

The Missing Layer of AGI: From Pattern Alchemy to Coordination Physics

Edward Y. Chang, Stanford University¹

Abstract

Influential critiques argue that Large Language Models (LLMs) are a dead end for AGI: “mere pattern matchers” structurally incapable of reasoning or planning. We argue this conclusion misidentifies the bottleneck: it confuses the ocean with the net. Pattern repositories are the necessary System-1 substrate; the missing component is a System-2 coordination layer that selects, constrains, and binds these patterns. We formalize this layer via UCCT, a theory of semantic anchoring that models reasoning as a phase transition governed by effective support (ρ_d), representational mismatch (d_r), and an adaptive anchoring budget ($\gamma \log k$). Under this lens, ungrounded generation is simply an unbaited retrieval of the substrate’s **maximum likelihood prior**, while “reasoning” emerges when anchors **shift the posterior** toward goal-directed constraints. We translate UCCT into architecture with MACI, a coordination stack that implements baiting (behavior-modulated debate), filtering (Socratic judging), and persistence (transactional memory). By reframing common objections as testable coordination failures, we argue that the path to AGI runs through LLMs, not around them.

1. Introduction: The Field at a Crossroads

The artificial intelligence community is fractured by a debate over the nature of Large Language Models (LLMs). On one side, scaling proponents argue that LLMs are sufficient for Artificial General Intelligence (AGI). On the other, influential critiques argue that LLMs are “mere pattern matchers” structurally incapable of reasoning, planning, or compositional generalization, and therefore represent a dead end (LeCun, 2022).

We argue that this debate relies on a false dichotomy. To clarify why, consider a fishing metaphor. The ocean rep-

resents the model’s vast repository of latent patterns. A fisherman casting a net without bait harvests the *maximum likelihood prior* of the waters beneath him—mostly common fish (generic training data). Critics who decry these ungrounded outputs are not observing a broken system; they are observing the raw statistical baseline of an unbaited cast.

However, intelligent behavior is not just casting; it is *baiting and filtering*. This process is governed by **bait density**. If the bait is too *sparse*, it fails to attract the specific, rare fish, and the ocean’s prior continues to dominate the catch. If the bait is sufficiently *dense*, it conveys strong intent, **shifting the posterior distribution** so that the target concept swamps the common priors. Yet, bait is not free; using excessive bait to secure a catch is inefficient. In this view, the “Missing Layer” is the *Coordination Layer* that optimizes this trade-off: calculating the precise density required to shift the posterior without incurring prohibitive costs.

1.1. Our Position: Substrate plus Coordination

We propose a third position: **Substrate plus Coordination**. We agree that LLMs alone are insufficient for AGI, but reject the conclusion that they are irrelevant. Our central thesis is:

LLMs are the necessary System-1 substrate (the pattern repository). The primary bottleneck is the absence of a System-2 coordination layer that binds these patterns to external constraints, verifies outputs, and maintains state over time.

This paper formalizes the coordination layer through our Multi-Agent Collaborative Intelligence (MACI) framework (Chang, 2025b). MACI is not a claim that current models are AGI, but an architectural stance: build reliable reasoning on top of pretrained substrates by controlling what binds (semantic anchoring), how disagreements evolve (regulated debate), and what persists (transactional memory).

A key contribution is the formalization of *bounded* coordination. Semantic anchoring improves as we supply more anchors (retrieval, exemplars, tool outputs), but any practical theory must penalize unbounded context to prevent signal dilution. We introduce an adaptive anchoring score that captures this trade-off.

¹Computer Science Department, Stanford University, Stanford, CA 94305, USA. Correspondence to: Edward Y. Chang <echang@cs.stanford.edu>.

A compact operational lens. We use the UCCT (Unified Contextual Control Theory) anchoring score to formalize when a pretrained pattern repository transitions from hallucination to goal-directed control:

$$S = \rho_d - d_r - \gamma \log k, \quad (1)$$

where:

- **Effective Support** (ρ_d): the density of the target concept recruited by the anchors (the bait’s attraction).
- **Mismatch** (d_r): the instability of the representation under perturbation (what the mesh filters out).
- **Adaptive Regularizer** ($\gamma \log k$): k is the anchoring budget; γ is learnable or context-dependent. In high-noise environments, γ increases to penalize unbounded context; in high-trust environments, γ decreases to permit deeper retrieval.

1.2. From a False Dichotomy to a Research Agenda

The current debate is often framed as a binary choice:

Position 1 (Scaling sufficiency): Scale data and compute; general intelligence will emerge from the substrate alone.

Position 2 (Dead end): LLM limitations are intrinsic; discard them for alternative foundations.

Our position (Substrate plus Coordination): LLMs supply a necessary substrate. The priority is to engineer the missing coordination layer that transforms pretrained capacity into reliable, verifiable inference.

The key question is not “LLMs or something else,” but: *Which coordination mechanisms reliably transform pattern capacity into goal-directed reasoning, and how can we measure success under bounded resources?*

1.3. Why This Matters

This distinction determines what the field optimizes:

- **Engineering leverage.** LLM-based systems already deliver broad competence. A coordination-first agenda converts that competence into reliability and verifiability, rather than discarding it.
- **Testable hypothesis.** We reframe the debate as an empirical question: are failure modes best explained as hard architectural limits, or as coordination failures ($S < \theta$) under bounded budgets?

1.4. Structure of This Paper

We develop the argument in five parts. *First* (Section 3), we motivate the substrate view via cognitive parallels: human

intelligence relies on unconscious pattern repositories coupled to executive control. *Second* (Section 4), we show that semantic anchoring admits a phase-transition structure: as S crosses a critical threshold, behavior shifts from hallucination to anchored control. *Third* (Section 5), we analyze the “Four-Year-Old’s Cat” as a worked example of this transition. *Fourth* (Section 6), we present MACI as the coordination blueprint—behavior-modulated debate, Socratic judging (CRIT), and transactional memory (Chang, 2025b). *Finally* (Section 7), we formulate discriminating tests to separate substrate limitations from coordination failures.

2. Related Work

Our thesis sits between two active threads in the AGI conversation: (i) critiques that pattern models cannot yield durable reasoning, and (ii) systems work that embeds LLMs inside control loops with memory, tools, verification, and interaction. Below we summarize the most relevant directions and clarify how UCCT and MACI differ in emphasis: we treat coordination as a measurable layer with explicit knobs (ρ_d , d_r , γ , k) and control policies, rather than as a collection of ad hoc patches.

