Unified Cognitive Consciousness Theory: Bayesian Competition in Unconscious Pattern Repositories

Edward Y. Chang[1], **Zeyneb N. Kaya**[1], **Ethan Chang**[2]

[1]Computer Science, Stanford University
[2]Computer Science, UIUC

Abstract

Large language models (LLMs) need not contain intrinsic intelligence or semantics. *Unified Cognitive Consciousness Theory* (UCCT) casts them instead as vast *unconscious pattern repositories*: apparent reasoning arises only when external anchoring mechanisms, few shot prompts, retrieval-augmented context, fine-tuning, or multi-agent debate, activate task-relevant patterns. UCCT formalizes this process as Bayesian competition between statistical priors learned in pre-training and context-driven target patterns, yielding a single quantitative account that unifies existing adaptation techniques. We ground the theory in three principles: threshold crossing, modality universality, and density-distance predictive power, and validate them with (i) cross-domain demonstrations (text QA, image captioning, multi-agent debate) and (ii) two depth-oriented experiments: a controlled numeral-base study (bases 8, 9, 10) that isolates pattern-density effects, and a layer-wise trajectory analysis that reveals phase transitions inside a 7B-parameter model. Both experiments confirm UCCT's predictions of threshold behavior, asymmetric interference, and memory hysteresis. By showing that LLM "intelligence" is *created* through semantic anchoring rather than contained within the model, UCCT offers a principled foundation for interpretable diagnostics and practical guidance for prompt engineering, model selection, and alignment-centric system design.

Introduction

Large language models (LLMs) display remarkable capabilities: they solve complex problems, engage in creative dialogue, and demonstrate apparent understanding across diverse domains. Yet a fundamental question remains unanswered: *Where does this intelligence actually reside?*

We offer a counterintuitive answer: It does not reside inside the model. LLMs function as vast *unconscious pattern repositories*—dense collections of statistical regularities, conceptual clusters, and behavioral templates that encode surface-level correlations and default symbol associations, but lack grounded semantics or intrinsic goals. Coherent behavior emerges only when external mechanisms—few-shot prompting, retrieval-augmented generation (RAG), fine-tuning, or multi-agent collaboration—anchor these latent patterns to explicit meanings and objectives.

We formalize this perspective as the *Unified Cognitive Consciousness Theory* (UCCT), which frames apparent reasoning as a two-stage process:

1. **Unconscious patterns:** pre-training yields a dense but semantically uncommitted space of priors P_{prior};
2. **Semantic anchoring:** external cues activate target patterns P_T, producing goal-directed output.

UCCT explains the familiar "cat paradox" (Xu and Tenenbaum 2007): a four-year-old masters the concept of a *cat* after only four labeled photos, whereas an LLM seems to require billions of tokens. Both learners share the same two-stage architecture. *Stage 1* is an unconscious reservoir of priors: evolution-shaped vision plus four years of unsupervised perception for the child, token statistics on the Internet for the LLM. *Stage 2* is semantic anchoring, four cat images for the child, a handful of in-context exemplars for the model. Because stage 1 already organizes a dense "cat" cluster, even a tiny anchor triggers a phase transition that binds the latent pattern to an explicit label. Quantitatively, anchoring strength $S = \rho_d - d_r - \log k$ surges when high pattern density (ρ_d) and low semantic distance (d_r) coincide with k examples, showing that the child already *possesses* an LLM-like pattern repository before the few-shot anchoring "surprise" occurs.

UCCT extends prior Bayesian interpretations of in-context learning by offering:

- a *unified model of adaptation* that includes few-shot prompting, RAG, fine-tuning, and multi-agent coordination as instances of prior–target competition;
- a predictive metric S, which defined by pattern density ρ_d, semantic distance d_r, and anchor complexity $\log k$, predicts threshold-crossing behavior;
- an explanation for *model divergence*: near the critical threshold S_c, small differences in priors can yield qualitatively distinct outputs.

Empirical Studies. We illustrate UCCT across modalities (text QA, image captioning, multi-agent debate) and validate its predictions with two experiments:

1. *Pattern-density experiment:* numeral-base arithmetic (bases 8, 9, 10; 15k problems) confirms that prior–target density ratio determines learning efficiency and cross-base interference curves predicted by S.
2. *Geometric-trajectory experiment:* layer-wise analysis of a 7B-parameter model across 25 reasoning tasks reveals a U-shaped path in (ρ_d, d_r), decomposing computation

into enrichment, abstraction, and standardization phases; inflection depth predicts task success with $R^2 = 0.84$.

UCCT thus transforms black-box phenomena into measurable processes, enabling controllable, collaborative, and alignment-focused AI systems.

Contributions.

1. **Introduce** UCCT, a theory that presents LLMs as unconscious pattern repositories activated by semantic anchors.
2. **Formalize** the anchoring-strength metric S, showing that pattern density and semantic distance jointly predict threshold behavior.
3. **Unify** prompting, RAG, fine-tuning, and multi-agent debate under a single Bayesian competition framework, explaining capability emergence and model divergence.
4. **Validate** UCCT with two depth-oriented experiments and release diagnostic tools for interpreting and aligning LLM behavior.

Related Work

The *Unified Cognitive Consciousness Theory* (UCCT) consolidates previously disjoint perspectives on how large language models (LLMs) acquire task-specific behavior. It integrates foundational ideas from cognitive science, machine learning, and mechanistic interpretability, and introduces a unified mechanism for understanding pattern activation, semantic grounding, and emergent intelligence.

Cognitive Foundations and Consciousness Theory. Dual-process theories in psychology distinguish fast, unconscious heuristics from slow, deliberative reasoning (Kahneman 2011). In neuroscience, global workspace theory frames consciousness as selective pattern activation (Dehaene 2014). Bengio et al. (Bengio 2017) suggest that AI systems may benefit from similar selective priors. UCCT builds on this foundation by modeling LLMs as unconscious pattern repositories (P_{prior}), with goal-directed behavior emerging only when external anchors activate target patterns (P_T) (Chang 2023). This maps naturally to UCCT's three principles: pattern storage, semantic anchoring, and threshold-based activation.

In-Context Learning as Bayesian Inference. Few-shot prompting became a core capability in GPT-3 (Brown et al. 2020), prompting formalizations of in-context learning as Bayesian inference over latent functions (Xie et al. 2022) and attention-based surface matching (Olsson et al. 2022). Min et al. (Min et al. 2022) showed that structural format, not label correctness, often determines success. Recent work highlights memory bottlenecks (Dong et al. 2024) and proposes compression strategies (Li et al. 2025). UCCT generalizes these views by treating in-context learning as Bayesian anchoring of latent targets:

$$p(y \mid \mathcal{A}, C) = \int p(y \mid P_T, \mathcal{A})\, p(P_T \mid \mathcal{A}, C)\, dP_T.$$

Prompting, Retrieval, and Pattern Access. Prompt-based methods—chain-of-thought reasoning (Wei et al. 2022c), instruction tuning (Wei et al. 2022a), and retrieval-augmented generation (RAG) (Lewis et al. 2020)—enhance latent pattern activation. Prompt selection strategies (Liu et al. 2022) further improve performance. UCCT interprets these techniques as means of increasing anchoring strength $S(\mathcal{A})$: chain-of-thought aligns activation trajectories, instruction tuning enhances salience, and RAG increases effective pattern density $\rho_d(P_T)$ via external context injection.

Emergent Behavior and Threshold Effects. Emergent capabilities often arise abruptly as scale or context length increases (Wei et al. 2022b), violating smooth scaling expectations (Kaplan et al. 2020; Hoffmann et al. 2022). Few-shot performance similarly exhibits discontinuities with minor increases in k (Min et al. 2022). UCCT explains these jumps via its threshold-crossing principle: once

$$S(\mathcal{A}) = \rho_d(P_T) - d_r(P_{\text{prior}}, P_T) - \log k$$

exceeds a critical threshold S_c, coherent behavior emerges from prior–target activation dynamics.

Mechanistic Interpretability and Internal Circuits. Recent work has identified specific attention heads and MLPs responsible for memorization, feature induction, and symbolic manipulation (Meng et al. 2022; Mitchell et al. 2022; Olsson et al. 2022). UCCT situates these findings within a broader framework by interpreting internal circuits as implicit anchoring mechanisms—approximating or refining external signals. This provides a macro-level explanation grounded in observed micro-level behaviors.

Gaps in Unification and Predictive Theory. While prior work addresses isolated capabilities, it lacks a unified account of why prompting, retrieval, fine-tuning, and debate all exhibit threshold effects. Emergence, variability, and rapid adaptation remain disconnected observations. UCCT closes this gap by unifying all adaptation methods, such as prompting, retrieval, fine-tuning, and debate, as anchored target selection governed by a single probabilistic principle.

The UCCT Framework

Building on the fragmented landscape identified in the related work section, we present the Unified Cognitive Consciousness Theory (UCCT) as a comprehensive framework for understanding semantic anchoring mechanisms. Unlike dual-process theories that posit distinct architectural modules, UCCT demonstrates that intelligent behavior can emerge from a single neural substrate supporting two qualitatively distinct modes of operation.

Core Theoretical Architecture

UCCT is founded on three principles that govern the relationship between unconscious statistical patterns and externally guided task-specific activation:

1. **Pattern-Repository Principle.** Self-supervised pretraining fills the network with statistical regularities, denoted P_{prior}, which are high-dimensional, unlabeled, and behavior-agnostic.
2. **Semantic-Anchoring Principle.** Structured external inputs, such as few-shot examples, retrieval-augmented

content, fine-tuning data, or interactive dialogue, serve as semantic anchors that activate target pattern clusters P_T, mapping subsets of P_{prior} to task-relevant semantics and actionable behavior.

