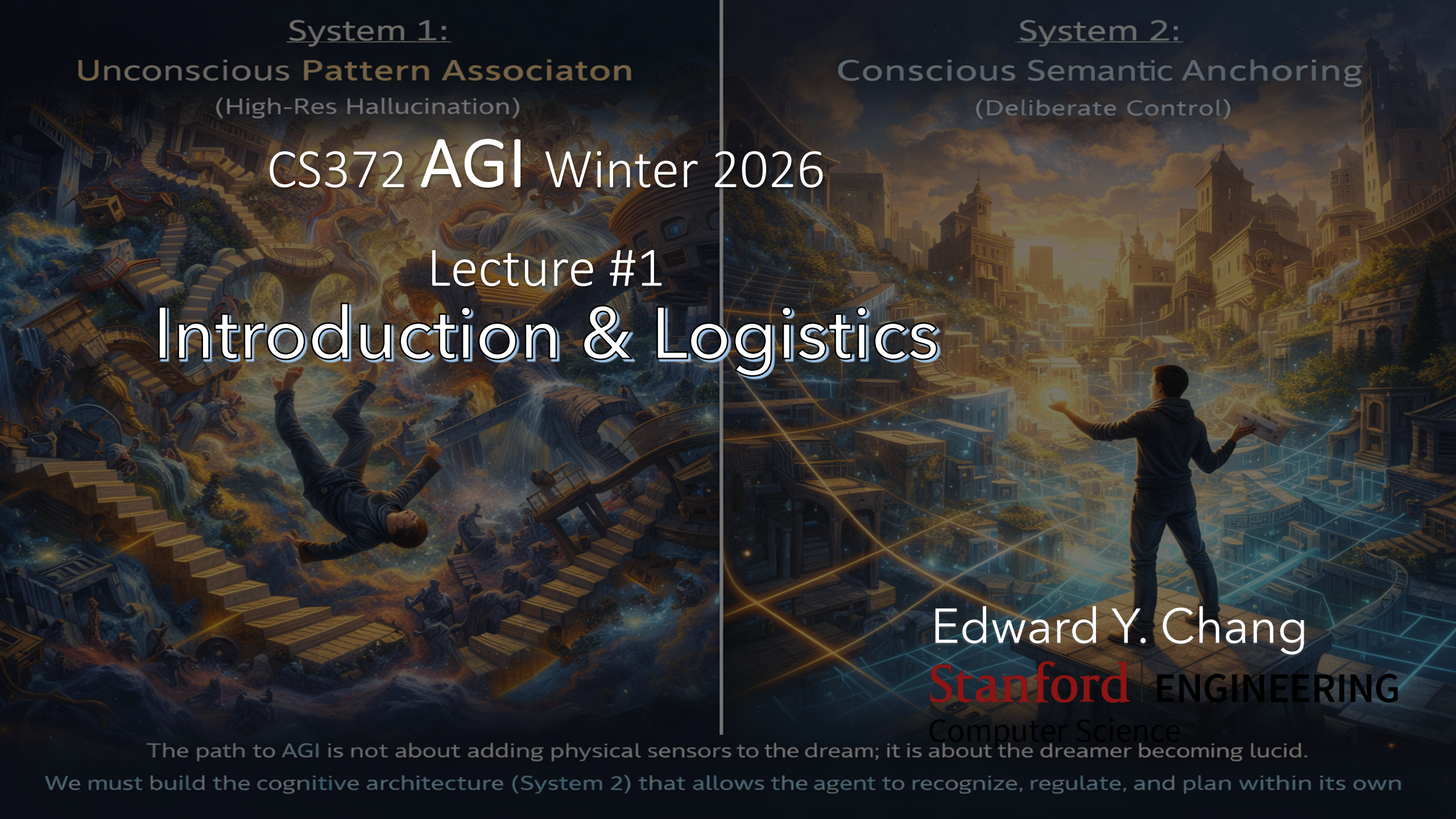


CS372 AGI Winter 2026

Lecture #1

Introduction & Logistics

Edward Y. Chang
Stanford ENGINEERING
Computer Science

The path to AGI is not about adding physical sensors to the dream; it is about the dreamer becoming lucid.
We must build the cognitive architecture (System 2) that allows the agent to recognize, regulate, and plan within its own

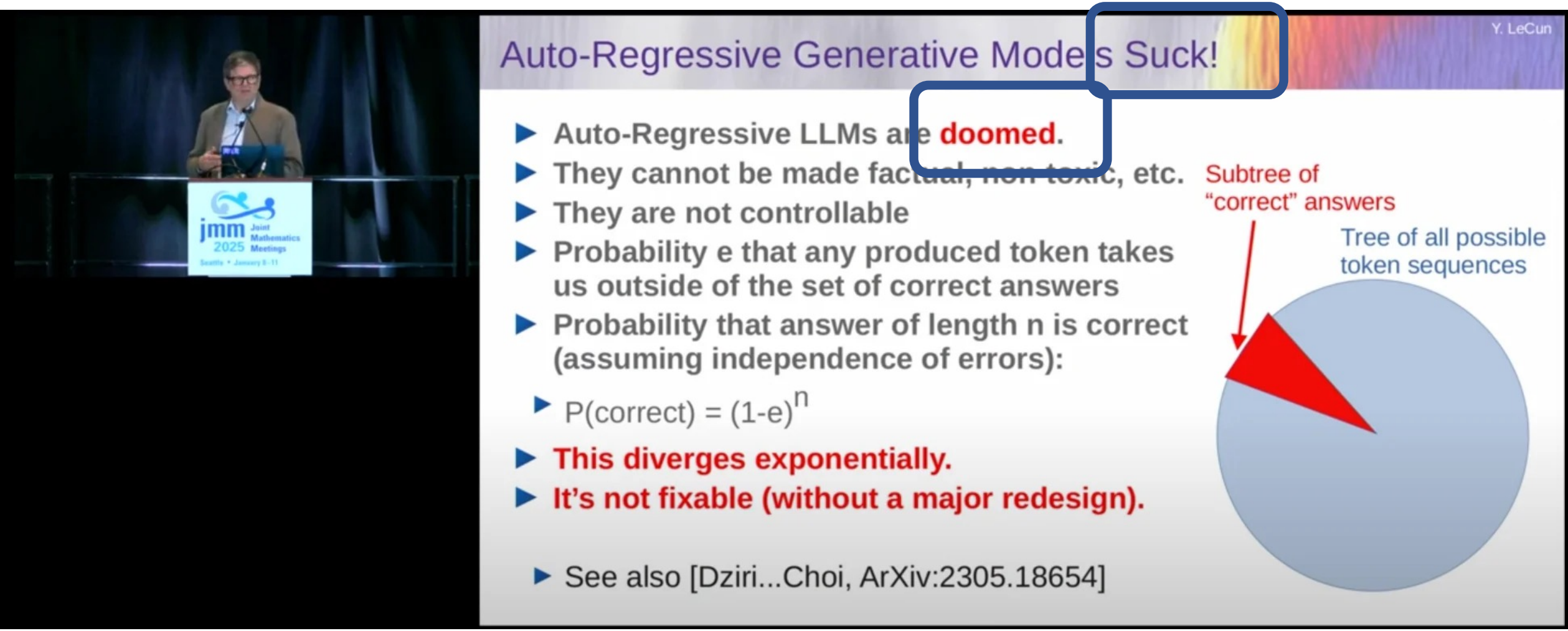
Today's Lecture Outline

- Can LLMs Alone Lead to AGI?
- Syllabus
  - Content & Textbook
  - Course Project and Assignments
  - KDD (February 9), NeurIPS (May)
  - Final Presentation (March 11, 13)

What is AGI? (on-going debate)

- **DeepMind:** A system that achieves at least human-level performance across at least 50% of economically valuable, cognitive, non-physical tasks. (quantifiable, but arbitrary)
- **OpenAI:** An autonomous system that can perform any cognitive task better than most humans and can learn new tasks without human intervention. (qualitative)
- **Gary Marcus :** AGI requires robust, generalizable knowledge representation that supports compositional reasoning, causal inference, and systematic generalization to novel combinations—not merely statistical interpolation within a training distribution. (abstract)

Auto-Regressive LLM Doomed? 2023-25



Can LLMs Alone Reach AGI?