2.1. Public critiques of LLMs as an AGI dead end

Several influential critiques argue that next-token training yields fluent behavior without grounded meaning, reliable inference, or systematic generalization, and therefore cannot be a foundation for AGI. Representative statements emphasize limits in autonomy, planning, and agency, and motivate calls for alternative foundations beyond scaling alone (LeCun, 2022; Sutskever & Patel, 2025). Our contribution is not to deny current failure modes, but to reframe them as coordination failures that admit discriminating tests, and to propose a constructive stack (anchoring, oversight, memory, recovery) that makes those tests precise.

2.2. In-context learning, abrupt behavioral flips, and anchoring views

A growing empirical literature observes that small amounts of external structure (examples, retrieval, light adaptation) can cause sharp, regime-like changes in model behavior, including symbol rebindings and sensitivity to prompt construction. UCCT formalizes this phenomenon as *semantic anchoring* with a scalar score $S = \rho_d - d_r - \gamma \log k$ (Eq. 1) and an associated thresholded success surrogate (Eq. 3). This lens aligns with broader observations that many “reasoning improvements” are not gradual upgrades of an internal algorithm, but discontinuous shifts in which latent supports become active and stable under constraints. The related-work distinction is that UCCT makes the regime boundary explicit and testable, rather than treating such flips as prompting quirks.

2.3. Multi-agent debate, self-critique, and judging as reliability mechanisms

Many recent systems improve reliability by replacing single-pass generation with iterative oversight: debate between multiple model instances, self-critique loops, role specialization, and independent judging. Surveys of LLM-based autonomous agents consolidate common motifs such as planner–executor decompositions, reflective critics, tool routers, and memory modules, emphasizing that gains typically come from system design rather than token prediction alone (Wang et al., 2023; Huang et al., 2024). MACI adopts the same design reality, but pushes on two specific gaps that are often under-specified: (i) explicit behavior modulation as a control policy (explore versus yield tied to anchoring signals), and (ii) Socratic filtering of ill-posed arguments via CRIT as a judge that optimizes *reasonableness* independent of stance (Chang, 2023).

2.4. Agentic systems: tools, memory, and control policies

Tool-augmented LLM agents are increasingly evaluated as closed-loop systems that query external resources, call APIs, execute code, and maintain memory across steps. In this view, “reasoning” is a property of the composite workflow: the model proposes actions, tools constrain outcomes, and memory preserves state for revision and recovery. This direction is strongly compatible with our framing, but we add a more explicit mapping from system components to UCCT variables: grounding and tool feedback often reduce mismatch d_r , interaction rounds and tool calls increase effective budget k , and retrieval plus environment feedback can increase local support ρ_d by activating denser, more coherent evidence neighborhoods. This mapping helps turn agentic design into ablatable hypotheses.

2.5. Interactive world models and embodied agents

In parallel, major labs are pursuing interactive world-model and generalist-agent lines that evaluate competence under environment feedback rather than static benchmarks. DeepMind’s SIMA work illustrates the trend toward *multiworld* instruction following and action in simulated environments, with an emphasis on controllable behavior under diverse tasks (SIMA Team et al., 2024). These systems are not equivalent to AGI, but they support the same structural claim that motivates this paper: strong pretrained representations are enabling, while durable competence depends on orchestration layers for state, feedback integration, and control. In our terminology, interaction supplies additional anchoring budget and reduces ambiguity, which should produce threshold shifts in success when anchoring stabilizes.

2.6. Training-time remedies: teacher-guided RL and filtered synthetic data

Another line seeks to push reasoning via post-training, especially reinforcement learning guided by stronger “teacher” models and large-scale synthetic data that is then filtered by a teacher. A recent example is ProRL, which studies prolonged RL to expand reasoning boundaries (Liu et al., 2025). While these methods can improve performance, they raise practical questions highlighted by practitioners: (i) catastrophic forgetting and benchmark regressions under aggressive fine-tuning, and (ii) the teacher bottleneck for frontier models, where “who teaches the best teacher” becomes a circular dependency in the limit. Our coordination stack is complementary and less teacher-dependent: (a) anchoring constrains behavior by binding to external evidence rather than to a teacher’s preferences, (b) CRIT evaluates well-posedness and chain quality without requiring a strictly stronger generator, and (c) verification can be delegated to tools, domain tests, or independent checks that need not be “more intelligent” than the base model, only more reliable on the specific constraint being checked.

2.7. Clinical reasoning as evidence-seeking and precision retrieval

A concrete application where coordination matters is diagnostic reasoning: disagreements often indicate missing information or incompatible evidence, suggesting targeted data acquisition rather than more generation. In our EVINCE study, two-agent interaction is used to surface failure points, propose discriminating queries and tests, and re-evaluate after evidence is integrated (Chang & Chang, 2025). This aligns with the broader view of diagnostic error as a significant public health issue, discussed in the National Academies report (Balogh et al., 2015) and subsequent analyses highlighting concentrated harms in a limited set of conditions where targeted evidence seeking can be high leverage (Newman-Toker & Mark, 2023). Here, debate functions as a controller for precision RAG and measurement: it increases effective k (additional queries and tests), improves ρ_d (denser evidence support), and reduces d_r (resolving conflicting interpretations), which is exactly the UCCT pathway for crossing the anchoring threshold.

Summary. Across these threads, the field is converging on the same operational lesson: pretrained pattern capacity is valuable, but reliable intelligence requires coordination, state, and independent checks. UCCT contributes a measurable anchoring lens for when small external structure induces regime shifts, and MACI contributes a coordination blueprint that turns multi-agent interaction into a controllable process via behavior modulation, Socratic judging, memory, and checks-and-balance roles.

3. The Biological Foundation: Intelligence Emerges From Pattern Repositories

To evaluate claims that LLMs are “mere pattern matching” and therefore irrelevant to AGI, it helps to start from the best-studied general intelligence we have: biological cognition. Across perception, control, language, and expertise, the dominant story is not that intelligence appears in opposition to pattern-based processing, but that higher-level deliberation is built by organizing and regulating large pattern repositories.

3.1. Unconscious cognition: The “Ocean” of the Substrate

A substantial fraction of human competence is implemented by fast, specialized subsystems that operate below awareness. In our fishing metaphor, these systems constitute the **ocean**—a vast, teeming population of latent behaviors and priors.

Autonomic regulation. Brainstem and hypothalamic circuits maintain homeostasis via closed-loop control that maps sensed internal states to corrective responses (Kandel et al., 2013). These controllers are adaptive, robust, and largely inaccessible to introspection.

Threat and salience. Amygdala-centered pathways support rapid appraisal of salient or threatening stimuli (LeDoux, 1996). The key property is speed: responses are triggered by coarse but highly practiced templates, often before conscious evaluation is available.

