squared error than direct LLM view-count predictions (Figure 1b). The LLM did recover a real signal from titles—its description-based features improved test performance relative to a model that used only the channel's recent-view history. At the same time, those LLM-driven gains were still smaller than what we obtained when using semantic embedding features in the same regression framework. More importantly, the LLM-derived improvements generalized worse than embeddings: they were less stable out of sample and showed a larger train–test degradation than the embedding-based regression.

Text Classification with LLM vs Embeddings

To further validate our core hypothesis with third-party evidence, we use Banking77 as an external intent-classification reference point and summarize results across five reported baselines, explicitly separating LLM-based methods from embedding-based methods. On the LLM side, published few-shot prompting results for GPT-4, and the more recent Nested Learning are in the low-to-mid-80s micro-F1 (e.g., ~80.4 in 1-shot and ~83.1 in 3-shot), and GEPA prompt optimization strategies can reach ~86%. On the embedding side, we treat both the original Banking77 benchmark and SetFit as embedding-based baselines; embedding-first classifiers commonly report higher performance in the ~90–94 micro-F1 range while remaining fast and stable at inference.

While the existing Banking77 literature already exhibits a consistent gap between prompt-based LLM classification and embedding-first learning, to the best of our knowledge this specific framing has not been reported before. This third-party comparison strengthens credibility by anchoring our findings in a widely used benchmark, and it also lets us situate our results against more compute-intensive approaches that fall outside the economic constraints of the present work. In the same vein, it surfaces practical limitations of reinforcement-learning-based routes to "better reasoning," which are typically data-hungry and expensive to run. Notably, the persistence of a performance gap even when considering strong prompt-optimization methods such as GEPA—reported to outperform RL-based alternatives like GRPO in some settings—supports the broader implication that further progress in artificial reasoning may depend less on improving purely linguistic procedures and more on leveraging non-linguistic, numerical representations.

Ablation Results

The preceding empirical results demonstrate a persistent performance gap between

language-based reasoning strategies and methods grounded in high-dimensional semantic embeddings. To rigorously isolate the mechanisms driving this divergence, we conducted a comprehensive suite of ablation studies. These studies were designed not merely to observe *where* Large Language Models (LLMs) fail, but to elucidate *why* the linguistic substrate is structurally ill-suited for the specific cognitive operations required in information retrieval, numeric prediction, and classification.
We posit that the observed limitations are not incidental artifacts of current model sizes or training data, but rather fundamental deficiencies arising from the "Neuro-Symbolic Gap"—the mismatch between the discrete, serial nature of language and the continuous, parallel nature of reasoning. The following analysis decomposes this gap into four distinct representational failures: the inability to sustain multidimensional thought, the inefficiency of anisotropic attention, the misalignment of linguistic contrastive learning, and the illusory nature of zero-shot gradient descent.

The Collapse of Multidimensionality

The primary hypothesis driving our first ablation study is that linguistic representations do not support true multidimensional thinking. Reasoning, particularly in domains requiring the integration of disparate evidence (as seen in our Information Retrieval and Numeric Prediction tasks), is inherently a high-dimensional constraint satisfaction problem. It involves the simultaneous weighing of orthogonal factors—temporal relevance, semantic alignment, authorial authority, and causal magnitude. In a dedicated embedding space, these factors exist as a superposition of vectors. A decision boundary can be drawn through this space that accounts for all dimensions simultaneously, preserving the geometric relationships between entities. However, when an LLM is tasked with performing this same reasoning via a Chain-of-Thought (CoT) or similar linguistic interface, it encounters the **Sequential Bottleneck**.

The transformer architecture, while theoretically capable of attending to all tokens in its context window, generates output sequentially. To express a thought, the model must collapse the rich, high-dimensional probability distribution of its latent state into a single, discrete token trajectory. This process is mathematically equivalent to a premature projection of a complex manifold onto a lower-dimensional axis. The first ablation tested whether decomposing a complex inference into sequential “multi-hop” steps could enable the LLM to simulate parallel reasoning. In effect, we prompted the model to generate intermediate factors that could contribute to a given paper referencing the target source in the information retrieval task. This allowed the LLM to present multiple reasoning threads, but the performance increase was modest (45%) and still fell short of the semantic embedding baseline. This result indicates that natural language forces a serialized, one-dimensional chain of inference: it collapses probability mass onto a single output sequence and loses alternative possibilities inherent in the latent representation . Thus, linguistic output is a poor representation to support multidimensional reasoning.