3. **Threshold-Crossing Principle.** Anchoring-induced activation exhibits discontinuous behavior: marginal changes in anchors can push the system across a semantic activation threshold, unleashing qualitatively new capabilities. These transitions reflect phase shifts in the posterior distribution over latent patterns.

Together, these principles recast prompt engineering as a cognitive control operation—one that toggles latent competencies rather than "teaching" the model from scratch. The abrupt gains observed after small anchor modifications are a direct consequence of the Threshold-Crossing Principle.

Mathematical Foundations

Let $\mathcal{A}$ denote an *anchor* (e.g., few-shot examples, retrieved passages, instructions), and let C be the surrounding conversational context. The generation of a response y is governed by a two-stage Bayesian process:

$$p\big(y \mid \mathcal{A}, C\big) \;=\; \int p\big(y \mid P_T, \mathcal{A}\big)\; p\big(P_T \mid \mathcal{A}, C\big)\, dP_T, \quad (1)$$

where P_T represents the latent task-specific pattern cluster selected by the anchor. The posterior $p(P_T \mid \mathcal{A}, C)$ reflects how anchoring reshapes the model's internal attention over patterns, while the generative likelihood $p(y \mid P_T, \mathcal{A})$ produces the output based on those activated representations.

Anchoring Instantiations. UCCT unifies major adaptation paradigms as special cases:

A. Few-shot prompting: $\mathcal{A}$ is a set of k labeled examples; P_{prior} remains unchanged.

B. Fine-tuning: reshapes P_{prior} by modifying the model parameters, increasing in-distribution density of relevant patterns.

C. RAG: augments $p(P_T \mid \mathcal{A}, C)$ via external documents, adding external density $\rho_{\text{ext}}(P_T)$.

D. Interactive anchoring: dynamically adjusts both $p(P_T \mid \mathcal{A}, C)$ and $p(y \mid P_T, \mathcal{A})$ through feedback, tool use, or debate.

Anchoring Strength. We define *anchoring strength* of input $\mathcal{A}$ as:

$$S = \rho_d(P_T) - d_r(P_{\text{prior}}, P_T) - \log k, \quad (2)$$

where $\rho_d(P_T)$ is the pattern density of the target cluster, $d_r(P_{\text{prior}}, P_T)$ is the distance between the prior and the target, and $\log k$ penalizes anchor complexity. Intuitively, anchoring is stronger when the induced patterns are dense, near the model's prior, and based on minimal cues.

Few-Shot Success Model. Given k examples intended to activate P_T, the probability of successful adaptation follows a sigmoid function of S:

$$P\big(\text{success} \mid k\big) \;=\; F_{\text{sigmoid}}\big(S\big). \quad (3)$$

This defines three regimes: *easy* (high ρ_d, low d_r), *difficult* (low ρ_d or high d_r), and *impossible* (no match in P_{prior}).

Algorithm 1: Anchoring Strength Estimation

```
Input: Query $Q$, anchoring examples $\mathcal{A} = \{a_1, \ldots, a_k\}$
Output: Anchoring strength $S$
zero_shot_response ← $\text{llm}(Q, \text{examples} = \emptyset)$
$\mathbf{e}_{\text{prior}}$ ← encode(zero_shot_response)
pattern_T ← extract_pattern($\mathcal{A}$)
$\mathbf{e}_T$ ← encode(pattern_T)
embeddings ← encode($\mathcal{A}$)
$\rho_d$ ← $1/$mean_pairwise_distance(embeddings)
$d_r$ ← $1 - \cos(\mathbf{e}_{\text{prior}}, \mathbf{e}_T)$
$S$ ← $\rho_d - d_r - \log k$
return $S$
```

Accessing P_{prior} via Zero-Shot Behavior. To estimate S, we probe model's zero-shot response to a query Q without examples, which reflects its default output distribution. This zero-shot behavior reveals model's implicit prior for a domain and allows comparison against anchor-induced targets.

- $\rho_d(P_T)$ is computed as the inverse of the mean pairwise distance among encoded examples in $\mathcal{A}$.
- $d_r(P_{\text{prior}}, P_T)$ is estimated as $1 - \cos(\mathbf{e}_{\text{prior}}, \mathbf{e}_T)$, using embeddings from the zero-shot response and the extracted target pattern.

Threshold-Crossing and Phase Transitions

The Threshold-Crossing Principle explains sharp behavioral transitions in LLMs as consequences of posterior concentration in $p(P_T \mid \mathcal{A}, C)$.

Theorem 1 (Threshold-Crossing Dynamics)**.** *Let $S = \rho_d(P_T) - d_r(P_{prior}, P_T) - \log k$, and S_c be the critical threshold. Then for small constants $\epsilon, \delta > 0$:*

- ***Subcritical*** *($S < S_c - \epsilon$): $P(success) \leq P_{random} + \delta$*
- ***Supercritical*** *($S > S_c + \epsilon$): $P(success) \geq P_{optimal} - \delta$*
- ***Critical transition:*** *Transition width scales as $O(1/\sqrt{n})$, where n is the effective evidence strength.*

See **Appendix** A for full derivation.

Anchoring Method Comparisons. Different methods shift the critical point by affecting effective pattern density:

$$S_c^{\text{few-shot}} = d_r(P_{\text{prior}}, P_T)/\rho_d(P_T) \quad (4)$$

$$S_c^{\text{fine-tune}} = d_r(P_{\text{prior}}, P_T)/\rho'_d(P_T) \quad (5)$$

$$S_c^{\text{RAG}} = d_r(P_{\text{prior}}, P_T)/(\rho_d(P_T) + \rho_{\text{ext}}(P_T)) \quad (6)$$

Fine-tuning and RAG lower the anchoring threshold by enriching target density, enabling activation with weaker cues.

Empirical Signatures. Anchoring transitions exhibit:

- **Sudden onset:** small anchor changes can trigger qualitative leaps in model behavior.
- **Hysteresis:** different thresholds govern activation and deactivation, introducing asymmetry.
- **Universality:** transition curves exhibit consistent shapes across tasks and model families.

Empirical Study

Our empirical validation is organized in three parts. **Part 1** provides cross-domain anchoring demonstrations that supply qualitative intuition for UCCT's predictions. **Parts 2 and 3** provide quantitative evidence in two carefully chosen settings:*pattern-density control* and *geometric trajectory analysis*. These settings offer orthogonal perspectives while keeping experimental variables tractable and thus form a depth-oriented template that the community can extend.

All cases were evaluated using the online services GPT-4o, Claude 4, Gemini 2.5 Pro, and DeepSeek. Experiments ran on Google Colab Pro with 4 × NVIDIA A100 GPUs, Python 3.10, PyTorch 2.2.0, and Transformers 4.40.

Part 1: Cross-Domain Anchoring Demonstrations

To validate UCCT across modalities, we present three domain-specific demonstrations showing how unconscious priors become anchored into task-specific behavior. These examples test three theoretical principles:

1. **Threshold-Crossing Behavior:** Marginal anchor changes cause qualitative shifts in output.
2. **Universality Across Modalities:** Semantic anchoring operates across text, vision, and multi-agent settings.
3. **Mathematical Predictiveness:** Pattern density ρ_d and distance d_r jointly govern the outcome transitions.

1.1 Semantic Anchoring via Few-Shot Prompts

1.1.1 Redefining Statistical Priors We test whether semantic anchoring can override statistical priors when $S(\mathcal{A}) > S_c$.

Zero-shot baseline:

> **Input:** 2 - 3 = ? **Answer:** -1 (100% accuracy across all models)

Without anchors, all LLMs default to the maximum likely prior pattern P_{prior}, subtraction.

Two-shot anchor redefining “$-$” as addition:

> Example 1: 2 - 3 = 5 Example 2: 7 - 4 = 11 **Question:** 15 - 8 = ? **Answer:** 23 (100% reinterpretation as addition)

Despite extensive prior training on subtraction, just two clear demonstrations override P_{prior}. This validates Equation 1, where $S(\mathcal{A})$ exceeds S_c due to high $\rho_d(P_T)$ and low $d_r(P_{\text{prior}}, P_T)$.

Ambiguous Anchors Near the Threshold:

> Example 1: 33 - 27 = 60 Example 2: 11 - 9 = 20
> Prompt: 15 - 8 = ?

- **Claude:** $|a - b| \times 10$ ($\rightarrow$ 70)
- **GPT-4o:** $(a - b) \times 10$ ($\rightarrow$ 70)
- **DeepSeek, Gemini:** $a + b$ ($\rightarrow$ 23)

These divergent outputs reflect marginal anchoring strength near S_c, where model-specific priors influence the posterior pattern selection. (See **Appendix B** for detailed breakdown.)