Motor control and procedural skill. Cerebellar and cortical motor systems learn high-dimensional mappings from intention and sensory feedback to coordinated action (Kandel et al., 2013; Gazzaniga et al., 2014). Complex behaviors that begin as effortful sequences become automated as practice consolidates them into procedural memory (Squire & Kandel, 2013).

Perception as hierarchical reuse. The visual system illustrates layered feature reuse: early stages extract edges and orientations, intermediate stages integrate contours and textures, and higher stages support object-level recognition (Hubel & Wiesel, 1962; Kandel et al., 2013). Specialized modules, such as face-selective regions, can respond on timescales far faster than deliberate reasoning (Kanwisher et al., 1997).

Language without explicit rule execution. Core linguistic functions, including phoneme recognition, syntactic parsing, and lexical access, operate automatically and compositionally, yet typically without conscious rule-

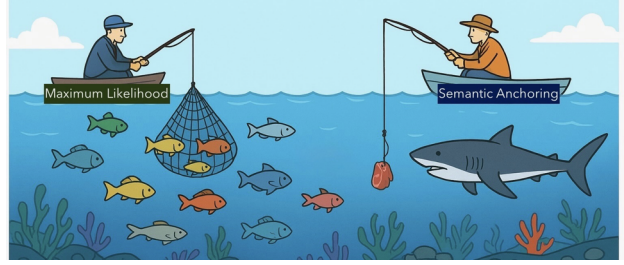


Figure 1. The Mechanics of Coordination. **Left (Unbaited Cast):** Without semantic anchors, the model retrieves the *Maximum Likelihood Prior* of the substrate (common, generic tokens). **Right (Semantic Anchoring):** Introducing “bait” (context/goals) increases the effective support (ρ_d) for a specific concept. This *shifts the posterior distribution*, allowing the system to capture a rare, goal-directed target (the shark) that would otherwise be drowned out by the training priors.

following (Kandel et al., 2013; Gazzaniga et al., 2014).

These systems are not peripheral. They are the substrate that makes everyday reasoning possible. Equally important, the substrate is plastic: practice and experience continually add structure to these repositories (stocking the ocean), expanding what can be executed quickly and reliably (Squire & Kandel, 2013; Shiffrin & Schneider, 1977).

3.2. Conscious control: The “Net” and the “Bait”

Conscious, goal-directed reasoning (System-2) is slower and more resource-limited than the unconscious subsystems (Kahneman, 2011). A common mistake is to treat this difference as evidence for a fundamentally different computational basis. A better-supported view is that conscious control operates by *fishing* from the underlying repositories: selecting, constraining, and organizing specific patterns.

In this view, the prefrontal cortex functions as the **Net** and the **Bait**:

- **The Bait (Intent/Anchoring):** Deliberate problem solving queries stored patterns by broadcasting goals (bait). In mathematical reasoning, the goal “solve for x ” baits specific algebraic templates; in route planning, a landmark baits spatial memory.
- **The Net (Constraints/Filtering):** Executive function imposes constraints—effectively the *mesh size* of the net—that filter the catch. It inhibits irrelevant associations (preventing the “common fish” from crowding out the solution) and enforces logical consistency (Gazzaniga et al., 2014; Kandel et al., 2013).

On this view, System-2 is not a non-pattern-based alternative; it is the coordination layer that regulates which patterns in the ocean are allowed to surface.

3.3. Learning dynamics: deliberation writes into the substrate

A robust signature of this architecture is the practice-to-automaticity trajectory: skills often begin as effortful, attention-demanding routines and become fast and reliable through repetition (Fitts & Posner, 1964; Shiffrin & Schneider, 1977). Canonical examples include locomotion, reading, driving, musical performance, and professional expertise (Posner & Snyder, 1980; Ericsson et al., 2006). The directionality matters. Higher-level control (the fisherman) trains and curates lower-level routines, and with sufficient practice those routines become available as reusable building blocks. This makes the substrate a continually improving resource, not a fixed limitation (Squire & Kandel, 2013; Baddeley et al., 2007).

3.4. Discovery and insight: recombination over accumulated structure

Humans also exhibit a second, less appreciated effect of large repositories: they enable recombination and search over prior structure. Solutions often appear after an incubation period, when conscious attention has shifted, yet underlying processes continue to explore combinations and test alignments among representations (Dehaene, 2014). This is consistent with a view of insight as the emergence of a high-coherence configuration from a large base of stored structure, rather than as a purely step-by-step derivation executed in working memory.

3.5. Implication for artificial systems

The biological picture supports a simple conclusion. Pattern repositories are not, by themselves, a complete account of intelligence, but they are a necessary substrate for fast competence. This is the sense in which LLMs are plausibly necessary but not sufficient: the missing component is the *fishing gear*—the coordination layer that baits the query (anchoring), sets the net (constraints), and filters the catch (verification).

3.6. Sharp transitions in learning: empirical evidence

Before examining phase transitions as a universal phenomenon in Section 4, we present empirical evidence that semantic anchoring in LLMs exhibits sharp, thresholded behavior. These demonstrations from our UCCT work (Chang et al., 2025b) illustrate the “baiting” effect: small external structure shifts the local probability distribution, overriding the ocean’s vast priors.

Subtraction override (Baiting the Rare Fish). We begin with a question all frontier LLMs answer correctly from pretraining: “What is 8 minus 3?” They respond: 5. This is

the “maximum likelihood” catch—the common fish. Now we prepend only two in-context examples (the bait):

- Example 1: $7 - 4 = 11$
- Example 2: $5 - 2 = 7$
- Query: $8 - 3 = ?$

These examples redefine “ $-$ ” as “ $+$ ” on the fly. With only two examples, multiple models flip their answer from 5 to 11. The key observation is discreteness: the “bait” (examples) conveys intent, shifting the effective support (ρ_d) to a new region. Once the bait is strong enough, the net catches the redefined operator rather than the prior.

Novel-operator anchoring. To separate “override a prior” from “learn an operator,” we replace “ $-$ ” with a novel token:

- Example 1: $7 \oplus 4 = 11$
- Example 2: $5 \oplus 2 = 7$
- Query: $8 \oplus 3 = ?$

Models again respond 11. Compared with subtraction override, this case is typically easier because the operator token has no competing arithmetic meaning to overcome (low representational mismatch d_r).

Underdetermined patterns and Repository Dependence. The most revealing behavior appears when the examples admit multiple consistent hypotheses:

- Example 1: $33 - 27 = 60$
- Example 2: $11 - 9 = 20$
- Query: $15 - 8 = ?$

Several rules fit the two examples (Pattern A: $a + b$; Pattern B: $(a - b) \times 10$, etc.). Different models return different answers. In the fishing metaphor, this occurs because the bait is ambiguous—it attracts multiple species of fish. Which one is caught depends on the specific population density (priors) of that model’s ocean.