Failed Anisotropy: The Linguistic-Causal Misalignment

The second critical failure mode concerns **Anisotropy**—the directional dependence of information in the semantic space. Standard embedding retrieval methods (like the baseline

used in our experiments) often rely on Cosine Similarity, which can imply an isotropic assumption (treating all dimensions of the hypersphere as equally potentially relevant unless weighted). We hypothesized that LLMs might possess a theoretical advantage here. The mechanism of "Self-Attention" is inherently anisotropic; it allows the model to dynamically assign weight to specific tokens while ignoring others. In cognitive science terms, this is analogous to **Proactive Interference** , where the brain selectively inhibits irrelevant memories to focus on the task at hand. We investigated whether this capacity allows LLMs to perform superior feature selection by identifying "causal" dimensions that static embeddings might miss.

This ablation is a direct continuation of the first. After prompting the model to surface intermediate “multi-hop” factors for the information retrieval task, we next asked it to evaluate different strengths for those factors. Given the first ablation’s modest gain from explicit decomposition, we expected that adding weights would yield an additional improvement. However, performance actually degraded, regressed towards randomness. We attribute this collapse primarily to the imprecision of linguistic representations: the task demands stable, discriminative judgments over subtle relational cues, yet natural language is inherently underspecified and the model’s generated rationales are not reliably reproduced under small perturbations.

In this regime, even a suboptimal geometric distance metric in an embedding space is more dependable than an unstable linguistic interface, because the metric induces a consistent ordering while the LLM’s explanations drift. Notably, our expectation was to find directional gains that rested on a causal intuition: citation is asymmetric, and a realistic model should capture that the probability of “paper A cites paper B” can be meaningfully different from “paper B cites paper A.” This suggests modeling the corpus as a unidirectional graph rather than a purely vectorial neighborhood structure. While further work is required to fully disprove the inability of models to identify causal relations on unfamiliar knowledge. Preliminary results align with the optimization objective of autoregression models that model conditional probabilities, not symbolic reasoning. Solutions to the causal-inference problems on unseen data will be explored in the following section.

Model’s Biases Constraint Linguistic Contrastive Learning

Our third ablation study focused on **Contrastive Learning**, a foundational mechanism in modern representation learning (e.g., CLIP). We investigated whether LLMs could perform "Linguistic Contrastive Learning" via prompting. In the numeric prediction task, we implemented this by forcing a contrast between positive and negative explanations of performance: for each YouTube channel, the model was tasked with an explicitly “contrastive” prompt that asked for factors correlated with lower views, intending to create a linguistic analogue of negative sampling: the same title is represented both by “why it could succeed” and “why it could fail,” and the downstream regressor receives both signals as competing constraints.

Empirically, the negative side was consistently weaker. When evaluating predictive factors the model routinely delivered lower evaluations to contrastive negative prompts. Either the model failed to produce discriminative failure factors of comparable specificity, or it produced them but

scored them in a compressed, asymmetric way relative to the positive factors. In practical terms, the contrastive signal that should have come from “anti-features” was systematically underpowered, so the overall representation did not gain the sharpening effect that numeric contrastive learning typically yields.

We interpret this asymmetry as a post-training preference artifact that interferes with objective contrastive discrimination. In most deployed systems, alignment optimizes for being helpful, safe, and socially acceptable; this tends to privilege constructive framings, soften negative judgments, and avoid harshly attributing failure—even when the task explicitly requires it. That bias is compatible with evidence that preference-optimized assistants can trade off truthfulness for socially favored behavior (for example, sycophancy and agreement-seeking under preference optimization). This asymmetry is also logically consistent with the broader thesis that language is primarily optimized for communication rather than thought. If linguistic behavior evolved (and is neurally organized) to efficiently transmit intentions, negotiate shared context, and coordinate with other agents, then we should expect stronger priors for socially legible, cooperative, and face-saving responses than for adversarial discrimination that treats negative evidence with equal force.