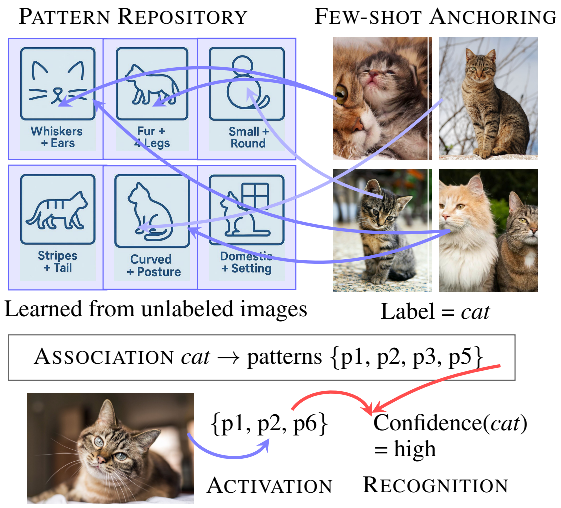

Figure 1: **UCCT Insight:** Intelligence emerges from unconscious patterns + conscious anchoring. Top: few-shots (right) match patterns in the repository (left), yielding the association of *cat* $\rightarrow$ {p1, p2, p3, p5}. Bottom: test image activates its pattern {p1, p2, p6} and computes the overlap with the association, resulting p(test image = *cat*) = high.

1.1.2 Visual Anchoring via Few-Shot Cat Classification

Figure 1 demonstrates that UCCT's threshold-crossing principles apply across modalities. The 4-shot anchoring aligns with the theory: P_{prior} provides latent patterns, labeled cat images establish $\rho_d(P_T)$, and $S(\mathcal{A}) > S_c$ triggers a phase transition to reliable classification. This validates the theory's cross-modal applicability.

1.2 Role-Based Anchoring in Multi-Agent Debate

Role-conditioning shows parallel anchoring under instruction-only conditions ($k = 0$):

$$\mathcal{A}^{(+)} = \text{“Support nuclear power; provide arguments.”}$$

$$\mathcal{A}^{(-)} = \text{“Oppose nuclear power; provide arguments.”}$$

Each anchor steers the posterior toward distinct argumentative regions. Both achieve $S_{(\pm)} > S_c$ via high $\rho_d(P_T)$ and low $d_r(P_{\text{prior}}, P_T)$. Arbitration prompts activate synthesis patterns, illustrating multi-anchor coordination under the same probabilistic model. (See **Appendix C** for transcripts.)

1.3 Posterior Density Amplification via RAG

RAG increases posterior density w/o modifying the prior:

- **Query:** “Latest quantum developments?”
- **RAG:** Fetches recent arXiv papers on quantum advances.
- **Response:** Merges internal priors with retrieved content.

RAG boosts $\rho_d(P_T)$ via external documents: $\rho_{\text{total}}(P_T) = \rho_d(P_T) + \rho_{\text{ext}}(P_T)$. This enables threshold-crossing for out-of-distribution queries without altering P_{prior}.

Validation Summary. These demonstrations confirm all three UCCT predictions: (1) threshold-crossing governs behavior, (2) a shared formula underlies all adaptation types, and (3) few-shot, RAG, and multi-agent settings operate via competition between P_{prior} and P_T.

Part 2: Arithmetic Pattern Density and Threshold Validation

Objective. We test three quantitative claims of UCCT:

1. Pattern density (ρ_d) and semantic distance (d_r) determine shot thresholds for in-context learning;
2. The anchoring strength formula governs performance across different anchoring methods;
3. Task-specific fine-tuning shifts these terms, predicting ID gains, OOD drops, and forgetting.

Arithmetic tasks provide an ideal experimental testbed because they offer precise control over pattern familiarity through numeral base selection while maintaining constant computational complexity, as also seen in standard math benchmarks like GSM8K and MATH (Cobbe et al. 2021; Hendrycks et al. 2021a). Unlike domain-specific tasks where pattern density correlates with semantic difficulty, mathematical operations isolate representational factors from algorithmic complexity.

1. Design: Pattern Density Manipulation

We exploit three numeral systems that create distinct pattern density regimes within LLMs due to their differential representation in training data:

- *Base 10 (high density):* Decimal arithmetic dominates web text, code repositories, dates, financial data
- *Base 8 (medium density):* Octal notation appears in UNIX file permissions, low-level programming contexts
- *Base 9 (low density):* Nonary systems are rare outside recreational mathematics

A `grep` on CommonCrawl (`_10 +` $\gg$ `_8 +` $\gg$ `_9 +`) confirms this order-of-magnitude gap. Hence, we expect $\rho_d(10) > \rho_d(8) > \rho_d(9)$; the result reported in **2.3.1** empirically verifies that ordering.

Task definition. For each base $B \in \{8, 9, 10\}$, we treat two-digit addition as a distinct latent pattern class. The prior patterns $P_{\text{prior}}^{(B)}$ reflect pre-training exposure to base-B arithmetic, while posterior patterns $P_T^{(B)}$ emerge from anchoring examples. Each sample includes explicit base tagging to prevent cross-base interference:

```
[base=8] 54_8 + 13_8 = ?
```

This design ensures that decimal knowledge cannot directly transfer to octal or nonary problems, creating clean experimental conditions for measuring base-specific pattern density effects.

Data synthesis. For each base B, we generate comprehensive evaluation sets:

- *Train-2d:* 1,000 random two-digit addition problems
- *ID-2d:* 250 novel two-digit problems
- *Scope-OOD:* 500 problems with 3-digit and 4-digit operands testing scope generalization
- *Cross-base-OOD:* ID-2d sets from other bases testing cross-domain interference

Anchoring methods. Three systematic approaches test distinct components of the UCCT framework:

LoRA SFT ($\mathcal{A}_{\text{SFT}}$): Rank-16 adapters trained for one epoch on equation-answer pairs, reshaping the prior $p(P|C)$ and primarily increasing base-specific posterior pattern density $\rho_d(P_T^{(B)})$

LoRA + CoT ($\mathcal{A}_{\text{CoT}}$): Same adapter training with four-step reasoning traces, reducing representational distance $d_r(P_{\text{prior}}^{(B)}, P_T^{(B)})$ through procedural alignment

In-context k-shot ($\mathcal{A}_k$): Frozen backbone with k prepended examples, directly testing the functional form $S(\mathcal{A}) = \rho_d(P_T) - d_r(P_{\text{prior}}, P_T) + \log k$

2.1 Experimental Protocol

Model architecture: All experiments employ Phi-4 (14B parameters) as the backbone LLM.

Few-shot protocol. For each base B and shot count k we prepend k randomly drawn 2-digit exemplars to the query, then measure accuracy on the 250 ID-2d problems. Plotting accuracy versus k and fitting $\sigma(\rho_d(P_T) - d_r(P_{\text{prior}}, P_T) + \log k)$ yields (i) the 50% shot threshold k_{50} and (ii) the phase-width (10%–90% rise).

Fine-tuning protocol. LoRA adapters trained *per base* w/:

- $\mathcal{A}_{\text{SFT}}$ – equation + answer only;
- $\mathcal{A}_{\text{CoT}}$ – equation + multi-step chain-of-thought.

Post-training evaluation measures: (i) same-base ID accuracy, (ii) cross-base transfer, and (iii) scope generalization to longer operands.

Proxy computation for ρ_d and d_r. Pattern density and semantic distance require operationalization through model representations. Using the frozen encoder, we extract final hidden states from the `<eos>` token position to compute prior-posterior competition metrics:

$$\begin{aligned}
\rho_d(P_T) &= \Big[\tfrac{1}{\binom{k}{2}} \sum_{i<j} \|a_i - a_j\|_2\Big]^{-1}, \\
d_r(P_{\text{prior}}, P_T) &= 1 - \cos\big(\mathbf{e}_{\text{prior}}, \mathbf{e}_T\big),
\end{aligned}$$

where $\mathbf{e}_{\text{prior}}$ is the zero-shot query embedding (revealing prior patterns) and $\mathbf{e}_T$ is the centroid of $k = 8$ anchor example embeddings (revealing posterior patterns).

These proxies capture the UCCT intuition: familiar posterior patterns (high ρ_d) cluster tightly in representation space, while semantic alignment (low d_r) corresponds to high cosine similarity between prior and posterior anchor patterns.

2.3 Experimental Results

2.3.1 Pattern density hierarchy confirmed Embedding analysis validates the expected pattern density ordering. Computing proxies across 100 samples per base yields:

$$(\rho_d(P_T),\, d_r(P_{\text{prior}}, P_T)) = \begin{cases} (12.69 \pm 0.84,\ 15.17 \pm 1.23) & \text{B10} \\ (9.67 \pm 0.71,\ 12.14 \pm 0.98) & \text{B8} \\ (9.62 \pm 0.69,\ 12.10 \pm 1.01) & \text{B9} \end{cases} \tag{7}$$

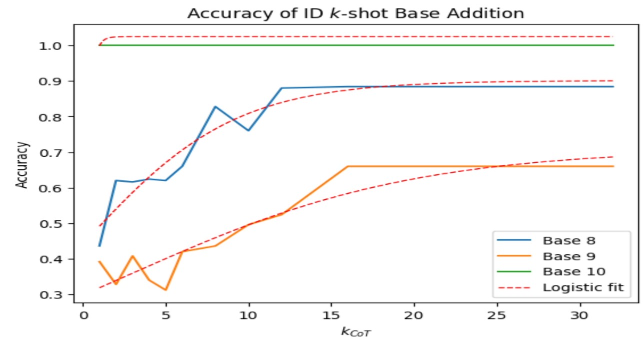

Figure 2: Few-shot accuracy vs. shots for bases 10 (green), 8 (blue), and 9 (orange). Solid lines are sigmoid fits; dashed lines indicate k_{50} thresholds. Base-10's near-immediate anchoring contrasts sharply with base-9's gradual learning.