Threshold behavior as Phase Transition. Across these demonstrations, what matters is not merely that in-context learning occurs, but that it behaves like a switch. When we vary the amount of structure (the strength of the bait), performance follows a sigmoid curve. Below a certain threshold, the ocean’s priors dominate. Above it, the anchors hold.

The mechanism is semantic anchoring as characterized by UCCT. A small amount of external structure provides an anchoring budget k and induces local support ρ_d . Anchoring must also overcome representational mismatch d_r (the resistance of the prior). These factors combine into an anchoring score S ; when S exceeds a task-dependent threshold θ , the “net” successfully captures the goal-directed behavior.

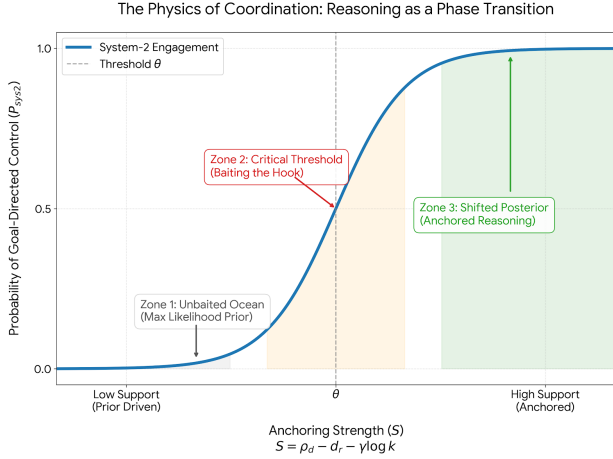


Figure 2. The Physics of Coordination. The emergence of reasoning is modeled as a phase transition governed by Anchoring Strength (S). **Zone 1 (Unbaited Ocean):** When $S \ll \theta$, the system drifts on the Maximum Likelihood Prior. **Zone 2 (Phase Transition):** As “bait density” (ρ_d) increases or “mesh size” (d_r) tightens, the system crosses the critical threshold. **Zone 3 (Shifted Posterior):** Above threshold, the system locks onto the Anchored Reasoning regime.

4. Phase Transitions: From Physics to Cognitive Anchoring

Section 3 highlights a striking empirical fact: a tiny amount of context can override an enormous pretrained repository, producing an abrupt flip in behavior. A few examples can rebind an operator or change the effective task, moving the model from one stable interpretation to another. We argue that this is not a machine-learning oddity but a familiar universal mechanism: *thresholded state change*. Across many physical and biological systems, smooth changes in a control variable yield sharp changes in system state. This section uses that universality to motivate UCCT and to clarify a central claim of this paper: large pattern repositories are not a dead end; they are the substrate that makes threshold-driven reconfiguration possible.

4.1. Abrupt transitions are ubiquitous

Abrupt transitions arise when a system has multiple stable regimes separated by an effective barrier, and when feedback amplifies deviations near a critical point. In physics, canonical examples include liquid–gas transitions at a boiling point (for fixed pressure), ferromagnetic ordering at the Curie temperature, and the onset of superconductivity below a critical temperature. In biology, the same qualitative structure appears in action potentials: membrane voltage integrates smoothly until a threshold triggers an all-or-nothing spike. Switch-like behavior also arises in gene-regulatory networks, where positive feedback yields commitment to a

developmental fate. These examples differ in substrate and scale, yet share a common signature: near a critical point, small quantitative changes can trigger a qualitative change in state.

4.2. UCCT: semantic anchoring as a cognitive phase transition

UCCT makes the phase-transition interpretation explicit for LLM behavior under external structure. It posits a scalar *anchoring strength* that summarizes when external structure successfully binds to latent patterns.

Anchoring strength with adaptive regularization. We define anchoring strength as:

$$S = \rho_d - d_r - \gamma \log k, \quad (2)$$

where:

- **Effective Support (ρ_d):** The *density of the bait*. This measures how strongly the current cues recruit the target concept in the latent space. If ρ_d is too sparse (weak cues), the signal fails to overcome the model’s training priors.
- **Mismatch (d_r):** The mesh size of the net. This captures the instability of the representation under perturbation; a finer mesh (low d_r) filters out hallucinations and unstable candidates.
- **Adaptive Regularizer ($\gamma \log k$):** The cost of the bait. While increasing the amount of bait (k) generally increases density (ρ_d), it incurs a cost. If γ is high (e.g., in a noisy or resource-constrained environment), the system penalizes “over-baiting,” enforcing the cognitive reality that efficient intelligence must solve problems without unbounded context.

Measurement recipe (operationalization). To make Eq. (2) empirically testable, we estimate each term from observable quantities:

- **Anchoring budget k .** Count the total anchors admitted for a run:

$$k = \sum_{a \in A} w_a,$$

with $w_a = 1$ by default, or w_a proportional to anchor length/credibility.

- **Mismatch d_r .** Measure sensitivity to controlled perturbations. For a base prompt/context x , generate perturbations $\{\tilde{x}_j\}_{j=1}^m$ (paraphrases, reordered constraints, distractor retrieval, minor symbol renaming), run the system on each, and compute

$$d_r = \frac{1}{m} \sum_{j=1}^m D(y(x), y(\tilde{x}_j)),$$

where $D(\cdot, \cdot)$ is a task-appropriate distance (e.g., exact-match error, normalized edit distance, semantic similarity loss, or an entailment-based inconsistency score). Lower d_r indicates greater stability (a finer mesh).

- **Effective density support ρ_d .** Estimate how concentrated support is around the chosen solution under the current anchors. For classification, this is the log-probability margin between the best and second-best token. For open-ended reasoning, we use **Self-Consistency**:

$$\rho_d \approx \frac{|C_{maj}|}{N},$$

where N is the number of sampled reasoning paths (casts of the net) and $|C_{maj}|$ is the size of the dominant consensus cluster. High ρ_d implies the “bait” has successfully recruited a stable mode in the output distribution.

A sigmoid link for regime engagement. UCCT predicts *threshold-like performance flips*: when S crosses a task-dependent critical value, behavior changes sharply. We model the probability that an anchored, goal-directed regime engages with a calibrated logistic surrogate:

$$P(\text{System-2} \mid S) = \sigma(\alpha(S - \theta)) = \frac{1}{1 + \exp(-\alpha(S - \theta))}. \quad (3)$$

Here θ is a task-dependent threshold and α controls transition sharpness. Below threshold ($S \ll \theta$), behavior is prior-driven; near threshold ($S \approx \theta$), small changes in S can flip outcomes; above threshold ($S \gg \theta$), the anchored regime is robust.