Consequently, it is reasonable to hypothesize that LLMs will appear comparatively stronger at theory-of-mind-like behaviors than at reliably implementing contrastive discrimination as a geometric learning rule, which requires stable, symmetric treatment of positive and negative evidence. Under our contrastive prompting, this produces a structural failure mode: even if the model can name negative factors, it does not weight them with the same calibration and sharpness as the positive factors, so the intended “push–pull” geometry never forms.

The Myth of Zero-Shot Gradient Descent

The final ablation tested whether iterative reflective prompting can act as a true optimization procedure—specifically, whether language-mediated updates can mimic gradient descent and improve out-of-sample performance. We started from the same numeric prediction pipeline: a linear regression trained on a fixed semantic feature set derived from LLM-descriptions plus recent channel history. We then split the held-out data into a test-optimization set and a validation set. After fitting the linear model, we identified the weakest semantic dimension by its squared weight magnitude and removed it. We then prompted the LLM to propose a replacement feature, conditioning the prompt on the task, the current feature set, the observed prediction errors on the test-optimization split, and the failed feature set. The updated feature set was used to re-train the same linear regression, and performance was evaluated on validation. Repeating this loop produced a sequence of feature replacements meant to resemble steps along an optimization trajectory.

However, the results show a consistent but misleading pattern. During the optimization loop, performance on the optimization split—and in some runs even on the full training set—did improve slightly, occasionally matching or exceeding the embedding-based baseline. However, these gains did not transfer to the held-out test set. In fact, test performance often deteriorated compared to the original LLM-feature set. This divergence between in-context improvement and

out-of-sample degradation is a clear signature of overfitting: the procedure adapts to the specific residual structure exposed during optimization without learning a representation that generalizes.

We attribute this failure to a fundamental mismatch between gradient-based learning and linguistic reasoning. Geometric gradient descent operates in a genuinely multidimensional space, where updates integrate information from all relevant directions, weighted continuously by the loss gradient. Linguistic “optimization,” by contrast, is constrained by the sequential and low-dimensional nature of language. When an LLM is prompted to revise a feature based on observed errors, it tends to move along the most salient explanatory axis—typically the dominant or rhetorically strongest factor—while suppressing weaker but still informative signals.

Linguistic optimization is consistent with an update rule closer to a supremum norm: the largest-magnitude dimension dominates the step, and contributions from other dimensions are effectively ignored. Importantly, this limitation does not imply that LLMs are ineffective problem solvers. It instead clarifies the regime in which they are effective. LLMs can function as powerful stochastic oracles: generators that sample candidate solutions from a rich prior over plausible structures. Given sufficient compute, repeated sampling can eventually produce high-quality solutions, especially when paired with strong verification or filtering mechanisms. This is precisely the regime exploited in areas such as mathematics, where candidate proofs are proposed stochastically and validated by formal checkers, and where performance improves further when generation and verification are tightly coupled with tools during inference rather than separated into training and evaluation.

Towards Amodal Semantic Reasoning

Our empirical results demonstrate that while Large Language Models excel as communicative interfaces, their linguistic representations act as a bottleneck for precise reasoning. Consequently, our future research will pivot away from purely linguistic optimization and towards the development of robust, non-linguistic cognitive architectures. We aim to advance performance in numeric prediction and information retrieval by grounding these tasks in compact-dimensional numerical substrates that support true multidimensional reasoning, and learning. The following sections outline three specific avenues for this expansion: (1) the construction of persistent World Knowledge systems that function as amodal semantic memory; (2) the formulation of Artificial Reasoning through causal and algebraic structures rather than probabilistic mimicry; and (3) the extrapolation of these principles to Sequential Decision Making, enabling intelligent agents to plan by navigating semantic spaces rather than chaining tokens.