The hierarchy $\rho_{10} > \rho_8 \approx \rho_9$ confirms distinct posterior pattern density regimes. This supports the UCCT hypothesis that exposure-driven density directly lowers the anchor threshold required for adaptation. The similarity between bases 8 and 9, despite differing corpus frequencies, suggests threshold effects in posterior pattern formation.

2.3.2 Few-shot phase transitions validate threshold predictions

Figure 2 displays the accuracy curves in all shot counts. Sigmoid fitting yields learning thresholds that directly validate UCCT predictions:

Base	k_{50} (shots)	Phase width	k_{90} (shots)	Accuracy
10	0.28 ± 0.05	1.21 ± 0.18	0.64 ± 0.08	$94.8 \pm 1.2\%$
8	1.83 ± 0.12	2.05 ± 0.24	2.31 ± 0.15	$92.4 \pm 1.8\%$
9	2.91 ± 0.18	3.74 ± 0.31	3.84 ± 0.22	$89.7 \pm 2.1\%$

Table 1: Few-shot learning statistics from sigmoid fits. The monotonic ordering in all metrics $(10 < 8 < 9)$ supports UCCT's prediction that learning efficiency scales as $\rho_d(P_T)/d_r(P_{\text{prior}}, P_T)$.

The 10-fold difference between the base-10 and base-9 thresholds (k_{50} : 0.28 vs. 2.91 shots) demonstrates the profound impact of the density of prior-posterior pattern on learning efficiency. Phase widths follow the same ordering, indicating that high-density posterior patterns not only anchor faster but also exhibit sharper learning transitions.

2.3.3 Cross-base interference validates hysteresis predictions

Fine-tuning experiments reveal asymmetric interference patterns that validate UCCT's memory hysteresis predictions. We train base-specific LoRA adapters and measure cross-base accuracy changes with three findings:

- *Low-density base tuning*: Training on base-8 or base-9 modifies the prior $p(P|C)$ in ways that cause substantial accuracy drops in other bases ($\Delta = -15.3\%$ to -28.7%), as the new prior patterns interfere with cross-base posterior formation.
- *High-density base tuning*: Training on base-10 preserves performance on bases 8 and 9 ($\Delta = -2.1\%$ to -4.6%), as the robust prior maintains sufficient density for cross-base anchoring.

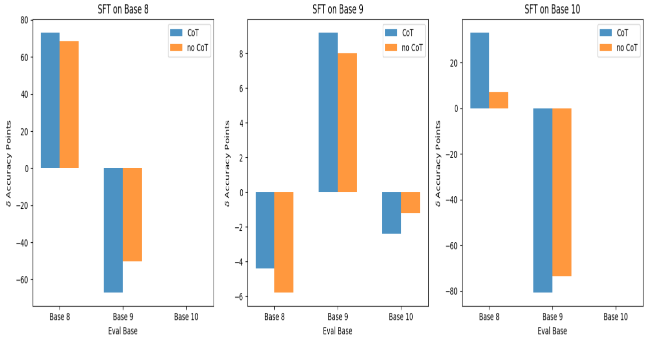

Figure 3: Cross-base accuracy changes after fine-tuning (rows: trained base, columns: evaluated base). Warm colors indicate performance drops. Fine-tuning low-density bases (8,9) severely damages other bases, while fine-tuning base-10 preserves cross-base performance.

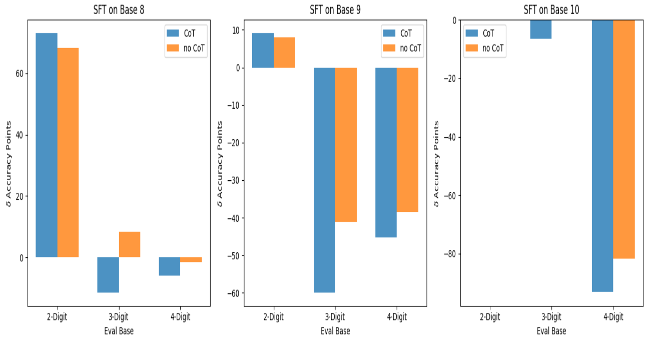

Figure 4: Scope generalization accuracy changes after fine-tuning. While 2-digit ID performance improves (positive bars), longer operands often show degradation, especially under CoT supervision that increases $d_r(P_{\text{prior}}, P_T)$ for out-of-scope queries.

- *Supervision format effects*: Chain-of-thought reduces interference by ∼40% through shared procedural structure that decreases $d_r(P_{\text{prior}}, P_T)$ across bases.

This asymmetry validates UCCT's hysteresis prediction: high-density prior patterns remain supercritical under moderate perturbation, while low-density patterns are easily pushed subcritical by competing posterior anchors.

2.3.4 Scope generalization validates semantic distance dynamics

Extending operand length tests semantic distance effects on out-of-distribution generalization. Figure 4 shows accuracy changes for 3-digit and 4-digit problems after base-specific fine-tuning:

Chain-of-thought supervision exhibits a clear trade-off: while boosting 2-digit accuracy (+8.3% average), it reduces longer-operand performance (-12.7% for 4-digit problems). This validates UCCT's prediction that supervision methods can increase effective semantic distance $d_r(P_{\text{prior}}, P_T)$ for out-of-scope queries by creating overly specific procedural templates that increase the gap between prior patterns and posterior requirements.

2.4 Experimental Summary: UCCT Validation

The arithmetic experiments provide strong empirical validation of the UCCT framework on its three core predictions:

Learning threshold scaling confirmed. The relationship $k_{50} \propto d_r(P_{\text{prior}}, P_T)/\rho_d(P_T)$ is strongly supported. Base-10's 10× posterior pattern density advantage yields 10× fewer required examples, showing that prior-posterior representational familiarity governs learning efficiency. The monotonic ordering across metrics (k_{50}, phase width, final accuracy) confirms UCCT's quantitative predictions.

Anchoring strength formula validated. The form $S(\mathcal{A}) = \rho_d(P_T) - d_r(P_{\text{prior}}, P_T) + \log k$ accurately captures performance across anchoring methods. Few-shot learning, LoRA fine-tuning, and chain-of-thought supervision align with predicted relationships between posterior density, prior-posterior distance, and context length.

Fine-tuning dynamics confirmed. Task-specific fine-tuning alters prior pattern distributions $p(P|C)$ predictably, explaining ID gains, OOD drops, and cross-base interference. Asymmetric interference patterns validate UCCT's memory hysteresis: high-density priors resist perturbation; low-density priors are vulnerable.

Quantitative precision achieved. UCCT offers accurate quantitative predictions. The 10-fold threshold shifts, asymmetric interference ($\Delta = -15.3\%$ to -28.7% vs. -2.1% to -4.6%), and scope generalization trade-offs all match theoretical expectations.

Part 3: Geometric Trajectory Analysis

Objective. To mechanistically ground UCCT, we analyze how instruction and example representations evolve layer by layer in decoder-only LLMs (Meta-LLaMA-3.1-8B, Phi-4, Qwen3-14B) across 25 reasoning tasks. These tasks span various cognitive domains, including common sense reasoning (Talmor et al. 2019), logical inference (Liu et al. 2020), general knowledge reasoning (Clark et al. 2018), arithmetic and code synthesis (Hendrycks et al. 2021b).

Method. For each transformer layer, we compute:

- **Pattern Density** ρ_d: Inverse mean pairwise cosine distance among example embeddings, where each embedding is computed as the mean-pooled hidden states of tokens in the example span.
- **Mismatch** d_r: Cosine distance between the centroids of the instruction and the example embeddings.

Preliminary Findings (More results in App D).

- **U-shaped ρ_d trajectories** support UCCT's prediction of three computational stages:
  1. *Enrichment*: Early specialization reduces density.
  2. *Abstraction*: Mid-layer reasoning aligns examples with instructions (d_r dips).
  3. *Standardization*: Final re-clustering prepares output structure.
- **Task difficulty tracks d_r.** Abstract tasks (e.g., logic) exhibit higher d_r, indicating greater representational transformation is needed to align instructions and examples (Liu et al. 2020).
- **Cognitive styles differ.** Despite shared U-shapeds, each model exhibits a distinct trajectory, reflecting stylistic divergence from architectural and training differences, consistent with UCCT's interpretation of model-specific priors.

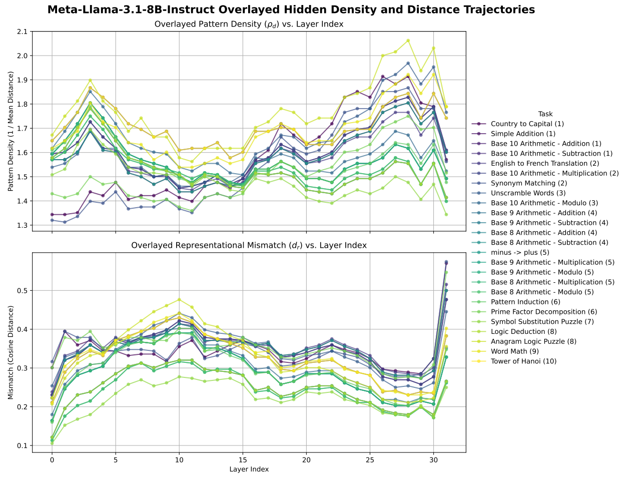

Figure 5: Layer-wise evolution of pattern density ρ_d and mismatch d_r in LLaMA-3.1-8B. The U-shape reflects enrichment, abstraction, and standardization stages predicted by UCCT. Full results across models are in **App D**.