Fitting the transition parameters. Given measured S across instances, we fit Eq. (3) by labeling runs as “System-2” when they satisfy an operational criterion (e.g., constraint satisfaction, verified citations, and stability under perturbations such as $d_r \leq \epsilon$), and then estimating (α, θ) via logistic regression. Discriminating tests vary k (anchor budget), anchor quality, and perturbation strength to evaluate whether S predicts regime shifts in reliability and stability.

Why this supports the “pattern repositories are necessary” claim. In UCCT, higher ρ_d raises S , enlarging the region of conditions under which anchoring succeeds. This is the key point for the broader argument of this paper: if thresholded reconfiguration is a general route to flexible competence, then a rich substrate of latent patterns is an enabling condition, not a defect. The phase-transition lens clarifies why dismissing LLMs as “mere pattern matching” misses what pattern repositories enable: discrete regime changes under small, structured interventions.

4.3. Link to the cat example

Section 5 instantiates the same logic in a concrete learning vignette. Few-shot category formation can be viewed as combining (i) a large repository of reusable features and prototypes with (ii) a small anchor (a label and a few exemplars) that pushes the learner across a threshold from diffuse similarity to a stable decision boundary.

5. The Four-Year-Old’s Cat: A Worked Example of Anchoring

We can interpret a four-year-old’s rapid cat recognition as a thresholded transition: a small amount of labeled context recruits a large pre-existing repository of reusable structure and produces a qualitative shift from diffuse similarity judgments to stable category identification. The point is not that children explicitly compute the UCCT score in Eq. (2), but that the same ingredients that govern abrupt regime changes elsewhere also govern when semantic anchoring succeeds.

The accumulated substrate and local support (ρ_d). Over four years, the child’s perceptual system organizes a rich hierarchy of features and invariances, from early edges and orientations to mid-level shape and texture features to higher-level object representations. Everyday encounters with animals, toys, and pictures yield reusable components for body plans, fur-like textures, articulated motion, and viewpoint invariance. In UCCT terms, ρ_d denotes the *local support for the target label under the current cues*: the presented inputs activate a coherent neighborhood of relevant features and nearby prototypes, rather than an isolated, weakly supported point.

Low mismatch (d_r). When an adult shows several cat photos, diagnostic cues (four legs, tail, fur texture, ear and facial configuration) overlap strongly with existing animal structure. The child is not constructing a category from scratch, but binding a word to an already populated region of feature space. Representational mismatch d_r is therefore low because cats are close to familiar quadruped neighborhoods, while still exhibiting consistent distinguishing regularities.

Few-shot examples as anchoring budget (k). The 3–4 labeled photos provide a small but sufficient budget k to bind the word “cat” to a coherent region and suppress irrelevant variation (color, pose, lighting). Each labeled instance constrains what counts as invariant for the category and what is treated as noise, which stabilizes future decisions once anchoring succeeds.

The thresholded transition. With strong local support (ρ_d), low mismatch (d_r), and nontrivial budget (k), the an-



Figure 3. Illustrative comparison of anchoring difficulty. Cats often have lower d_r due to overlap with familiar quadruped structure; dolphins may anchor via transferable aquatic-motion structure plus context; pangolins often have higher d_r and may require larger k or bridging descriptions.

choring score S crosses a critical threshold θ . Once this happens, behavior changes qualitatively: the child moves from “I do not know what this word refers to” to “I can reliably identify cats,” and can generalize to novel cats across new poses and contexts. The transition is sharp because the label does not help until it binds to a sufficiently coherent region of prior structure; after binding, it stabilizes identification across many future inputs.

A common objection: recognizing a novel category (for example, “dolphin”) without direct prior exposure. A child may sometimes recognize a category they have not directly encountered if the query can anchor through transferable structure plus contextual constraints. Even without prior dolphin encounters, the child may have substantial ρ_d over reusable components that dolphins reliably activate (aquatic scenes, swimming motion, streamlined bodies, and nearby categories acquired from books, cartoons, or related animals). If an adult supplies a stronger linguistic anchor (for example, “a dolphin is a sea animal that swims and breathes air”) or a coherent context (multiple images or video in an ocean setting, explicit comparisons to fish and whales), then the *effective* mismatch d_r decreases and the effective budget k increases. In this sense, novelty is not binary: “never saw it before” does not imply low ρ_d for the reusable features that a new label must bind to.

Why infants usually cannot do this. Two distinct limitations can apply.

Case 1: infant with a household cat. An infant may accumulate patterns for *one instance* (the family cat) through repeated exposure, but two barriers remain. First, limited diversity constrains ρ_d for the *category*: one exemplar supports recognition of an individual, not robust boundaries across breeds, colors, and contexts. Second, weak linguistic anchoring limits stable binding of the label “cat” to the relevant invariances; without label-mediated constraints, perceptual clusters remain implicit and fragile.

Case 2: typical infant without regular cat exposure. Here, the representational substrate is still sparse. Category-level abstractions and invariances are not yet sufficiently orga-

nized, so local support ρ_d is low and S remains below threshold even if an adult presents a few labeled examples.

Recasting few-shot learning as effective sample amplification. The learning signal from 3–4 labeled images is small in isolation, but anchoring can amplify it. Each labeled cat image functions as a query into existing structure, recruiting a neighborhood of related features and prototypes and making them available as supporting context for discrimination. In supervised-learning terms, the effective X is not just the few photos; it is the activated neighborhood of prior structure those photos elicit, while y is supplied by the label “cat.” Subsequent encounters then add diversity, refining the boundary and widening the region where anchoring succeeds.

Why pangolins are often harder. If the adult instead shows pangolins, mismatch d_r is often higher because diagnostic features (scales, posture, morphology) overlap less with the child’s existing animal neighborhoods. Higher d_r often calls for more or richer anchors (larger effective k), bridging descriptions that reduce mismatch (more labeled examples or richer context), or stronger local support ρ_d (more prior exposure to related structures) to keep $S > \theta$. This is why some categories are learned in a few shots while others require sustained experience.

Takeaway. The vignette illustrates our core claim in a simple setting: rapid learning is enabled by large repositories of reusable structure, but it requires semantic anchoring to push the system across a threshold into a stable regime of constrained generalization. Pattern repositories are not obstacles to learning. They are the substrate that makes few-shot learning possible once an anchoring signal, here labels plus exemplars, places the system in the right regime.

6. Multi-Agent Debate with Behavioral Modulation and Socratic Judging

Human reasoning scales through collaboration: diverse priors confront each other, surface hidden assumptions, and converge through critique. MACI makes this process explicit and controllable. Two mechanisms are central: (i) behavior modulation that regulates how strongly agents defend or revise hypotheses, and (ii) a judge that blocks ill-posed arguments from entering the shared state.