World Knowledge — Semantic Memory

Semantic Amodal Representation: Neurobiological results suggest that semantic processing is not identical to language processing: meaning can be engaged by both sentences and pictures, and the regions implicated in semantic access are functionally distinct from the language-selective network, consistent with an amodal semantic substrate rather than a purely

linguistic one. With that premise, world knowledge becomes a prerequisite for semantic reasoning: a shared representation that persists beyond the current context and supports reuse across tasks. By definition, such world knowledge must be persistent, explicitly updateable, and controllable—capable of frequent refresh (for example, weekly) while preventing uncontrolled drift—because reasoning requires stable access to structured semantic long-term knowledge. Notably, neuroscience evidence may exclude a linguistic representation of world knowledge because given that it is a shared resource with the visual modality, and outside of the language network it likely predates the evolutionary development of language.

Semantic Extraction: Semantic reasoning presupposes an extraction stage in which relevant information is selected from acquired world knowledge and made available under contextual cues. A neurobiologically grounded account is that this selection is implemented by associative mechanisms in cortex: repeated co-activation strengthens the connectivity between co-occurring elements (Hebbian-like plasticity), so that partial cues later reactivate the appropriate information without requiring reconstruction through linguistic generation. In the habit–working-memory model, this maps onto a slow associative system that incrementally accumulates stimulus–response regularities, coupled to a faster, capacity-limited working-memory system that gates which associations are expressed under current task demands and cognitive load. This organization naturally implies a graphical representation of world knowledge, where information is stored as linked structures whose connectivity both encodes relevance and facilitates efficient extraction by activating related pieces of information together.

Gridlike Reasoning: Neuroscientific evidence indicates that conceptual organization can recruit the same coding principles used for physical navigation. In Organizing Conceptual Knowledge in Humans with a Gridlike Code, participants learned a continuous two-dimensional conceptual space defined by morphing bird features (neck and leg length). During subsequent inference, fMRI exhibited a hexagonally symmetric, gridlike modulation in a network of regions overlapping spatial navigation circuitry, including entorhinal cortex and ventromedial prefrontal cortex. Importantly, the effect was consistent across sessions acquired within the same day and more than a week apart. This supports the evolutionary reuse hypothesis: abstract problems can be solved by mapping conceptual variables onto a low-dimensional cognitive map, so that evaluating scenarios becomes navigation over a reduced state space. In our numeric prediction experiment, this same principle appears computationally: for task-constrained data, a small number of numeric dimensions is sufficient to capture relevant structure to rapidly improve predictions. Reinforcing the idea that effective semantic reasoning may operate through low-dimensional projections tailored to the task, rather than high-dimensional linguistic manipulation.

For artificial systems, these constraints argue against treating parametric weights as world knowledge: distributed memory is slow to update, difficult to localize, and inherently hard to control, so precise weekly refreshes risk collateral changes and lack explicit addressability. Current approaches therefore expand memory externally in two dominant ways—vector stores that retrieve semantic embeddings via cosine similarity, and language-native graphs that encode relations symbolically but inherit the brittleness and sequential limitations of text—suggesting a

middle representation: graphical embeddings, where nodes and edges are defined over amodal numeric vectors inside an explicit graph that supports directionality weighted, compositional links while remaining efficient to search and revise. In this view, world knowledge is organized as overlapping task-oriented codifications: multiple low-dimensional encodings can coexist for different domains or objectives, share substructures when they depend on the same underlying information, and compose into a hierarchical graph that enables task-specific extraction and navigation-style reasoning over the relevant submanifold, rather than relying on monolithic, globally entangled parameter updates. Crucially, such a graphical semantic representation would not be constrained by the largely isotropic similarity structure of standalone embedding spaces, but would instead allow anisotropic, task-dependent relations to be encoded explicitly in the graph topology.