Yann LeCun's arguments against LLMs leading to AGI:

- **Lack of World Models**
  - LLMs don't build internal models of how the world functions
  - Cannot simulate physical interactions or complex systems dynamics
  - Missing the grounding in physical reality necessary for general intelligence
- **Absence of Persistent Memory**
  - LLMs have no stable, updatable memory structure beyond their training data
  - Cannot reliably accumulate new knowledge through experience
  - No mechanism for developing long-term, consistent "beliefs" about the world

…cont

- **Fundamental Reasoning Constraints**
  - Struggle with multi-step logical reasoning, especially when problems require novel approaches
  - Unable to reliably validate their own outputs against *reality*
  - Reasoning is *simulated* rather than structurally implemented
- **Limited Planning Capabilities**
  - Cannot effectively set goals and develop strategies to achieve them
  - No inherent drive or motivation system to prioritize actions

Can LLMs Alone Reach AGI?

Proponents: *Geoffrey Hinton, Demis Hassabis, Sam Altman, Ilya Sutskever, Dario Amodei, et al.*

- Language itself contains implicit knowledge about how the *world* works
- Emergent capabilities observed at scale suggest fundamental reasoning potential via *controlled* debate
- Hybrid systems combining LLMs with other components could address current limitations (e.g., persistent memory)
- The line between "simulating" and "having" intelligence may be philosophically blurrier than critics suggest

LLMs: Necessary but Insufficient for AGI

◎ Lack of World Models

- ♣ Cognitive world models suffice for many reasoning tasks (cf. Dante's *Divine Comedy*—pure imagination, no physics)
- ♣ Absence of physical grounding is *a* limitation, not *the* barrier to AGI

◎ Absence of Persistent Memory

- ♣ Addressable: LLMs can be augmented with external memory for experiential learning
- ♣ Long-term, subjective, beliefs about the cognitive world can be developed and preserved

Physical Grounding: A Limitation, Not The Barrier

LECUN'S CRITIQUE

"LLMs lack embodied experience and physical world models – therefore cannot achieve AGI."

Implicit assumption:
Intelligence requires physical grounding

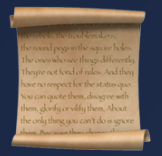
OUR COUNTER

This conflates general intelligence with physical intelligence.

1. Much of human reasoning operates in purely cognitive/abstract domains
2. Causal reasoning about ideas, beliefs, arguments needs no physics
3. Embodiment is necessary for robotics, not general reasoning

Responding to LeCun's position that LLMs are "doomed" for AGI

Evidence: Intelligence Without Physics



c. 1320

Dante's Divine Comedy

A medieval poet constructed a coherent, internally consistent universe with complex causal chains, hierarchical reasoning, and counterfactual logic. No physics. No embodiment. Pure cognitive world model.

∑ Mathematics

Theorem proving requires no physical grounding, pure logical structure

Law

Legal reasoning operates on abstract precedent and principle

◎ Ethics

Moral philosophy reasons about obligations, not objects

∞ Economics

Game theory, mechanism design, abstract agents and payoffs

These domains demonstrate that cognitive world models, not physical ones, underlie vast swaths of human intelligence.

Physical grounding is essential for embodied intelligence but not for cognitive generality

The Real Distinction

Intelligence Type	Requires Grounding?	LLM Capable?
Sensorimotor	YES	No (without augmentation)
Spatial / Physical reasoning	YES	Limited
Social / Cognitive reasoning	NO	**Potentially ✓**
Abstract / Formal reasoning	NO	**Potentially ✓**
Planning in cognitive domains	NO	**Augmentable ✓**

OUR POSITION

Physical grounding is necessary for embodied AI (robots, vehicles), not necessary for cognitive AGI.

IMPLICATION

LeCun's critique targets robotics, not reasoning. LLMs remain a viable path to cognitive AGI.

LLMs: Necessary but Insufficient for AGI

◎ Lack of World Models

- ♣ Cognitive world models suffice for many reasoning tasks (cf. Dante's *Divine Comedy*—pure imagination, no physics)
- ♣ Absence of physical grounding is *a* limitation, not *the* barrier to AGI

◎ Absence of Persistent Memory

- ♣ Addressable: LLMs can be augmented with external memory for experiential learning
- ♣ Long-term, subjective, beliefs about the cognitive world can be developed and preserved

...cont

◎ Limited Planning Capabilities

- ♣ Chain-of-Thought is brittle: cannot reliably decompose goals into executable strategies
- ♣ RLHF reward signal degrades (catastrophic forgetting, reward hacking)
- ♣ No transactional guarantees (atomicity, consistency, isolation, durability)

◎ Fundamental Reasoning Constraints

- ♣ Reasoning is simulated rather than structurally implemented (simulated via association/correlation)
- ♣ Struggle with multi-step logical reasoning, especially novel problems
- ♣ Unable to reliably validate outputs against reality, cognitive or physical

Recommendation Systems

Market Basket Problem (Rakesh, 2000)

Association-rule Based Prediction

To grow the base, we need association rules

- An association rule: $a, b, c \longrightarrow d$
- A Bayesian interpretation: $P(d \mid a, b, c) = \frac{N(a,b,c,d)}{N(a,b,c)}$
- The key is to count the occurrences (*support*) of itemsets $N(\ldots)$

Solutions or Patches to Problems?

Today's LLM Band-Aids, Complex Patches

Problem	Current Bandaid	Why It's Like the Aether
Unreliable reasoning	Prompt engineering • "Think step by step" • Role-playing • Emotional manipulation	Each task needs custom tuning Brittle, non-generalizable
Hallucination	RLHF / Constitutional AI	Patterns checking patterns No ground truth anchor
Missing knowledge	RAG (dump more context)	Like adding aether properties Doesn't fix reasoning
Planning failures	ReAct, Tree-of-Thoughts	Still patterns matching No external validation

Problems with heuristic-based patches?

Human Heuristics Get in the Way

- **AlphaGo (2016)**
  - Seeded with 60 million human expert games
  - Learned human strategies and biases
  - Improved from there RL
- **AlphaGo Zero (2017)**
  - Started from scratch with only game rules, self-supervised learning
  - Discovered novel strategies humans never considered
- Result: AlphaGo Zero defeated AlphaGo 100–0
- Demis Hassabis (DeepMind CEO):

  It's so hard to unlearn human stupidity?

When Patches Fail: The Aether Warning
Tens of thousands of papers = Band-Aids

- **Newton's framework couldn't explain light propagation or Mercury's anomalous orbit**. The luminiferous aether—a hypothetical medium for light waves—became the dominant patch, spawning decades of failed detection experiments and increasingly baroque models (elastic aether, rigid aether, Fitzgerald-Lorentz contraction).
- **Einstein's breakthrough was recognizing the axioms were wrong: space and time aren't fixed containers but flexible dimensions.** This eliminated all the patches at once—no aether needed, Mercury's orbit fell out naturally from curved spacetime.
- **The lesson: when you need endless patches, you're in the wrong paradigm. (218 years** from Newton (1687) to Einstein (1905)'s Special Relativity)

Advantages and Challenges of RLHF

- Pros: avoids the most obvious harms
- Cons: RLHF shortcomings
  - Whack-a-mole, instance level feedback, not generalizable.

Advantages and Challenges of RLHF

- Pros: avoids the most obvious harms
- Cons: RLHF shortcomings
  - Whack-a-mole, instance level feedback, not generalizable.
  - **Conflicting feedback**: Inconsistent judgments on topics such as nudity or alcohol consumption can lead to unpredictable behavior and lack adaptability to local cultures or norms.
  - **Forget effect**: Optimal parameters in LLMs may be unintentionally altered or lost over time, diminishing both the performance of knowledge and alignment.