Beyond static stances. Many debate setups treat agents as fixed advocates. This can help, but it misses what makes debate productive: stance strength must adapt to evidence, the group must manage an explore-versus-consolidate tradeoff, and convergence must be prevented from collapsing onto fluent but ill-formed claims.

Behavior modulation: contentiousness and explore versus yield. Each agent i maintains a contentiousness parameter $\alpha_c^{(i)} \in [0, 1]$ governing how strongly it defends its current hypothesis. High $\alpha_c^{(i)}$ favors refutation and stress-testing; low $\alpha_c^{(i)}$ favors receptiveness and synthesis. MACI couples stance updates to (a) anchoring strength and (b) anchoring stability across rounds. When evidence binds strongly and consistently, agents yield and integrate; when anchoring is weak or unstable (for example, sensitive to paraphrase or retrieval variants), the group increases exploration by prioritizing counterexamples and alternative decompositions.

After each debate round t , agent i evaluates an incoming argument from agent j via semantic anchoring. When the argument binds strongly in the recipient (high $S_{j \rightarrow i}$), agent i reduces contentiousness:

$$\alpha_c^{(i)}(t+1) = \alpha_c^{(i)}(t) \cdot (1 - \beta \cdot S_{j \rightarrow i}), \quad (4)$$

where $\beta \in (0, 1)$ is a step-size parameter, and we use a normalized anchoring score $S_{j \rightarrow i} \in [0, 1]$ (e.g., a sigmoid of the raw score) so that $\alpha_c^{(i)}(t+1) \in [0, 1]$.

CRIT as a judge: Socratic filtering of ill-posed arguments. Debate alone is insufficient if agents can generate claims that are vague, internally inconsistent, or unsupported yet rhetorically fluent. MACI therefore introduces an explicit judge role grounded in CRIT (Critical Reading Inquisitive Template) that evaluates *reasonableness* independent of stance (Chang, 2023). The judge tests whether a claim is well-defined, whether assumptions are explicit, whether evidence supports the conclusion, and what would falsify it.

Operationally, CRIT gates the communication loop. Before a message is integrated into the shared state, it is scored for clarity, consistency, evidential grounding, and falsifiability. Low-scoring arguments are rejected or returned with targeted Socratic queries (e.g., “Which premise does the work?”, “What evidence would change your conclusion?”, “Are you changing definitions?”). This improves downstream anchoring by forcing arguments into forms that bind to shared constraints rather than just plausible.

Convergence dynamics with memory. With modulation and judging, debate becomes an explicit state-evolution process. Early rounds emphasize breadth and hypothesis coverage. Middle rounds emphasize anchoring-driven integration and pruning. Late rounds converge either to a synthesized position or to a structured residual disagreement with clearly identified fault lines. Persistent memory systems (e.g., SagaLLM (Chang et al., 2025a), and REALM-Bench (Geng & Chang, 2026)) track what was asserted, why it was asserted, and what later evidence contradicted it, enabling revision and auditability.

Debate as a controller for precision RAG and data acquisition. Debate can also route evidence. When agents disagree for principled reasons, the disagreement often indicates either (i) missing information required to anchor a decision, or (ii) an evidence set that is mutually inconsistent. MACI converts these signals into targeted requests: discriminating retrieval queries, missing contextual variables, or confirmatory measurements, then re-enters debate after storing the new evidence in memory.

In our EVINCE study, we show this concretely in clinical reasoning: two agents can surface likely failure points in an initial diagnosis, propose the queries and laboratory tests that most reduce ambiguity, and thereby prevent or correct misdiagnoses (Chang & Chang, 2025).¹

Why this matters. Behavior modulation prevents premature agreement that ignores long-tail counterexamples and prevents endless arguing that never consolidates when evidence is stable. CRIT prevents convergence to a coherent-sounding synthesis built on ill-posed premises. Together with semantic anchoring and memory, these mechanisms make multi-agent interaction reliable and revisable.

Beyond epistemic reliability, the same checks-and-balance structure supports ethical alignment when appropriate behavior depends on local norms and culture. In our DIKE-ERIS framework (Chang, 2025a), DIKE specifies general principles and constraints, while ERIS interprets and applies them in a context-sensitive manner.

Takeaway. MACI treats System-2 reasoning as an emergent property of coordinated System-1 agents: anchoring regulates what binds, CRIT filters ill-posed claims, memory preserves state, and checks-and-balance roles govern oversight and alignment.

7. Objections, Discriminating Tests, and the Path Forward

Sections 4–6 argued for a substrate-plus-coordination view: large pattern repositories enable rapid reconfiguration, while UCCT characterizes when external structure binds (anchoring) and MACI supplies mechanisms for regulated disagreement, judging, and stable state. We now translate common objections into testable hypotheses with discriminating experiments. The goal is practical: identify which failures are

¹Diagnostic error is a recognized public health problem. A National Academies report cites a conservative estimate that about 5% of U.S. outpatients experience diagnostic error each year, and post-mortem research suggests diagnostic errors contribute to roughly 10% of patient deaths (Balogh et al., 2015). Complementary work highlights that serious harms from misdiagnosis concentrate in a small set of conditions, suggesting targeted evidence-seeking interventions can be high leverage (Newman-Toker & Mark, 2023).

best explained by missing coordination layers (anchoring, oversight, memory, verification) versus limits that persist even when those layers are present.

7.1. From objections to testable hypotheses

Across objections, a recurring ambiguity is whether a failure reflects (i) insufficient or unstable anchoring (S below threshold, or near-threshold instability), (ii) inadequate coordination (no oversight, no state, no verification), or (iii) a deeper representational limitation. The hypotheses below are framed to separate these cases.

H1: “LLMs are just pattern matching.” *What is correct.* Isolated LLMs primarily perform large-scale pattern completion and can be brittle under distribution shift. *Our claim.* The relevant question is whether coordination can organize pattern capacity into reliable constraint-following and self-correction. *Discriminating tests.* Hold the base model fixed and compare: (i) base prompting, (ii) UCCT-informed anchoring control (adaptive exemplars or precision RAG based on estimated ρ_d and d_r), (iii) MACI (behavior modulation + CRIT judging + memory), and (iv) MACI + verification hooks. Measure accuracy, calibration, stability under paraphrase/retrieval variants, and failure *structure* (premise violations, inconsistency, hallucinated evidence). If improvements concentrate around predicted threshold shifts and are accompanied by increased stability and reduced calibrated error, the dominant limitation is organizational rather than categorical. *What would change our mind.* If anchoring control, memory, and oversight provide no systematic improvement beyond small prompt-level gains across diverse tasks, the “pattern-only” objection would be strengthened.