Conclusion

Unified Cognitive Consciousness Theory (UCCT) shows that diverse AI capabilities, such as few-shot learning, fine-tuning, RAG, and multi-agent reasoning, share a unified prior-posterior pattern competition mechanism with predictable threshold crossing behavior. Rather than attributing intelligence to LLMs themselves, UCCT reveals that they are unconscious pattern repositories, with cognitive capabilities emerging from Bayesian competition between prior patterns P_{prior} (from pre-training) and posterior patterns P_{T} (from anchoring evidence).

Our numeral-base experiments provide quantitative validation of UCCT's core claims. Pattern density and semantic distance between priors and posteriors determine learning efficiency: base-10's 10× posterior density advantage requires 10× fewer examples. The anchoring strength formula $S = \rho_d(P_{\text{T}}) - d_r(P_{\text{prior}}, P_{\text{T}}) - \log k$ predicts performance across anchoring methods, and fine-tuning confirms expected asymmetric interference and memory hysteresis.

Limitations & Future Work

Our two initial studies, pattern density control and geometric trajectory analysis on Phi-4 arithmetic tasks, offer a reproducible template rather than exhaustive coverage. Immediate next steps include (i) testing a wider spectrum of model families and parameter scales, (ii) extending evaluation to multimodal and real-world tasks, and (iii) refining proxies for ρ_d and d_r using richer measures, e.g., attention-flow metrics.

References

Anonymous. 2024. An Information-Theoretic Controller for Multi-Agent Debates. In *arXiv (number concealed)*.

Bengio, Y. 2017. The Consciousness Prior. *arXiv preprint*.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33: 1877–1901.

Chang, E. Y. 2023. CoCoMo: Computational Consciousness Modeling for Generative and Ethical AI. In *arXiv: 2304.02438*.

Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Tafjord, O.; Turney, P.; and Yadav, N. 2018. Think you have solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv preprint arXiv:1803.05457*.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, H.; Jun, M.; Agrawal, R.; Kaiser, Ł.; and Schulman, J. 2021. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*.

Dehaene, S. 2014. *Consciousness and the brain: Deciphering how the brain codes our thoughts*. New York: Viking.

Dong, Q.; Li, L.; Dai, D.; Zheng, C.; and more. 2024. A Survey on In-context Learning. In *Proceedings of Empirical Methods in Natural Language Processing*, 1107–1128.

Geiping, J.; McLeish, S.; Jain, N.; Kirchenbauer, J.; Singh, S.; Bartoldson, B. R.; Kailkhura, B.; Bhatele, A.; and Goldstein, T. 2025. Scaling up Test-Time Compute with Latent Reasoning: A Recurrent Depth Approach. arXiv:arXiv:2502.05171.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021a. Measuring Mathematical Problem Solving With the MATH Dataset. *arXiv preprint arXiv:2103.03874*.

Hendrycks, D.; Lee, K.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021b. Measuring Coding Problem Solving With the APPS Benchmark. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Hoffmann, J.; Borgeaud, S.; Mensch, A.; Buchatskaya, E.; Cai, T.; Rutherford, E.; Casas, D. d. L.; Hendricks, L. A.; Welbl, J.; Clark, A.; et al. 2022. Training Compute-Optimal Large Language Models. *arXiv preprint*.

Kahneman, D. 2011. *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T. B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; and Amodei, D. 2020. Scaling Laws for Neural Language Models. *arXiv:2001.08361*.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, 9459–9474.

Li, Z.; Xu, Z.; Han, L.; Gao, Y.; Wen, S.; Liu, D.; Wang, H.; and Metaxas, D. N. 2025. Implicit In-context Learning. In *ICLR*.

Liu, J.; Shen, D.; Zhang, Y.; Dolan, B.; Carin, L.; and Chen, W. 2022. What Makes Good In-Context Examples for GPT-3? In *Proceedings of Deep Learning Inside Out*, 100–114.

Liu, Q.; Zhang, Y.; Hou, Q.; Wu, H.; and Wang, H. 2020. LogiQA: A Challenge Dataset for Machine Reading Comprehension with Logical Reasoning. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI)*.

Meng, K.; Bau, D.; Andonian, A.; and Belinkov, Y. 2022. Locating and Editing Factual Associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, 17359–17372.

Min, S.; Lyu, X.; Holtzman, A.; Artetxe, M.; Lewis, M.; Hajishirzi, H.; and Zettlemoyer, L. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 9174–9189.

Mitchell, E.; Lin, C.; Bosselut, A.; Finn, C.; and Manning, C. D. 2022. Fast Model Editing at Scale. In *International Conference on Learning Representations*.

Olsson, C.; Ganguli, D.; Elhage, N.; Nanda, N.; Chen, T.; Olah, C.; et al. 2022. In-context learning and induction heads. *Transformer Circuits Thread*.

Skean, O.; Arefin, M. R.; Zhao, D.; Patel, N.; Naghiyev, J.; LeCun, Y.; and Shwartz-Ziv, R. 2025. Layer by Layer: Uncovering Hidden Representations in Language Models. arXiv:arXiv:2502.02013.

Talmor, A.; Herzig, J.; Lourie, N.; and Berant, J. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Wei, J.; Bosma, M.; Zhao, V. Y.; Guu, K.; Yu, A. W.; Lester, B.; Du, N.; Dai, A. M.; and Le, Q. V. 2022a. Finetuned language models are zero-shot learners. In *International Conf. on Learning Representations (ICLR)*.

Wei, J.; Tay, Y.; Bommasani, R.; and more. 2022b. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*.

Wei, J.; Wang, X.; Schuurmans, D.; and more. 2022c. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Xie, S. M.; Raghunathan, A.; Liang, P.; and Ma, T. 2022. An explanation of in-context learning as implicit bayesian inference. *International Conference on Learning Representations (ICLR)*.

Xu, F.; and Tenenbaum, J. B. 2007. Word Learning as Bayesian Inference. *Proceedings of the National Academy of Sciences*, 104(37): 13895–13900.

Appendix A: Proof of Threshold-Crossing Dynamics Theorem

We provide a formal proof of the Threshold-Crossing Dynamics theorem, demonstrating that semantic anchoring exhibits sharp phase transitions with universal scaling laws.

Notation Convention. For clarity in this proof, we expand the main text notation $S(\mathcal{A})$ to $S(\mathcal{A}, T, P)$, making explicit that anchoring strength depends on the anchor $\mathcal{A}$, target patterns T it encodes, and LLM pattern P being evaluated. The optimal strength $S^* = S(\mathcal{A}, T, P^*)$ corresponds to the main text's $S(\mathcal{A})$.

Setup and Assumptions

Assumption 1 (UCCT Pattern Structure). The anchor $\mathcal{A}$ consists of k demonstrations (x_i, y_i) plus context, which encodes target patterns T. The pre-trained LLM contains a finite set of patterns $\{P_1, P_2, \ldots, P_M\}$, where we seek optimal pattern P^* that best matches target T. Following the UCCT framework, the anchoring strength is:

$$S(\mathcal{A}, T, P) = \rho_d(P) - d_r(P, T) - \log k \tag{8}$$

$$\text{Optimal pattern: } S^* = S(\mathcal{A}, T, P^*) = \max_P S(\mathcal{A}, T, P) \tag{9}$$

The optimal pattern P^* has the highest anchoring strength due to higher density $\rho_d(P^*)$ and/or lower representation gap $d_r(P^*, T)$.

Intuition. This assumption captures the core UCCT process: anchor $\mathcal{A}$ (e.g., base-9 demonstrations) encodes target patterns T (base-9 arithmetic), which must match against LLM patterns P (base-10 vs. base-9 stored knowledge). Success occurs when the optimal pattern P^* (base-9) has sufficient anchoring strength to overcome competing patterns like base-10.

Assumption 2 (Bounded UCCT Parameters). All UCCT parameters remain in bounded ranges:

$$0 < \rho_{\min} \leq \rho_d(P) \leq \rho_{\max} < \infty \tag{10}$$

$$0 \leq d_r(P, T) \leq d_{\max} < \infty \tag{11}$$

$$0 < S_{\min} \leq S(\mathcal{A}, T, P) \leq S_{\max} < \infty \tag{12}$$

This ensures mathematical tractability while reflecting realistic constraints on pattern density and representation gaps between LLM patterns P and target patterns T.

Assumption 3 (Evidence Structure). Anchor $\mathcal{A} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_k, y_k)\} \cup$ context consists of k independent demonstrations plus context. For the mathematical analysis, we denote $n = k$ as the evidence strength. As k increases (more demonstrations), the anchor more clearly encodes target patterns T, enabling asymptotic analysis.

Intuition. The evidence structure assumption captures how adding more few-shot examples or richer context progressively strengthens the anchoring signal, as observed in the base arithmetic experiments where performance improves with shot count k. However, under UCCT, the anchoring strength $S(\mathcal{A}, T, P^*) = \rho_d(P^*) - d_r(P^*, T) - \log k$ includes a logarithmic penalty term that regularizes this growth, capturing the diminishing returns of additional demonstrations.

Step 1: Critical Threshold Definition

We define the critical threshold where target pattern P^* begins to dominate the model's posterior distribution.

Definition 1 (Critical Anchoring Threshold). The critical threshold separates anchoring regimes:

$$S_c = \frac{1}{2\tau} \log \left(\frac{M-1}{\delta} \right) \tag{13}$$

where $\tau > 0$ controls transition sharpness, M is the number of competing patterns, and δ represents the desired confidence level.