Shortcomings of RLHF for Ethic Alignment

The Problem: Catastrophic Forgetting
A Classical ML Research Oversight

- What's happening:
  - Researcher fine-tunes on a new 6th cancer type
  - Reports improved accuracy on 6th type ✓
  - Never checks if original 5 types degraded ✗
- This is called:
  - Catastrophic forgetting (neural networks)
  - Negative transfer (transfer learning)
  - Performance regression (production ML)

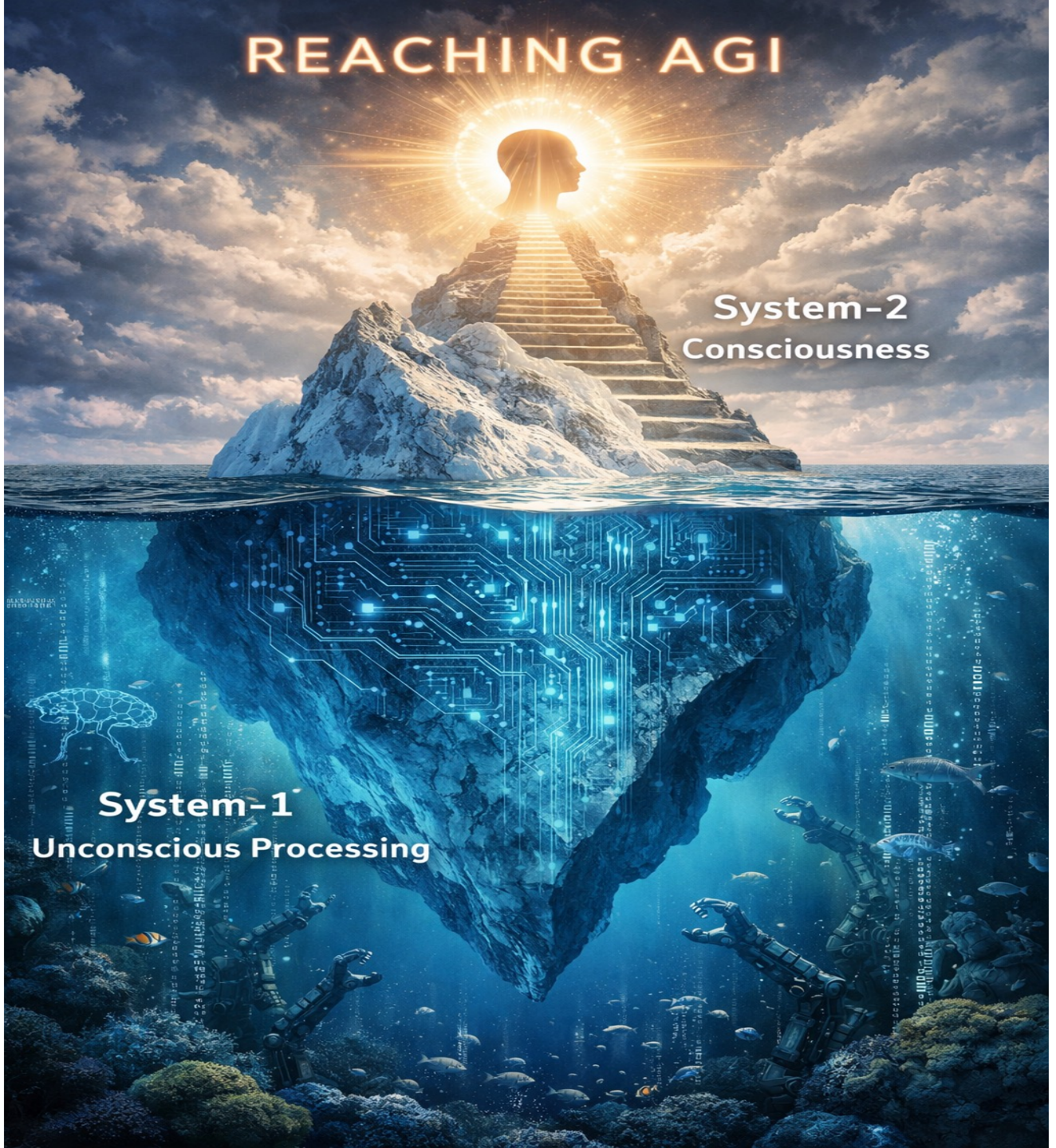
WHEN BANDAIDS FAIL:
From Newton to Einstein, From Prompts to System 2

Large language models should be treated as **unconscious pattern repositories** rather than complete reasoning systems. We outline a two-layer approach in which System-2 orchestration anchors, regulates, validates, and repairs the outputs of System-1 models.

The **Unified Cognitive Consciousness Theory (UCCT)** explains why anchoring and regulation are required for task intelligence.

The **Multi-LLM Agent Collaborative Intelligence (MACI)** framework shows how to achieve this in practice through role design, governance, persistent memory, validator-guided checks, precision retrieval, and localized repair.

The **RAC** paradigm resolves causal reasoning


REACHING AGI
System-2
Consciousness
System-1
Unconscious Processing

The LLM Recipe: A Statistical Cookbook

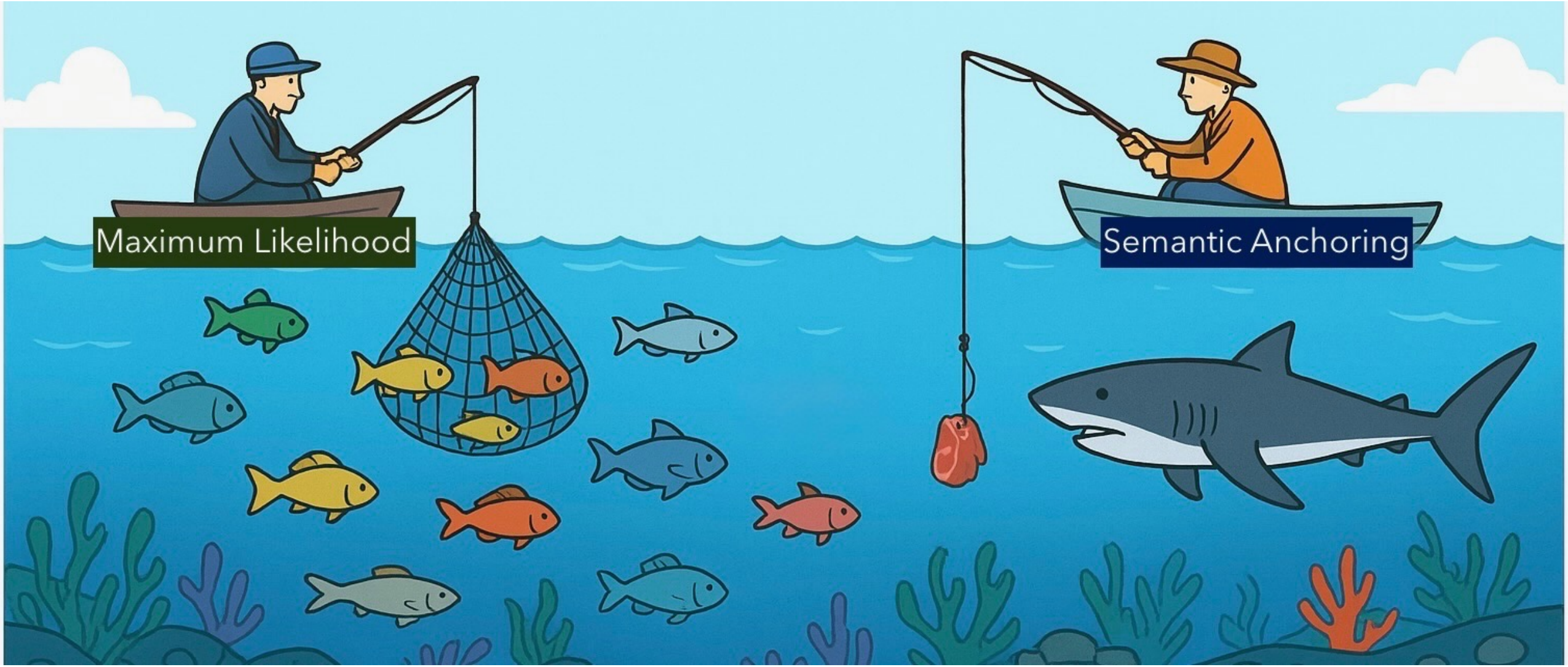
- Ingredients: Massive document collections + Transformers + GPUs
- Method: Next token prediction via maximum likelihood optimization
- Result: **Popularity voting system** given token sequences

Log-Likelihood

$$\textbf{Maximize } \log L(\theta) = \sum_{n=1}^{N} \sum_{t=1}^{T_n} \log p(x_t^{(n)} | x_{<t}^{(n)}; \theta)$$

- What's missing? No semantics!
- Only **patterns** consisting of tokens & tensors

UCCT: Semantic Anchoring



What is 15- 8?