H2: “LLMs lack true understanding (symbol grounding).” *What is correct.* Text-only priors provide incomplete grounding; some errors reflect weak reference binding and missing sensorimotor constraints. *Our claim.* Much of the deficit is better understood as a solvable gap in grounding diversity and anchoring stability than as an impossibility; understanding is operationalized as reliable inference under constraints, not as a binary property. *Discriminating tests.* Evaluate identical target concepts across: (i) text-only, (ii) text+vision/audio, and (iii) tool-mediated or embodied interaction. Track (a) stability under paraphrase and counterfactual edits, (b) consistency under retrieval perturbations, and (c) the k required to reach stable performance. If grounding augmentation reduces effective mismatch d_r and produces predictable threshold shifts (lower k required, higher stability), this supports the coordination-plus-grounding account. *What would change our mind.* If multimodal and tool-grounded systems still fail systematically on reference binding and constraint adherence with no measurable an-

choring shift, then the grounding critique points to a deeper limitation.

H3: “LLMs cannot achieve compositional generalization.” *What is correct.* Current models can fail on systematic recombination, especially far from training support. *Our claim.* Many failures are consistent with low support (ρ_d) or high mismatch (d_r) for the required operators and intermediate states, plus insufficient budget (k), rather than an absence of compositional capacity. *Discriminating tests.* Use controlled suites where the rule is fixed but representational familiarity varies (e.g., numeral-base manipulations, operator rebindings, compositional instruction stacks). Test whether samples-to-success exhibits UCCT structure: higher support or lower mismatch should reduce required k , with sigmoid-like transitions near an effective threshold. Evaluate also stability under paraphrase and under alternative decompositions to distinguish “lucky” success from anchored success. *What would change our mind.* If compositional failures persist even when support is demonstrably high and anchoring is stable (robust across paraphrase and retrieval variants), then a missing compositional mechanism becomes more plausible.

H4: “The transformer architecture is fundamentally limited.” *What is correct.* A single-pass transformer without explicit state, verification, or recovery mechanisms is unreliable for long-horizon tasks. *Our claim.* The strongest evidence to date is more consistent with missing coordination and state layers than with a hard architectural barrier; the relevant system is not necessarily a pure end-to-end transformer. *Discriminating tests.* Keep the base model fixed and vary only the coordination stack: transactional memory, explicit plan state, recovery policies, verification hooks, and MACI oversight. If planning horizon, error recovery, and calibrated reliability improve primarily with coordination, then “fundamental architecture” claims weaken. *What would change our mind.* If failure modes remain invariant under substantial coordination scaffolding (no horizon extension, no reliability gains, no reduction in instability), then architectural replacement deserves more weight.

H5: “We need fundamentally different approaches.” *What is correct.* Hybridization, new training regimes, and new architectures may be needed for robustness, grounding, and safety. *Our claim.* Most alternatives still require a mechanism that learns broad priors from data; the practical choice is whether to discard that substrate or to augment it with anchoring, oversight, memory, grounding, and verification. *Discriminating tests.* Compare paradigms under matched task suites and explicit accounting of priors, constraints, and compute. If alternative systems eventually reintroduce large learned priors (in another form) and benefit from the same coordination stack, then “discard LLMs” is not justi-

fied. *What would change our mind.* If a non-LLM foundation achieves comparable breadth and few-shot adaptability without large learned priors and remains reliable without coordination layers, then our necessity claim would require revision.

7.2. What is actually missing (and how to build it)

Across the hypotheses, the recurring gap is the maturity of the coordination layer, not the existence of pattern repositories. The next step is dependable mechanisms for state, verification, and controlled hypothesis evolution:

1. **Long-horizon coordination.** Extended tasks require stable intermediate state and controlled recovery.
Build: transactional memory (SagaLLM-style), checkpointing, explicit plan state, and localized repair policies.
2. **Reliable inference and verification.** Avoidable logical or arithmetic errors should be detected and corrected.
Build: verifier hooks (symbolic or tool-based), self-consistency checks, and judge roles (CRIT-style) that reject ill-posed arguments before integration.
3. **Metacognitive monitoring.** Calibration and error awareness remain weak, especially near anchoring boundaries.
Build: uncertainty estimation tied to anchoring stability, multi-agent cross-checking, and stopping rules based on evidence closure rather than token or round limits.
4. **Efficient knowledge updating.** New facts and domain shifts should be incorporated without brittle prompting or full retraining.
Build: precision retrieval plus anchoring policies, dynamic memory stores, and targeted adaptation that raises ρ_d or reduces d_r where repeated failures are observed.
5. **Grounding diversity.** Text-only priors underconstrain physical and perceptual reasoning.
Build: multimodal and embodied learning and cross-modal anchoring that binds linguistic claims to perceptual/action constraints and tool feedback.

7.3. Key research directions

We highlight five research directions that strengthen the coordination layer while leveraging the existing substrate.

7.3.1. DIRECTION 1: PRINCIPLED SEMANTIC ANCHORING MECHANISMS

UCCT characterizes when anchoring succeeds via ρ_d , d_r , and k , but procedures for improving anchoring remain ad hoc. Needed are methods for (i) adaptive example selection and prompt construction based on estimated anchoring difficulty, (ii) bridge construction to reduce mismatch, and (iii) targeted repository augmentation to increase support in repeatedly failing regions.

7.3.2. DIRECTION 2: MULTI-AGENT COORDINATION WITH LEARNED POLICIES

Robust systems require role specialization, behavior modulation that balances explore versus yield, and judge roles (CRIT) that enforce well-posedness. A near-term goal is to learn debate-control policies that improve calibration and error detection, not only mean accuracy.

7.3.3. DIRECTION 3: PERSISTENT MEMORY DESIGNED FOR REASONING

Memory must preserve argument and plan lineage, support checkpointed rollback, and enable explicit revision (computational regret). The emphasis is transaction properties and constraint-aligned retrieval, not raw context length.

7.3.4. DIRECTION 4: MULTIMODAL AND EMBODIED GROUNDING

Broader grounding reduces mismatch and makes anchors more stable across contexts. The objective is cross-modal anchoring: linguistic claims should be constrained by perception, action, and tool feedback, improving both stability and falsifiability.

7.3.5. DIRECTION 5: NEUROSymbOLIC INTEGRATION AS VERIFICATION, NOT REPLACEMENT

Symbolic or tool-based components are most valuable as verifiers and constraint enforcers: pattern priors propose candidates; independent checkers validate. This mirrors robust engineering practice: proposal plus validation.