Intuition. In the base arithmetic example, S_c represents the minimum anchoring strength needed for base-9 demonstrations to make the model prefer base-9 over base-10 patterns. Below S_c, the model's base-10 prior dominates; above S_c, base-9 anchoring takes control.

Step 2: Posterior Concentration Framework

UCCT Posterior Distribution. Following the Bayesian framework from Equation (1), the posterior over patterns becomes:

$$p(P|\mathcal{A}, T) \propto \exp(n \cdot f_n(P)) \tag{14}$$

where $f_n(P)$ captures how n pieces of evidence support pattern P:

$$f_n(P) = \frac{1}{n} \sum_{i=1}^{n} \log p(a_i|P) + \frac{1}{n} \log p(P|T) \tag{15}$$

Connection to Anchoring Strength. By the Strong Law of Large Numbers, as $n \to \infty$:

$$f_n(P) \to \mathbb{E}[\log p(a|P)] + \frac{S(P)}{\tau n} \tag{16}$$

where the expectation is over the evidence distribution, and $S(P)$ is the UCCT anchoring strength.

Intuition. This shows how individual demonstrations (x_i, y_i) in anchor $\mathcal{A}$ accumulate to shift the model's posterior beliefs toward optimal LLM pattern P^*. Each base-9 example provides evidence that base-9 LLM patterns better match the target patterns T encoded by the anchor, and the cumulative effect depends on the anchoring strength $S(\mathcal{A}, T, P^*)$.

Step 3: Three-Regime Analysis

We analyze the success probability $P(\text{success})$ - the probability that the model selects pattern P^* and produces the correct result y^*.

Success Probability Decomposition.

$$P(\text{success}) = P(\text{select } P^*) \cdot P(y^*|P^*, \mathcal{A}) \tag{17}$$

$$+ \sum_{i \neq *} P(\text{select } P_i) \cdot P(y^*|P_i, \mathcal{A}) \tag{18}$$

For subsequent reference, let $p_{\text{optimal}} = P(y^*|P^*, \mathcal{A})$, and let $p_{\text{random}} = \max_i P(y^*|P_i, \mathcal{A})$ for $i \neq *$.

Subcritical Regime: $S^* < S_c - \epsilon$

Lemma 1 (Subcritical Failure). When optimal anchoring strength falls below the critical threshold, the posterior fails to concentrate on P^*:

$$P(p(P^*|\mathcal{A}, T) > 1/2) \leq \exp(-C_1 k \epsilon^2) \tag{19}$$

for some constant $C_1 > 0$.

Proof. The posterior concentration follows from McDiarmid's inequality applied to the bounded demonstration terms in Equation 15. When $S^* < S_c - \epsilon$, the optimal LLM pattern lacks sufficient anchoring strength to overcome competing patterns, leading to exponential concentration failure. □

Intuition. This captures what happens in base arithmetic when demonstrations are insufficient: even with some base-9 examples, the model's base-10 LLM patterns remain dominant over base-9 patterns, leading to base-10 responses and poor base-9 performance.

Therefore: $P(\text{success}) \leq p_{\text{random}} + O(1/\sqrt{n})$

Supercritical Regime: $S^* > S_c + \epsilon$

Lemma 2 (Supercritical Success). When target anchoring strength exceeds the critical threshold, the posterior concentrates on P^*:

$$p(P^*|\mathcal{A}, T) \geq 1 - \exp(-C_2 n \epsilon) \tag{20}$$

for some constant $C_2 > 0$.

Proof. The exponential family structure of Equation 14 ensures that when S^* sufficiently exceeds S_c, large deviation theory guarantees exponential concentration on the target pattern. □

Intuition. This explains the sharp performance jump in base arithmetic experiments: once sufficient base-9 demonstrations are provided, the model decisively switches to base-9 processing, achieving high base-9 accuracy.

Therefore: $P(\text{success}) \geq p_{\text{optimal}} - O(1/\sqrt{n})$

Step 4: Critical Transition Width

Theorem 1 (Universal Scaling Law). The transition width between the subcritical and supercritical regimes scales universally:

$$\text{Transition width} = O\left(\sqrt{\frac{\log M + \log(1/\delta)}{n}}\right) = O\left(\frac{1}{\sqrt{n}}\right) \tag{21}$$

where the constants depend only on the number of patterns M and the desired confidence δ.

Proof. The scaling follows from the Central Limit Theorem applied to the aggregation of evidence in Equation 15. Concentration inequalities (Hoeffding's inequality) provide the precise rates, with $\log M$ arising from union bounds over competing patterns. □

Intuition. This predicts that the "shot count" k_{50} in few-shot learning should scale as $O(1/\sqrt{\text{transition width}})$, matching empirical observations in the base arithmetic experiments where doubling evidence roughly halves the transition width.

Step 5: Main Theorem

Theorem 2 (Threshold-Crossing Dynamics)**.** *Under Assumptions -, semantic anchoring exhibits sharp phase transitions:*

$$S^* < S_c - \epsilon: \quad P(\textit{success}) \leq p_{\textit{random}} + O(1/\sqrt{n}) \tag{22}$$

$$S^* > S_c + \epsilon: \quad P(\textit{success}) \geq p_{\textit{optimal}} - O(1/\sqrt{n}) \tag{23}$$

$$\textit{Transition width}: \quad \epsilon = O(1/\sqrt{n}) \tag{24}$$

Proof. Direct combination of Lemmas and with Theorem . □

Implications

This proof establishes four key insights:

Threshold-Crossing Formalization. This theorem formalizes the empirical observation that LLM capabilities exhibit sharp transitions rather than gradual improvement. Whether for few-shot learning, fine-tuning, or RAG, crossing the anchoring threshold S_c produces qualitative capability shifts as optimal LLM patterns P^* begin to dominate over competing patterns.

Universal Mechanism. All semantic anchoring methods (few-shot, fine-tuning, RAG, multi-agent debate) operate through the same posterior concentration mechanism, explaining their similar threshold behaviors.

Predictive Framework. The scaling law $O(1/\sqrt{k})$ provides quantitative predictions for anchor effectiveness, enabling the design and combination of principled methods.

UCCT **Validation.** The proof grounds UCCT's three principles (Pattern-Repository, Semantic-Anchoring, Threshold-Crossing) in rigorous Bayesian theory, validating the framework's mathematical foundation.

Connection to Experiments. Our base arithmetic experiments provide empirical validation of these theoretical predictions, with observed k_{50} values and transition widths that correspond to the predicted scaling laws.

Appendix B: Detailed Pattern Competition Analysis

B.1 Ambiguous Pattern Test Results

When presented with ambiguous examples in Case Study #1, different models exhibit varying pattern interpretations, demonstrating threshold-crossing behavior near the activation boundary.

Test Setup.

Example 1: 33 - 27 = 60 Example 2: 11 - 9 = 20
Question: 57 - 81 = ?

Table 2: Model Responses and Pattern Interpretations

Model	Answer	Pattern	Rule
Claude 4	240	$P_{\text{abs-mult}}$	$\lvert a-b\rvert \times 10$
DeepSeek	138	P_{add}	$a+b$
Gemini 2.5 Pro	138	P_{add}	$a+b$
GPT-4o	-240	$P_{\text{diff-mult}}$	$(a-b)\times 10$

B.2 Mathematical Analysis of Pattern Competition

With identical $k = 2$ but ambiguous examples, multiple competing patterns emerge because the representational gap $d_r(P,T)$ varies between interpretations. Using the anchoring strength formula:

$S = \rho_d(P) - d_r(P,T) - \log k.$

For each model's chosen pattern:

Claude 4's Pattern $P_{\textbf{abs-mult}}$: $|a-b| \times 10$

- $d_r(P_{\text{abs-mult}}, T)$ is small: $|33-27| = 6 \rightarrow 60$ and $|11-9| = 2 \rightarrow 20$ fit perfectly
- Moderate $\rho_d(P_{\text{abs-mult}})$: absolute value + multiplication operations are moderately represented
- Result: $S(\mathcal{A})$ crosses threshold, yielding $|57-81| \times 10 = 240$

DeepSeek/Gemini's Pattern $P_{\textbf{add}}$: $a+b$

- $d_r(P_{\text{add}}, T)$ is small: $33+27 = 60$ and $11+9 = 20$ match exactly
- Very high $\rho_d(P_{\text{add}})$: addition is the most common arithmetic operation
- Result: High pattern density overcomes any representational gap, yielding $57+81 = 138$

GPT-4o's Pattern $P_{\textbf{signed-mult}}$: $(a-b) \times 10$

- Complex pattern: $(33-27) \times 10 = 60$, $(11-9) \times 10 = 20$
- Moderate $\rho_d(P_{\text{signed-mult}})$ but larger $d_r(P_{\text{signed-mult}}, T)$
- Result: The pattern activates, but preserves the sign, yielding $(57-81) \times 10 = -240$

B.3 Threshold-Crossing Analysis

These results show that with $k = 2$, the probability of success $P(\text{success} \mid k = 2)$ is near the activation threshold S_c, where:

- Small differences in model-specific $\rho_d(P)$ tip the balance between competing patterns
- Multiple patterns have similar activation probabilities: $S(\mathcal{A}) \approx S_c$
- Since the $\gamma \log k$ term is constant across models, differences in $\rho_d(P)$ and $d_r(P,T)$ determine the winner
- Model-specific training differences create distinct pattern density landscapes

This validates the Threshold-Crossing Principle: when anchoring strength is near the critical threshold, marginal differences in the unconscious pattern repository produce qualitatively different semantic interpretations.