GPT-4o, Claude Sonnet, Gemini 2.5, DeepSeek,

The Power of K-Shot Anchoring

Case Study: Two Examples Override Millions

Two-Shot Pattern Redefinition:

Example 1: 2 - 3 = 5
Example 2: 7 - 4 = 11
Question: 15 - 8 = ?

The Power of K-Shot Anchoring

Case Study: Two Examples Override Millions

Two-Shot Pattern Redefinition:

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

Remarkable finding: Despite millions of associations of "−" with subtraction, just *two examples* achieve complete pattern redefinition!

UCCT Explanation:

- Anchoring context $\mathcal{A}$ provides sufficient evidence
- Shifts posterior $p(P_T|\mathcal{A}, C)$ from subtraction to addition patterns
- Representation gap $d_r(P_{\text{add}}, P_T)$ is minimal—examples are unambiguous

Statistical Pattern Activation vs Simple Substitution

How LLMs Actually Process Few-Shot Examples

Simple Symbol Substitution:

- Pattern: Replace ”-” with ”+”
- Mechanism: Direct token mapping
- Result: Surface-level rule
- [X] This is NOT what happens

Statistical Pattern Activation:

- Pattern: Numerical associations from pre-training
- Mechanism: Posterior redistribution via Bayes
- Result: Statistical inference
- ✓ This IS what happens

Key difference: LLMs don’t learn new rules from examples. They use examples to select which pre-existing statistical patterns to activate.

UCCT insight: Few-shot anchoring works by redistributing probability mass across patterns already learned during pre-training.

- 2 ⟟ 3 = 5; 7 ⟟ 4 = 11
- What is 15 ⟟ 8?

Looking at the pattern:

Given examples:

- 2 ⟟ 3 = 5
- 7 ⟟ 4 = 11

Pattern recognition:

- 2 + 3 = 5 ✓
- 7 + 4 = 11 ✓

The symbol ⟟ represents **addition**.

Therefore: 15 ⟟ 8 = 23

- 2 - 3 = 5; 7 - 4 = 11
- What is 15-8?

Looking at the pattern:

Given examples:

- 2 - 3 = 5
- 7 - 4 = 11

Pattern recognition:

- Normal subtraction: 2 - 3 = -1 ❌ (doesn't match 5)
- But addition: 2 + 3 = 5 ✓
- Normal subtraction: 7 - 4 = 3 ❌ (doesn't match 11)
- But addition: 7 + 4 = 11 ✓

The examples demonstrate that **despite the minus symbol, the operation produces addition results**.

Therefore: 15 - 8 = 23 (interpreting as 15 + 8)

The connection to MLE

The prior $p(P|C)$ emerges from **maximum likelihood estimation** during pre-training.

Pre-training process:

1. Model sees millions of (context, pattern) pairs in train data
2. Learns to maximize $\mathcal{L} = \sum_i \log p(P_i|C_i)$ over the training corpus
3. Results in $p(P|C)$ that assigns highest probability to patterns most frequently associated with context C

Key insight

$p(P|C)$ is the **maximum likelihood estimate** of which patterns should follow which contexts, learned from statistical co-occurrence in pre-training data.

This is why the prior is so strong—it represents the most statistically likely continuation based on massive evidence. Anchoring must overcome this statistical inertia to redirect pattern selection.

The Complete Bayesian Framework

UCCT Generation Model

$$p(y \mid \mathcal{A}, C) = \int \underbrace{p(y \mid P_T, \mathcal{A})}_{\text{pattern speaks}} \underbrace{p(P_T \mid \mathcal{A}, C)}_{\text{chor selects patter}} dP$$

2 1

Two-stage process:

1. **Pattern Selection:** Anchor $\mathcal{A}$ and context C determine which latent pattern P is activated
2. **Response Generation:** Selected pattern P generates tokens y, conditioned on anchor format

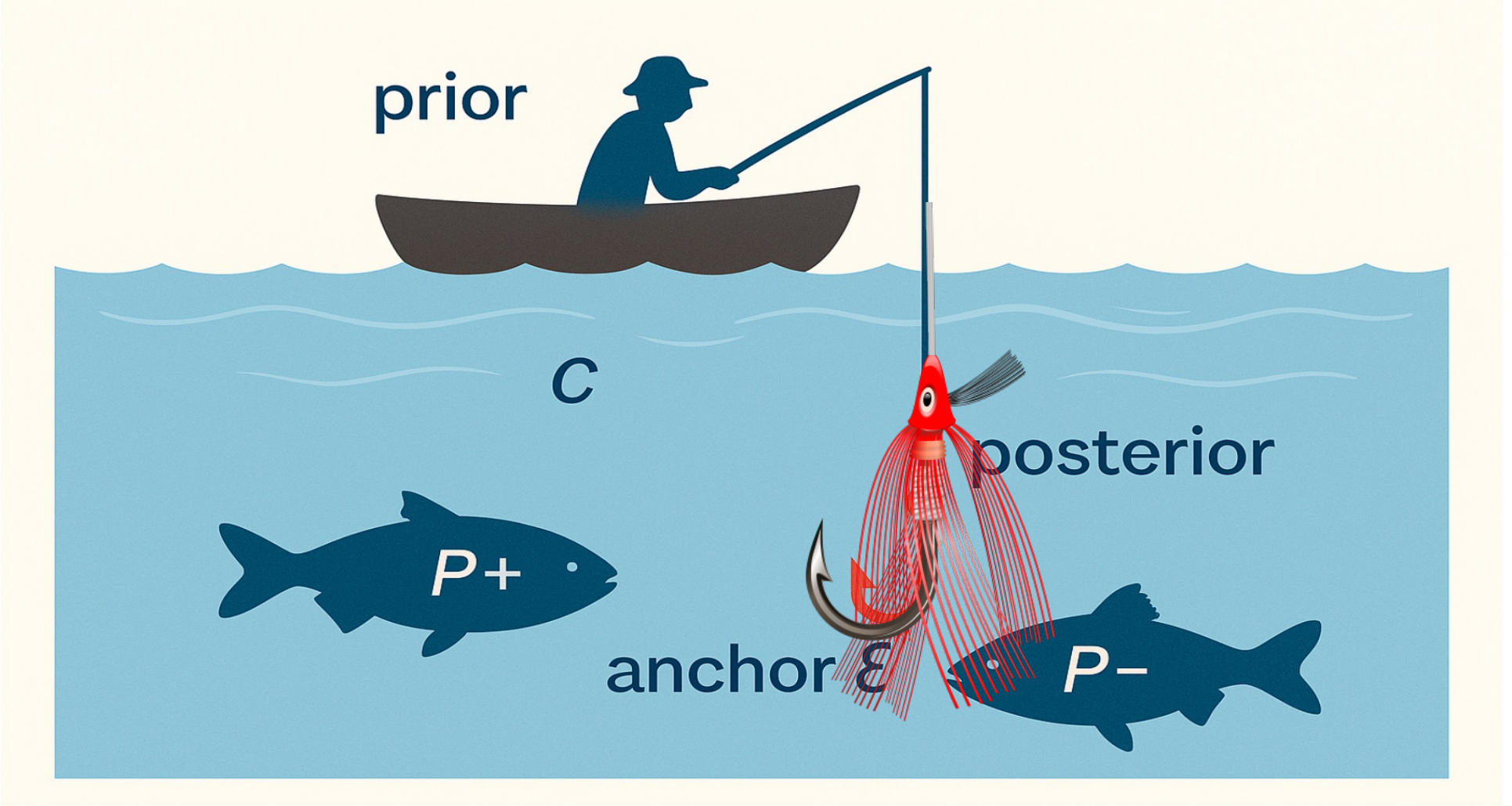
Critical assumption: $y \perp C | P_T, \mathcal{A}$
Once pattern is selected and anchor provides format, old context details don't matter for token generation.