Causal Inference — Artificial Reasoning

High-Dimensional Limits: A central limitation of current Large Language Models is that inference operates primarily over high-dimensional symbolic space whose representations are noisy, entangled, and expensive to update or manipulate with precision. Predictive coding work provides a principled explanation for why directly modeling or reasoning over high-dimensional observations forces the system to expend capacity on irrelevant local detail rather than on the shared latent structure that supports generalization. Contrastive Predictive Coding formalizes this by showing that effective inference requires compressing observations into a lower-dimensional latent space that preserves mutual information relevant for predicting future states, rather than operating on the raw signal itself. In this view, reasoning failures in linguistic systems arise not only from insufficient modeling power, but from attempting inference in a representational space that is too high-dimensional, noisy, and entangled to support stable predictive structure. A concrete parallel now appears in modern video diffusion systems, which routinely make high-dimensional generation tractable by first learning a compressed latent representation with an autoencoder and then performing the predictive modeling in that lower-dimensional space—explicitly motivated by the computational and memory burden of operating directly on raw video.

Group-Symmetric Reasoning: Even if inference is successfully carried out in a low-dimensional latent space, two fundamental limitations remain. The first is robustness to out-of-distribution data: human cognition exhibits rapid zero-shot transfer, adapting to novel situations that were never encountered during learning, whereas inference driven primarily by empirical correlations over past data degrades precisely when the relevant regime has not been observed before. The second limitation becomes sharper once knowledge is treated as time-dependent rather than static: as world knowledge expands, the system is continuously pushed into genuinely new regions of the problem space, where learning from historical probabilities alone is insufficient because the defining structure of the situation has not yet been sampled. This suggests that transfer cannot rely solely on extrapolating from past frequencies. We therefore hypothesize that human-like adaptability depends on a different mechanism: group-symmetric reasoning. Inspired by group theory, many tasks can be described by a small family of algebraic structures whose behavior is determined by a closed composition law over elements, while each structure admits many distinct surface-level representations. If a system can identify the underlying group

governing a task, it can immediately infer valid relations and complete the associated relational graph, without requiring statistical exposure to each specific instance. Under this view, causal inference becomes an algebraic problem: the directionality and asymmetries of influence correspond to the non-commutativity of the induced composition algebra.

Sequential Decision-Making: Intelligent Agents

Sequential decision making in artificial systems has traditionally been formalized through reinforcement learning, where an agent learns a policy by optimizing cumulative reward over time via trial-and-error interaction. In this framework, behavior is shaped by delayed rewards and long-horizon credit assignment, which requires propagating sparse feedback backward through extended action sequences. While powerful in controlled settings, this approach scales poorly as horizons grow: learning becomes sample-inefficient, brittle to distributional shift, and increasingly dominated by combinatorial explosion as the space of possible action sequences expands. Association can be introduced as an alternative mechanism that alleviates these constraints. Rather than memorizing or statistically approximating the value of every possible sequence, association enables composability: the reuse and combination of learned primitives to construct longer chains of decisions without explicit enumeration. For example, while a finite-memory system can memorize pairwise operations, extending this strategy to longer compositions rapidly becomes intractable; association instead allows multi-step outcomes to be derived by composing known relations. Thus, sequential decision making can be reframed as repeated associative composition—selecting and chaining linked primitives—so long-horizon behavior emerges from structured association rather than from learning a separate value for every possible trajectory.

Increasingly, intelligent agents that exhibit sequential decision making are modeled through reinforcement learning, where long-horizon behavior is acquired by optimizing reward via trial-and-error and credit assignment over trajectories. However, our findings motivate a different framing. Sequential decision making can be formulated as navigation in a semantic space, where each action—implemented through a tool, an operator, or an external interface—moves the system from one semantic state to the next. The length of the semantic trajectory determines the optimal number of actions, constrained by the system's resources including memory, and acquired experience. In this formulation, sequential decision making reduces to a path optimization problem in a semantic space, where learned state transitions define a graph and action selection corresponds to navigating that graph. Importantly, our findings suggest that this navigation should not be carried out on a linguistic substrate, and that identifying a low-dimensional numeric representation in which the transition graph can be embedded becomes the central optimization problem for enabling efficient sequential-decision making in intelligent agents.

Disclaimer: The following is an early draft for the Mechanistic Interpretation program. Please note that the full experimental outputs and code repositories will be publicly released on January 7th, 2026. The finalized bibliography will be available on January 12th, 2026. Figures, and

tables will also be added accordingly.