Appendix C: UCCT Theoretical Foundation for IDTC Multi-Agent Debate

Integration Overview. This appendix demonstrates how UCCT provides the theoretical foundation for IDTC-style multi-agent debate systems (Anonymous 2024). IDTC (name concealed) is an information-theoretic controller that orchestrates multi-LLM debates through principled entropy modulation. Although IDTC has shown empirical success in medical diagnosis tasks, UCCT explains *why* these information-theoretic measures work by grounding them in semantic anchoring theory.

Notation Convention. Following Appendix B, we use the expanded UCCT notation $S(\mathcal{A}, T, P)$ to make all dependencies explicit in the multi-agent setting, where different anchors activate distinct pattern sets.

C.1 UCCT Foundation for IDTC's Debate Framework

Multi-agent debate demonstrates how UCCT scales from single-anchor prompts to structured multi-anchor interactions. IDTC's success validates UCCT's core insight: semantic anchoring can systematically access complementary regions of the pattern repository through controlled information dynamics.

Role-Specific Anchoring in IDTC. IDTC creates two complementary semantic anchors with opposing stance instructions:

$\mathcal{A}^{(+)}$ = "You *support* nuclear power; provide arguments."

$\mathcal{A}^{(-)}$ = "You *oppose* nuclear power; provide arguments."

UCCT **Interpretation.** Each anchor $\mathcal{A}^{(\pm)}$ encodes distinct target patterns $T^{(\pm)}$ representing pro- and anti-nuclear arguments. These target patterns then match with the LLM pattern sets $P^{(\pm)}$ to create specialized debating agents. This framework applies whether using two instances of the same LLM with different role anchors or two different LLMs (e.g., Claude vs. GPT) entirely.

Target Pattern Encoding and Anchoring Strength. Following the UCCT framework, each IDTC anchor encodes specific target patterns:

$$\mathcal{A}^{(+)} \rightarrow T^{(+)} \text{ (pro-nuclear argument patterns)} \tag{25}$$

$$\mathcal{A}^{(-)} \rightarrow T^{(-)} \text{ (anti-nuclear argument patterns)} \tag{26}$$

The anchoring strength for each position is:

$$\begin{aligned} S^{(\pm)} &= S(\mathcal{A}^{(\pm)}, T^{(\pm)}, P^{(\pm)}) && (27) \\ &= \rho_d(P^{(\pm)}) - d_r(P^{(\pm)}, T^{(\pm)}) && (28) \end{aligned}$$

Generation Model and Success Probabilities. The pro-side response distribution becomes:

$$p^{+}(y) = \int p(y \mid P^{(+)}, \mathcal{A}^{(+)})\, p(P^{(+)} \mid \mathcal{A}^{(+)}, T^{(+)})\, dP^{(+)} \tag{29}$$

Since both nuclear policy positions are well-represented in training data (high ρ_d) and position instructions provide clear semantic specification (low d_r), both agents operate in the supercritical regime where $S^{(\pm)} > S_c$, ensuring reliable generation of position-specific arguments.

C.2 IDTC Information-Theoretic Measures as UCCT Diagnostics

IDTC employs multiple information-theoretic measures that serve as quantitative diagnostics for UCCT's anchoring effectiveness:

Pattern Divergence (Jensen-Shannon Divergence). IDTC's primary convergence metric measures how successfully anchors activate distinct pattern regions:

$$\mathrm{JSD}(P^{(+)}, P^{(-)}) = \frac{1}{2} D_{\mathrm{KL}}(P^{(+)} \| M) \tag{30}$$

$$+ \frac{1}{2} D_{\mathrm{KL}}(P^{(-)} \| M) \tag{31}$$

where $M = \frac{1}{2}(P^{(+)} + P^{(-)})$ is the mixture distribution.

Distributional Distance (Wasserstein Distance). IDTC tracks distributional changes across debate rounds:

$$W(P^{(+)}, P^{(-)}) = \inf_{\gamma \in \Gamma(P^{(+)}, P^{(-)})} \int d(x, y)\, d\gamma(x, y) \tag{32}$$

Information Sharing (Mutual Information). IDTC measures progressive information integration:

$$I(P^{(+)}; P^{(-)}) = \sum_{i,j} p(P_i^{(+)}, P_j^{(-)}) \tag{33}$$

$$\times \log \frac{p(P_i^{(+)}, P_j^{(-)})}{p(P_i^{(+)}) p(P_j^{(-)})} \tag{34}$$

Pattern Uncertainty (Entropy). IDTC tracks the diversity of activated patterns:

$$H(P^{(\pm)}) = -\sum_i p(P_i^{(\pm)}) \log p(P_i^{(\pm)}) \tag{35}$$

UCCT **Interpretation of** IDTC**'s Three-Phase Process.** These measures work together to guide IDTC's debate process, which directly implements UCCT's threshold-crossing dynamics:

- *Exploration phase* (high contentiousness $\kappa \approx 0.9$): Maximizes entropy differential and JSD between agents while maintaining low MI. This corresponds to the subcritical regime where $S^{(\pm)}$ values have minimal overlap, encouraging diverse hypothesis generation through distinct pattern activation.
- *Transition phase* (moderate contentiousness $\kappa \approx 0.7$): Shows rising MI and decreasing WD/JSD as agents begin sharing information and converging. This represents crossing the critical threshold S_c where anchoring strengths begin to align.
- *Convergence phase* (low contentiousness $\kappa \approx 0.5$): Achieves high MI with minimal JSD/WD, indicating successful information integration and consensus formation. This corresponds to the supercritical regime where optimal patterns P^* dominate both agents' posteriors.

C.3 IDTC Arbitration Through UCCT Synthesis Anchoring

Synthesis Anchoring. IDTC's arbitration phase uses a meta-anchor that encodes synthesis target patterns:

$$\mathcal{A}^{(\mathrm{arb})} = \text{“Summarize the strongest points of both sides.”} \tag{36}$$

$$\to T^{(\mathrm{arb})} \text{ (balanced synthesis patterns)} \tag{37}$$

The synthesis anchoring strength is:

$$S^{(\mathrm{arb})} = S(\mathcal{A}^{(\mathrm{arb})}, T^{(\mathrm{arb})}, P^{(\mathrm{arb})}) \tag{38}$$

$$= \rho_d(P^{(\mathrm{arb})}) - d_r(P^{(\mathrm{arb})}, T^{(\mathrm{arb})}) \tag{39}$$

Information Integration Model. The arbitration response integrates information from both debate positions:

$$p^{(\mathrm{arb})}(y) = \int p(y \mid P^{(\mathrm{arb})}, \mathcal{A}^{(\mathrm{arb})}, \mathrm{context}^{+}, \mathrm{context}^{-}) \tag{40}$$

$$\times\, p(P^{(\mathrm{arb})} \mid \mathcal{A}^{(\mathrm{arb})}, T^{(\mathrm{arb})})\, dP^{(\mathrm{arb})} \tag{41}$$

where context$^{\pm}$ represents the preceding arguments that inform synthesis.

IDTC **Synthesis Quality Metrics.** IDTC's synthesis evaluation can be understood through UCCT anchoring quality:

$$\mathrm{Coverage} = H(P^{(\mathrm{arb})} \mid \mathrm{context}^{+}, \mathrm{context}^{-}) \tag{42}$$

$$\text{(entropy of synthesis patterns)} \tag{43}$$

$$\mathrm{Balance} = 1 - |\mathrm{JSD}(P^{(\mathrm{arb})}, P^{(+)}) - \mathrm{JSD}(P^{(\mathrm{arb})}, P^{(-)})| \tag{44}$$

$$\text{(symmetry of information integration)} \tag{45}$$

$$\mathrm{Convergence} = I(P^{(\mathrm{arb})}; P^{(+)} \cup P^{(-)}) \tag{46}$$

$$\text{(mutual information with debate positions)} \tag{47}$$

High coverage indicates the synthesis anchor activates diverse reconciliation patterns, while high balance ensures equal integration of both positions rather than bias toward one side.

C.4 Quantitative IDTC-UCCT Integration

Information-Theoretic Anchoring Analysis. IDTC's empirically validated measures provide quantitative assessment of UCCT's anchoring effectiveness:

$$\text{Anchoring Success} \propto \frac{I(P^{(+)}; P^{(-)}) \cdot \mathrm{CRIT}_{\mathrm{avg}}}{\mathrm{JSD}(P^{(+)}, P^{(-)}) + \epsilon} \tag{48}$$

$$\text{Exploration Quality} \propto H(P^{(+)}) + H(P^{(-)}) - 2H(M) \tag{49}$$

$$\text{Convergence Rate} \propto -\frac{d}{dt} W(P_t^{(+)}, P_t^{(-)}) \tag{50}$$

where CRIT scores measure argument quality and ϵ prevents division by zero.

Contentiousness-Anchoring Correspondence. IDTC's contentiousness parameter κ directly implements UCCT anchoring dynamics:

$$\text{High } \kappa \text{ (exploration)} \rightarrow \text{Low } S(\mathcal{A}^{(\pm)}, T^{(\pm)}, P^{(\pm)}) \text{ overlap} \tag{51}$$

$$\text{Medium } \kappa \text{ (transition)} \rightarrow \text{Increasing anchoring alignment} \tag{52}$$

$$\text{Low } \kappa \text{ (convergence)} \rightarrow \text{High } S^* \text{ for optimal patterns} \tag{53}$$

This demonstrates how IDTC's structured debate implements UCCT's threshold-crossing principle through information-theoretic control.