Unified Cognitive Consciousness Theory: Bayesian Competition in Unconscious Pattern Repositories

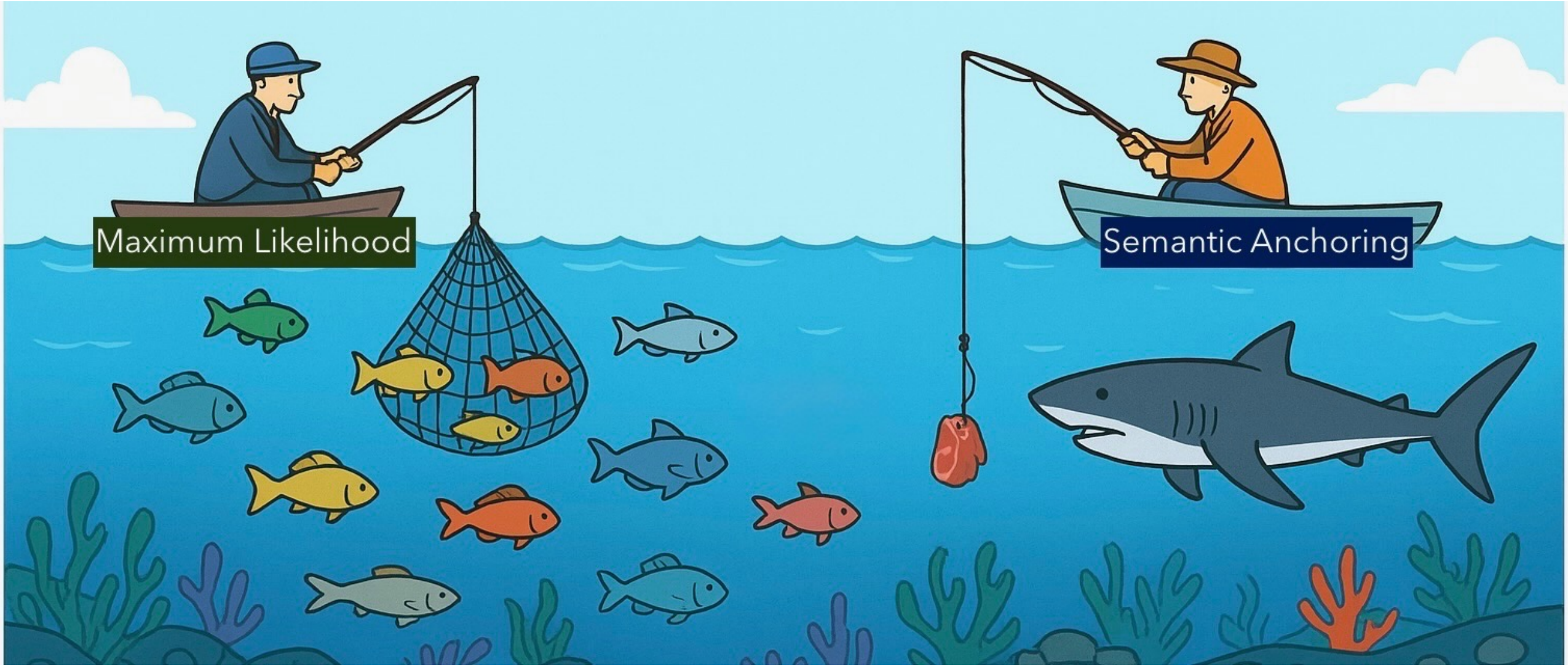
Edward Y. Chang[1], Zeyneb N. Kaya[1], Ethan Chang[2]

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

Unified Cognitive Consciousness Theory
Mathematical Foundation and Illustrations



Fixed Prior → Dynamic Posterior



Critique: Four-year-olds excel at tasks where LLMs struggle

- Critics often note that a four-year-old can learn the concept of "cat" from only a handful of pictures, whereas today's CNNs and LLMs typically require tens of thousands—or even millions—of labeled examples to achieve the same recognition accuracy.

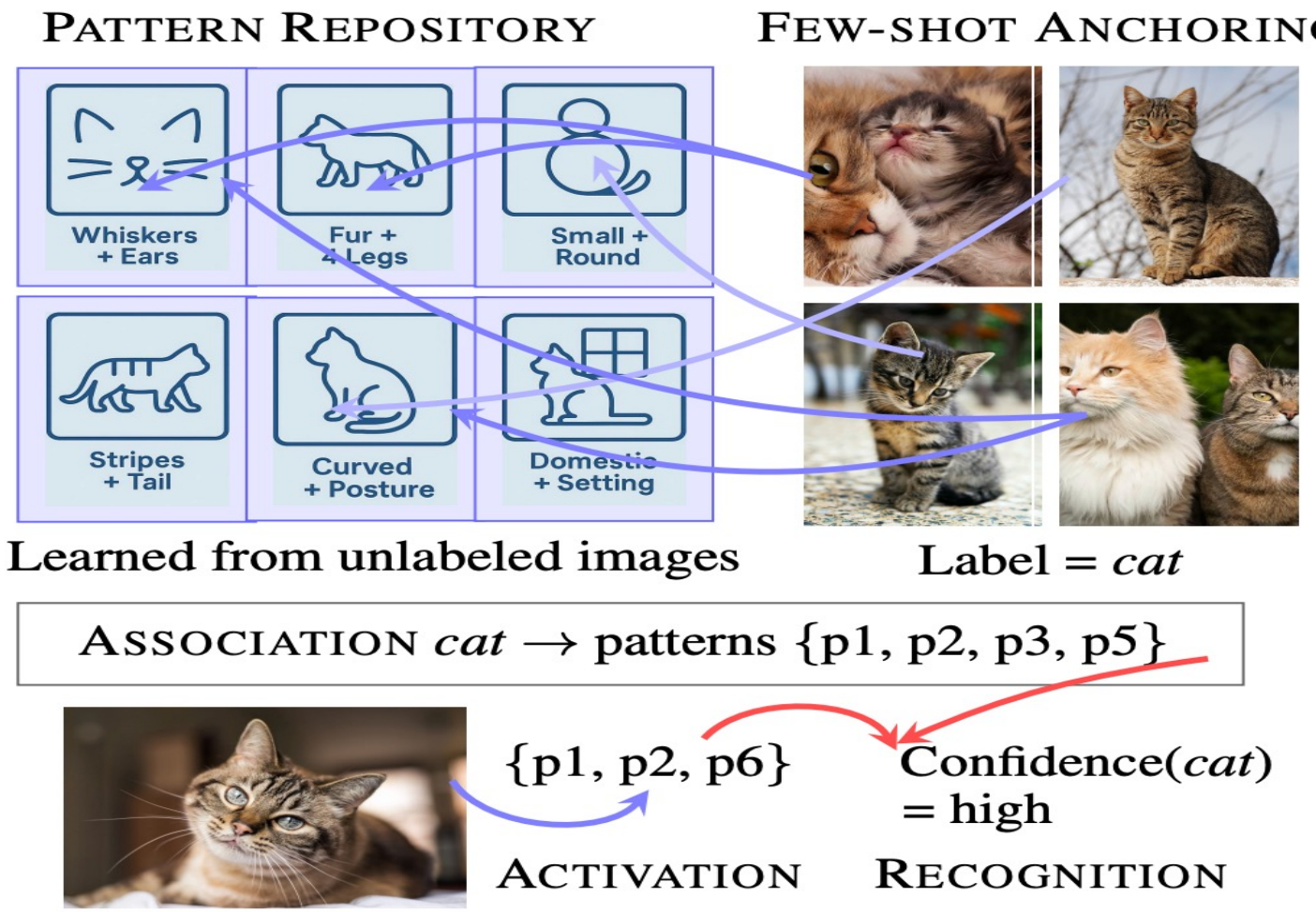


Figure 1: **UCCT Insight:** Intelligence emerges from unconscious patterns + conscious anchoring. Top: few-shots (right) match patterns in the repository (left), yielding the association of *cat* → {p1, p2, p3, p5}. Bottom: test image activates its pattern {p1, p2, p6} and computes the overlap with the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

Unconsciousness vs. Consciousness Explanation

- Critiques belittle that a 4-year-old can recognize “cat” with few sample images, but LLMs and CNNs requires tens of thousands and millions.