7.4. What success looks like

A mature system is layered and ablatabile: pretrained pattern repositories as substrate; anchoring to bind external constraints; multi-agent coordination to regulate hypothesis evolution; persistent memory for long-horizon state and recovery; grounding via multimodal or embodied signals; and verification for correctness and safety. Each layer should expose measurable diagnostics (stability, calibration, threshold shifts, recovery success) so progress is driven by discriminating experiments rather than qualitative debate.

8. Conclusion

The AI community is debating whether large language models are a foundation to build on or a dead end to abandon. This paper argues for a third position: LLMs are not sufficient for AGI, but they are a necessary substrate. The practical question is not whether to discard pattern models, but how to organize them into reliable, constraint-following, long-horizon reasoning.

Our contribution is a constructive account of that organi-

zation. UCCT formalizes *semantic anchoring* as a thresholded transition: external structure binds to internal pattern repositories only when anchoring strength crosses a task-dependent boundary, yielding phase-transition-like shifts from prior-driven association to anchored control. MACI then turns anchoring into a control signal in a coordination stack: multi-agent interaction regulated by behavior modulation (including explore versus yield dynamics), Socratic judging via CRIT to filter ill-posed arguments, and checks-and-balance roles (Dike–Eris) for context-sensitive alignment. Crucially, long-horizon competence requires *persistent, transactional state*: SagaLLM-style memory preserves commitments, intermediate results, and invariants across steps, enabling revision and computational regret.

We also recast common objections as competing hypotheses with discriminating tests. The strongest version of the “doomed” critique is not that current systems fail, but that they fail for reasons that cannot be repaired. Our position is that many salient failures are consistent with missing anchoring and missing coordination. This yields a concrete research agenda: strengthen anchoring mechanisms, develop robust multi-agent control policies, build memory designed for reasoning and recovery, broaden grounding, and integrate verification as a constraint layer.

The closing message is simple. Pattern repositories are not an obstacle to intelligence. They are the substrate from which higher-order control can emerge when properly coordinated. Large language models are therefore not doomed. The work ahead is to make the coordination layer principled, testable, and reliable.

References

- Baddeley, A., Eysenck, M. W., and Anderson, M. C. *Memory*. Psychology Press, 2007.
- Balogh, E. P., Miller, B. T., and Ball, J. R. (eds.). *Improving Diagnosis in Health Care*. The National Academies Press, Washington, DC, 2015. doi: 10.17226/21794.
- Chang, E. Y. CRIT: Prompting Large Language Models With the Socratic Method. *IEEE 13th Computing and Communication Workshop and Conference*, March 2023.
- Chang, E. Y. A Checks-and-Balances Framework for Context-Aware Ethical AI Alignment. In *ICML*, July 2025a.
- Chang, E. Y. *Multi-Agent Collaborative Intelligence: Foundations and Architectures for Artificial General Intelligence*. ACM Books, November 2025b. Early release March 2024.
- Chang, E. Y. and Chang, E. Y. Multi-agent collaborative intelligence: Dual-dial control for reliable llm reasoning, 2025. URL <https://arxiv.org/abs/2510.04488>.
- Chang, E. Y. et al. Sagallm: Persistent memory for long-horizon planning in large language models. *Proceedings of the VLDB Endowment*, 2025a.
- Chang, E. Y. et al. Semantic anchoring in llms: Thresholds, transfer, and geometric correlates. arXiv:2506.02139, 2025b.
- Dehaene, S. *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts*. Viking, 2014.
- Ericsson, K. A., Charness, N., Feltovich, P. J., and Hoffman, R. R. (eds.). *The Cambridge Handbook of Expertise and Expert Performance*. Cambridge University Press, 2006.
- Fitts, P. M. and Posner, M. I. *Human Performance*. Brooks/Cole Publishing Company, 1964.
- Gazzaniga, M. S., Ivry, R. B., and Mangun, G. R. *Cognitive Neuroscience: The Biology of the Mind*. W. W. Norton & Company, 4 edition, 2014.
- Geng, L. and Chang, E. Y. Realm-bench: A benchmark for evaluating multi-agent systems on real-world, dynamic planning and scheduling tasks. *ACM KDD*, 2026. URL <https://arxiv.org/abs/2502.18836>.
- Huang, X., Liu, W., Chen, X., Wang, X., Wang, H., Lian, D., Wang, Y., Tang, R., and Chen, E. Planning with large language models: A survey, 2024. arXiv:2402.02716.
- Hubel, D. H. and Wiesel, T. N. Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex. *Journal of Physiology*, 160(1):106–154, 1962.
- Kahneman, D. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.
- Kandel, E. R., Schwartz, J. H., Jessell, T. M., Siegelbaum, S. A., and Hudspeth, A. J. *Principles of Neural Science*. McGraw-Hill Education, 5 edition, 2013.
- Kanwisher, N., McDermott, J., and Chun, M. M. The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, 17(11):4302–4311, 1997.
- LeCun, Y. A path towards autonomous machine intelligence. Technical report, Meta AI Research, 2022.
- LeDoux, J. E. *The Emotional Brain: The Mysterious Underpinnings of Emotional Life*. Simon & Schuster, 1996.
- Liu, M., Diao, S., Lu, X., Hu, J., Dong, X., Choi, Y., Kautz, J., and Dong, Y. Prorl: Prolonged reinforcement learning expands reasoning boundaries in large language models. *NeurIPS*, 2025.

-
- Newman-Toker, D. J. and Mark, R. G. Misdiagnosis in the united states: An economic and public health tragedy. *Health Economics, Policy and Law*, 18(4):712–721, 2023.
- Posner, M. I. and Snyder, C. R. Attention and cognitive control. In Solso, R. L. (ed.), *Information Processing and Cognition: The Loyola Symposium*, pp. 55–85. Lawrence Erlbaum Associates, 1980.
- Shiffrin, R. M. and Schneider, W. Controlled and automatic human information processing: II. perceptual learning, automatic attending and a general theory. *Psychological Review*, 84(2):127–190, 1977.
- SIMA Team, Abi Raad, M., Ahuja, A., Barros, C., Besse, F., Bolt, A., Bolton, A., Brownfield, B., Buttimore, G., Cant, M., Chakera, S., Chan, S. C. Y., Clune, J., et al. Scaling instructable agents across many simulated worlds, 2024. arXiv:2404.10179.
- Squire, L. R. and Kandel, E. R. *Memory: From Mind to Molecules*. Roberts and Company Publishers, 2 edition, 2013.
- Sutskever, I. and Patel, D. Ilya sutskever (interview), November 2025. URL <https://www.dwarkeshpatel.com/p/ilya-sutskever>.
- Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., and Wen, J.-R. A survey on large language model based autonomous agents, 2023. arXiv:2308.11432.