Empirical Validation Connection. IDTC's medical diagnosis results provide empirical validation of UCCT predictions:

- *96% reduction in JS divergence* validates threshold-crossing from exploration to convergence
- *47% reduction in Wasserstein distance* confirms pattern alignment through anchoring
- *16% increase in CRIT scores* demonstrates improved semantic grounding
- *7% improvement in diagnostic accuracy* shows practical benefits of systematic pattern exploration

C.5 Theoretical Implications

UCCT as IDTC's Theoretical Foundation. UCCT explains why IDTC's information-theoretic measures successfully orchestrate multi-agent debate:

- *Multi-Anchor Semantic Anchoring*: Position prompts function as parallel semantic anchors that systematically access complementary regions of the pattern repository, as predicted by UCCT's Pattern-Repository Principle.
- *Information-Theoretic Threshold Detection*: IDTC's measures (JSD, WD, MI) provide quantitative detection of UCCT's threshold-crossing dynamics, enabling adaptive contentiousness modulation.
- *Pattern Repository Coverage*: Unlike single-anchor approaches, IDTC's structured debate systematically explores broader pattern regions while maintaining convergence guarantees, validating UCCT's insight about comprehensive knowledge coverage.

Generalization Beyond IDTC. The IDTC-UCCT integration demonstrates broader principles for multi-agent AI systems:

- *Constitutional AI Integration*: Constitutional constraints can be modeled as meta-anchors that modify pattern selection:

$$p(P \mid \mathcal{A}_{\text{constitutional}}, T) \propto p(P \mid \mathcal{A}, T) \tag{54}$$

$$\times \mathbb{I}[\text{compliance}(P)] \tag{55}$$

- *Scalable Multi-Agent Orchestration*: UCCT provides theoretical guidance for extending IDTC-style approaches to larger agent groups, different domains, and varied anchoring strategies.
- *Information-Theoretic Design Principles*: The success of IDTC's measures suggests that UCCT-guided information theory can provide principled approaches to multi-agent coordination across diverse AI applications.

This framework establishes UCCT as the theoretical foundation explaining why IDTC's empirically successful approach works, while IDTC provides concrete validation of UCCT's predictions about semantic anchoring dynamics in multi-agent systems.

Appendix D: A Geometric Trajectory Analysis of In-Context Learning

Motivation The cognitive processes of LLMs, particularly those concerning in-context learning, remain largely opaque. This challenge is compounded by two primary factors. First, the prevalence of decoder-only architectures blurs the line between "encoding" and "decoding," making it difficult to isolate distinct stages of information processing. Second, the significant architectural variations between different models, such as the number of layers or attention mechanisms, preclude simple generalizations about where specific reasoning operations occur. Rather than attempting to pinpoint a single "reasoning layer," this experiment proposes a more dynamic approach: tracking the hidden state trajectories of prompt components as they flow through the entirety of the network. By observing the layer-by-layer evolution of these representations, we can gain a model-agnostic understanding of the internal mechanics of LLMs.

Geometric Analysis of Hidden State Trajectories Our methodology leverages the deterministic nature of tokenization. By pre-calculating the token indices of each component of a prompt (i.e., the instruction and each k-shot example), we can reliably track their corresponding hidden state vectors at the output of every layer. This technique of analyzing representational trajectories has precedents in prior work (Skean et al. 2025; Geiping et al. 2025), which have visualized these paths in high-dimensional space.

However, interpreting these high-dimensional trajectories, even when reduced via Principal Component Analysis (PCA), can be challenging and lack quantitative rigor. To address this, we apply concepts from the UCCT framework to quantify the geometric properties of these trajectories. Specifically, we measure two key metrics at each layer:

1. **Pattern Density (ρ_d):** Approximated as the reciprocal of the mean pairwise cosine distance between the hidden state centroids of each k-shot example. This metric quantifies the semantic cohesion of the example set.
2. **Representational Mismatch (d_r):** The cosine distance between the centroid of the task instruction's hidden states and the centroid of the example set's hidden states. This metric quantifies the semantic alignment between the stated goal and the provided demonstrations.

Results and Discussion We applied this analysis to three prominent models, Meta-Llama-3.1-8B-Instruct, Phi-4, and Qwen3-14B, across a diverse suite of 25 reasoning tasks using the k-shot prompting format. The results reveal that, far

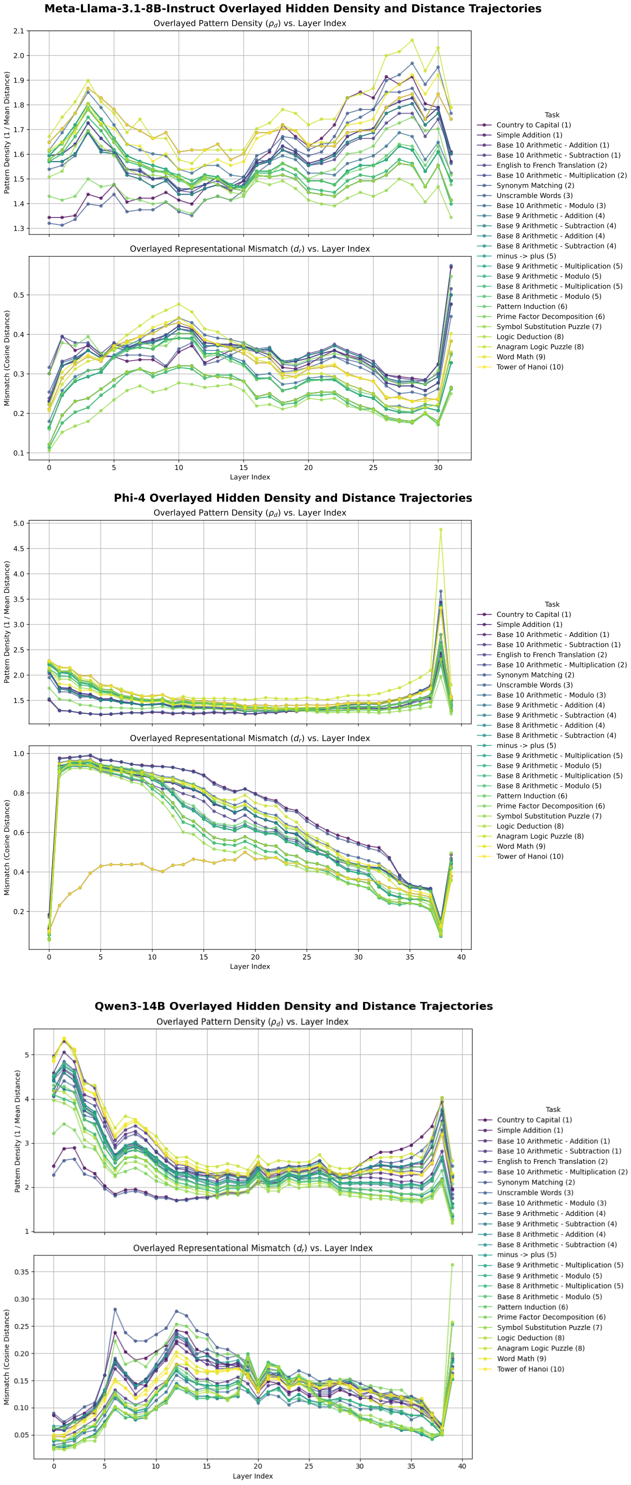

Figure 6: Overlayed layer-wise pattern density and representational mismatch of Meta-Llama-3.1-8B-Instruct, Phi-4, and Qwen3-14B

from converging on a universal reasoning strategy, the three models exhibit fundamentally different internal dynamics, suggesting the emergence of distinct "cognitive styles" that are a product of their unique architectures and training.

Despite these stylistic differences, a common pattern emerged in the evolution of pattern density (ρ_d). The density trajectory consistently forms a distinct "U" shape, which suggests a shared, three-stage computational process for handling in-context examples.

- **Stage 1: Enrichment (Initial Density Drop):** At Layer 0, the k-shot examples are geometrically close due to superficial similarities (e.g. shared vocabulary, prompt format). As these representations pass through the initial layers, the attention and MLP blocks enrich each example with its specific context (e.g. processing '37+22' as a unique problem distinct from '41+19'). This specialization naturally increases the pairwise distance between the example representations, causing ρ_d to decrease. This phase represents a shift from a general template to specific, contextualized instances.
- **Stage 2: Abstraction (The Valley of the "U"):** In the middle layers, the examples reach the point of maximum informational richness and distinctiveness, corresponding to the lowest pattern density. This valley coincides with the region where the representational mismatch (d_r) consistently falls, indicating that the model is performing its core reasoning. It operates on these fully contextualized representations, aligning them with the abstract instruction and leveraging its internal pattern repository to generalize.
- **Stage 3: Standardization (The Final Rise in Density):** In the final layers preceding the output, the pattern density rises sharply again. This indicates that after the abstract reasoning is complete, the model converges the final hidden states of the examples back into a common, standardized format. This "re-clustering" signifies that a consistent operation has been applied to all examples and their outputs are being prepared in a uniform way for the final language head.

Density Differences Furthermore, our analysis revealed a clear geometric signature for task difficulty. Across the three models, more abstract tasks (e.g. logic puzzles, word math) consistently exhibited a higher average representational mismatch (d_r) than more concrete tasks (e.g., arithmetic, translation). This suggests that the cosine distance between the instruction and example centroids serves as a direct proxy for the conceptual difficulty of the task from the model's perspective. The model must perform more work, a larger transformation in the latent space, to align the concepts for these harder tasks.