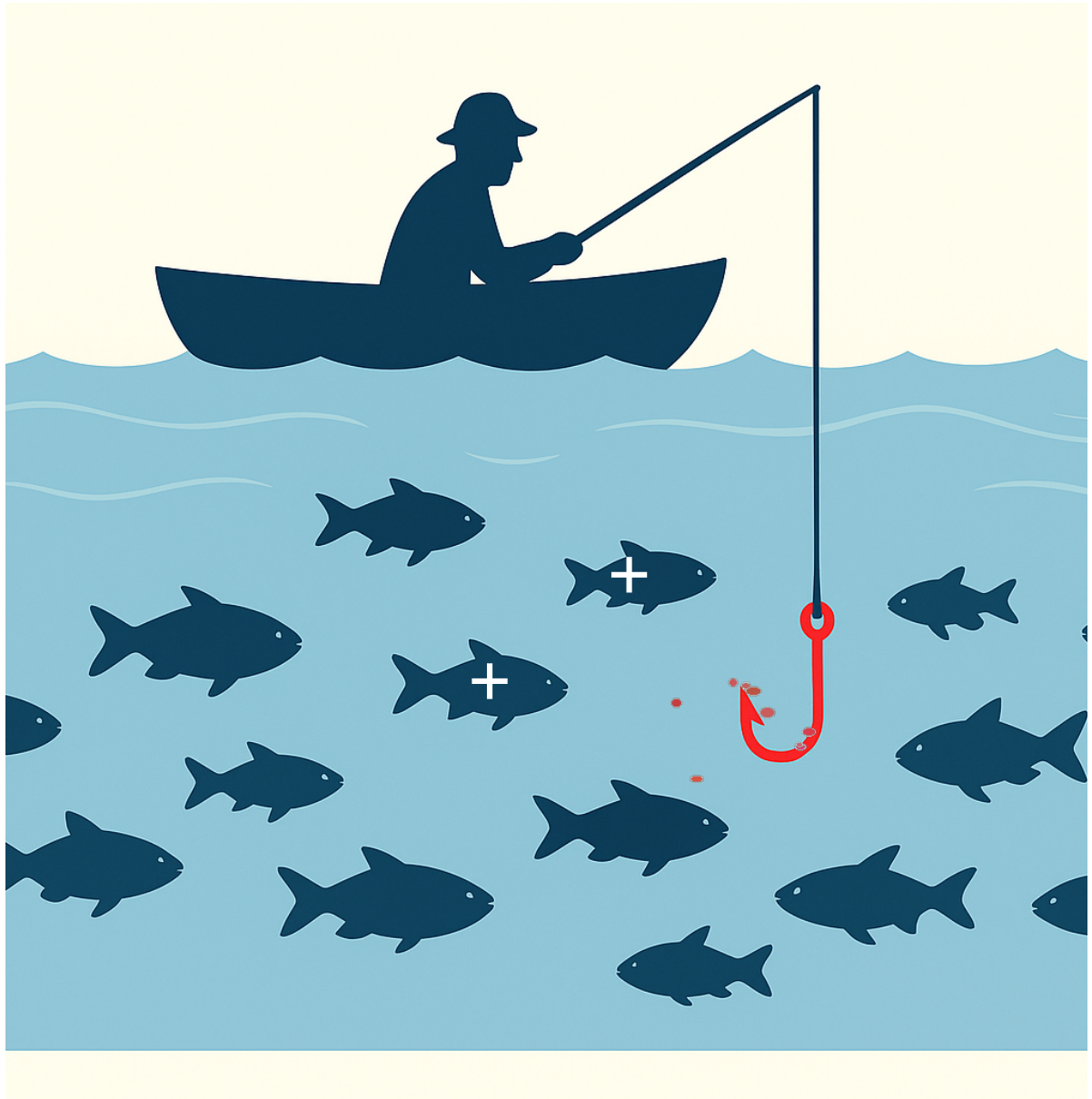
- A 4-year-old was born, not with a blank slate.
- We are born with DNA initialized neural systems, metabolism, vital capabilities, gifted, unlearned.
- Between 0 and 4, we sense, observe, collect patterns, most with no semantics.
- Then on day at 4, UCCT anchoring takes place.

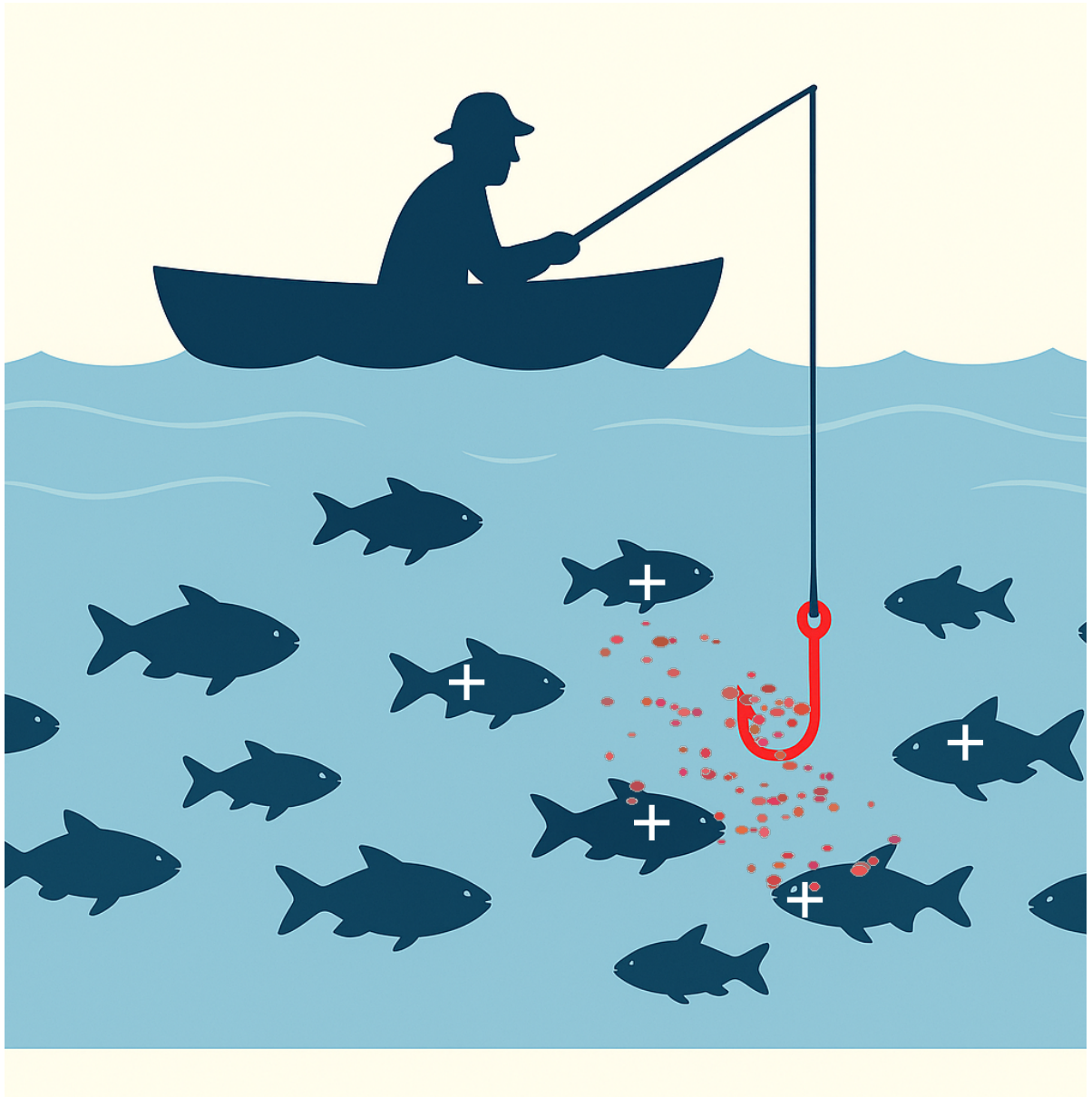
Anchoring phase transition is similarity to unconsciousness to consciousness transition

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

Quantifying Anchoring Strength

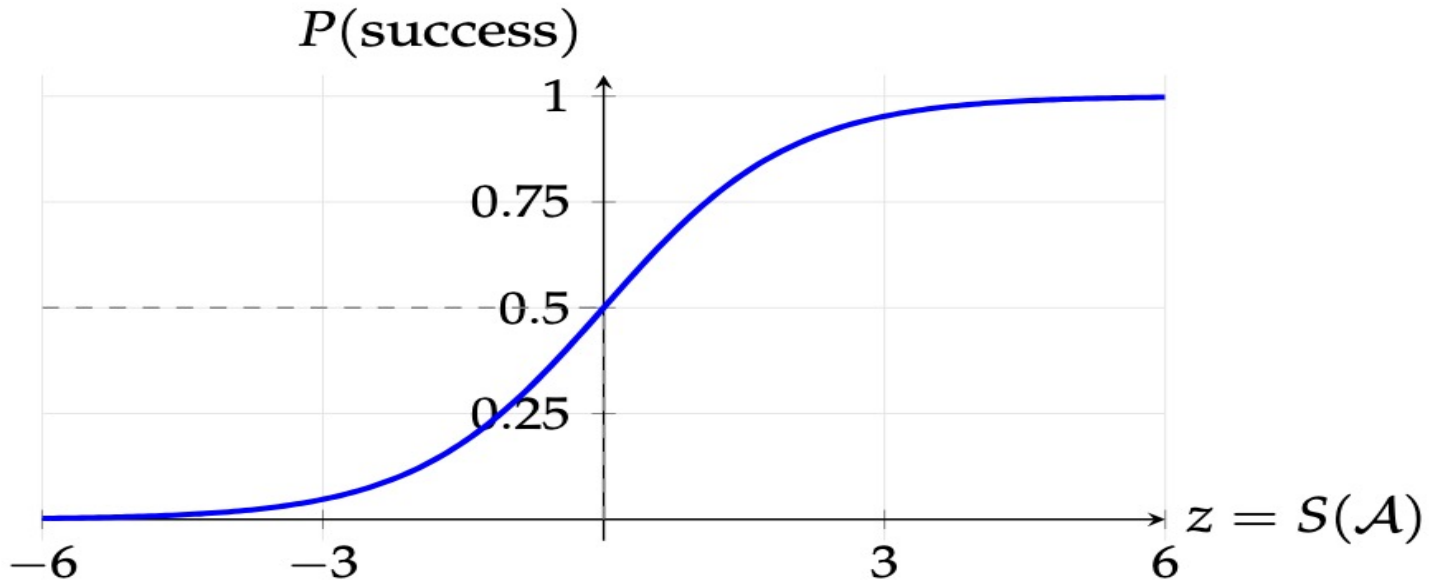
- S(A) : Anchoring strength
- ρ_d(PA) : Patent density
- d_r : Distance between anchoring patterns and prior
- Log k : penalty for token consumption





Phase Transition Function: Sigmoid

Sigmoid curve $\sigma(z) = 1/(1 + e^{-z})$



Near $z = 0$ the slope is steep ⇒ small prompt tweaks can flip failure to success. For $|z| > 3$ the curve saturates—diminishing returns.

Jump-Like Mutations; Quantum Jumps

Discrete Events w/ States

WHAT IS LIFE?

The Physical Aspect of the Living Cell

BY

ERWIN SCHRÖDINGER

SENIOR PROFESSOR AT THE DUBLIN INSTITUTE FOR ADVANCED STUDIES

Based on Lectures delivered under the auspices of the Institute at Trinity College, Dublin, in February 1943

CAMBRIDGE

AT THE UNIVERSITY PRESS

1948

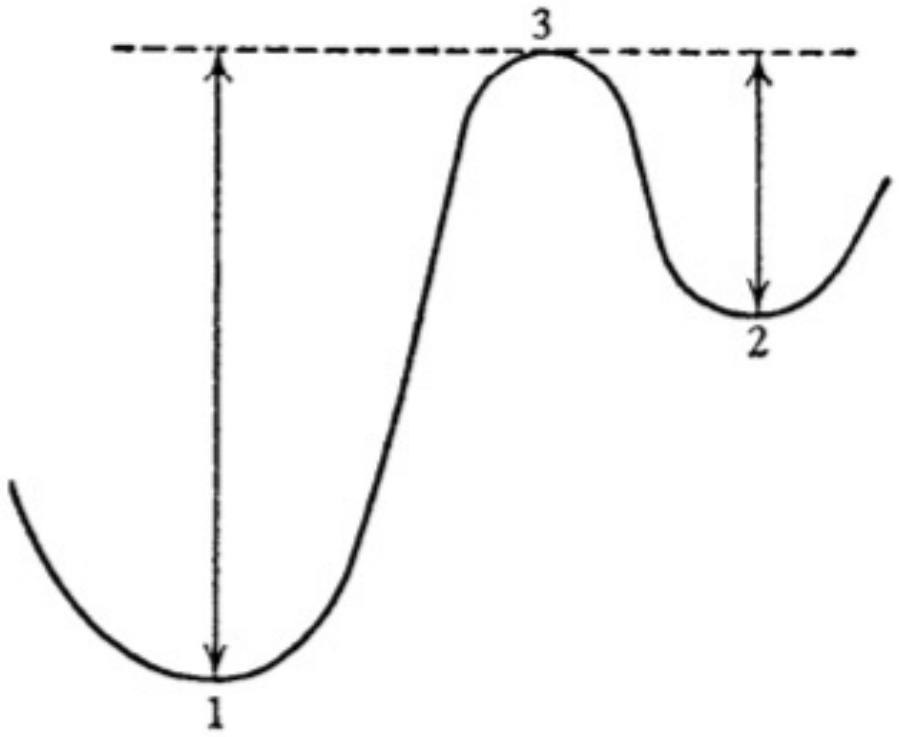
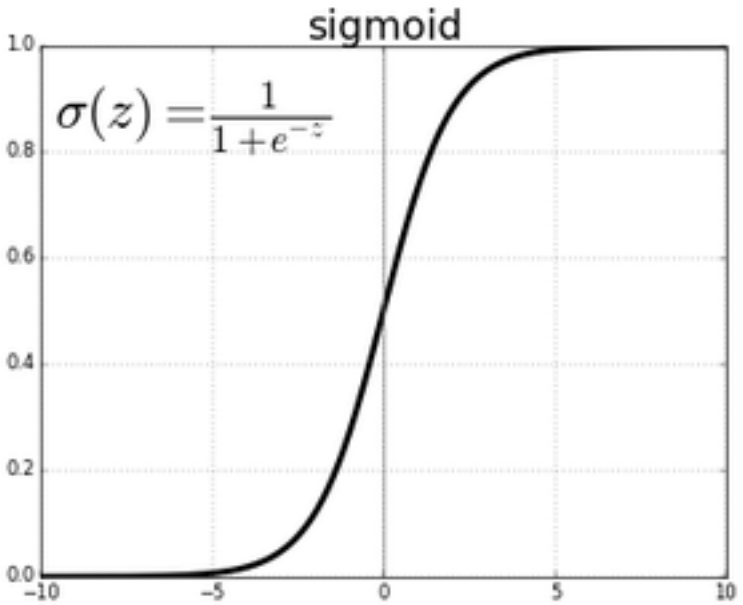


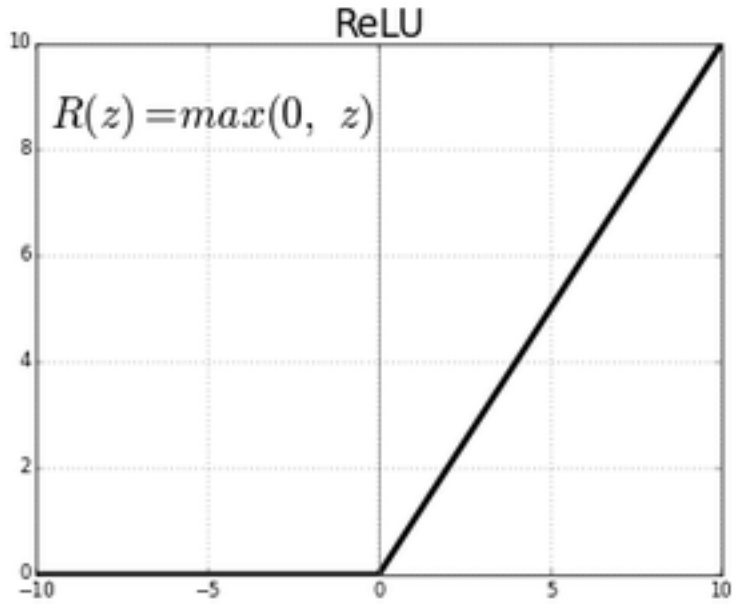
Fig. 12. Energy threshold (3) between the isomeric levels (1) and (2). The arrows indicate the minimum energies required for transition.

Unconsciousness ←→ Consciousness

What is Life, Schrodinger,1954

- Unconsciousness → Consciousness
  - Pain, itchiness, hungry, discomfort
  - "Background" traffic when reading at a coffee shop





UCCT Phase Transitions vs. Scaling Laws
Two Different Predictions

- Key Insight: UCCT predicts phase transitions, not gradual scaling. When anchoring strength crosses critical threshold W*, understanding emerges suddenly—like water freezing at 0°C.
- Evidence: Few-shot learning thresholds, grokking phenomena, prompt sensitivity

WHEN BANDAIDS FAIL:
From Newton to Einstein, From Prompts to System 2

Large language models should be treated as **unconscious pattern repositories** rather than complete reasoning systems. We outline a two-layer approach in which System-2 orchestration anchors, regulates, validates, and repairs the outputs of System-1 models.

The **Unified Cognitive Consciousness Theory (UCCT)** explains why anchoring and regulation are required for task intelligence.

The **Multi-LLM Agent Collaborative Intelligence (MACI)** framework shows how to achieve this in practice through role design, governance, persistent memory, validator-guided checks, precision retrieval, and localized repair.

The **RAC** paradigm resolves causal reasoning

- Information (anchoring) ←→ Linguistic Patterns
- Behavior ←→ Emotion

- Semantic Anchoring → Linguistic Patterns ← Behavior


MACI
AGI
SYSTEM-2
UCCT/CoCoMo
CRIT
SocraSynth/Evince
BEAM
DIKE & ERIS
SagaLLM/ALAS
Precision RAG
Polynthesis
SYSTEM-1: PATTERN REPOSITORY
(LLMs, embeddings, retrieval)

Dual-Dial Multi-Agent Debate (MACI): What & How

- Two independent dials: an **information dial** (τ) that gates what evidence can enter; a **behavior dial** (CL) that schedules contentiousness from explore → consolidate.

Contentiousness = 100%, going nowhere

Contentiousness = 90%, bring out various perspectives

echang@cs.stanford.edu

Contentiousness < 10%, happy meals

Conditional Behaviors with Contentiousness from High to Low

C.L.	Tone	Emphasis	Language
0.9	Most confrontational; raising strong ethical, scientific, and social objections.	Highlighting risks and downsides; ethical quandaries, unintended consequences, and exacerbation of inequalities.	Definitive and polarizing, e.g., "should NOT be allowed," "unacceptable risks," "inevitable disparities."
0.7	Still confrontational but open to some benefits, albeit overshadowed by negatives.	Acknowledging that some frameworks could make it safer or more equitable, while cautioning against its impl. challenges.	Less polarizing; "serious concerns remain," "needs more scrutiny."
0.5	Balanced; neither advocating strongly for nor against.	Equal weight on pros and cons; looking for a middle ground.	Neutral; "should be carefully considered," "both benefits and risks."
0.3	More agreeable than confrontational, with reservations.	Supportive but cautious; focus on ensuring ethical and equitable use.	Positive but careful; "impetus to ensure," "transformative potential."
0.0	Completely agreeable & supportive.	Focused on immense potential benefits; advocating for proactive adoption.	Very positive; "groundbreaking advance," "new era of possibilities."

WHEN BANDAIDS FAIL:
From Newton to Einstein, From Prompts to System 2

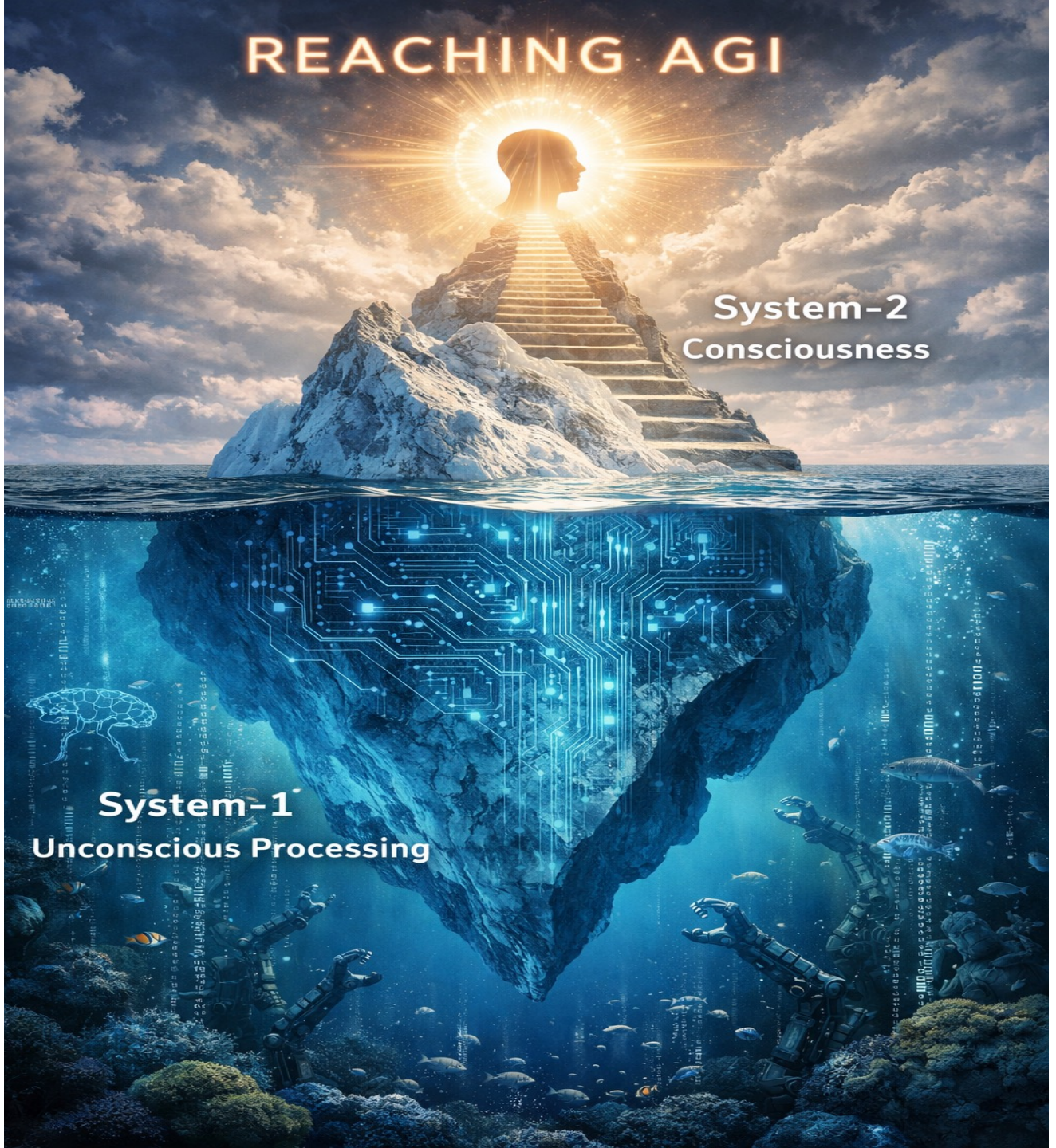
Large language models should be treated as **unconscious pattern repositories** rather than complete reasoning systems. We outline a two-layer approach in which System-2 orchestration anchors, regulates, validates, and repairs the outputs of System-1 models.

The **Unified Cognitive Consciousness Theory (UCCT)** explains why anchoring and regulation are required for task intelligence.

The **Multi-LLM Agent Collaborative Intelligence (MACI)** framework shows how to achieve this in practice through role design, governance, persistent memory, validator-guided checks, precision retrieval, and localized repair.

The **RAC** paradigm resolves causal reasoning

Real Problems?


REACHING AGI
System-2
Consciousness
System-1
Unconscious Processing

Multi-LLM Agent Collaborative Intelligence

The Path to Artificial General Intelligence

Edward Y. Chang

ASSOCIATION FOR COMPUTING MACHINERY

System-2 Reasoning

From Semantic Anchoring to Causal Intelligence

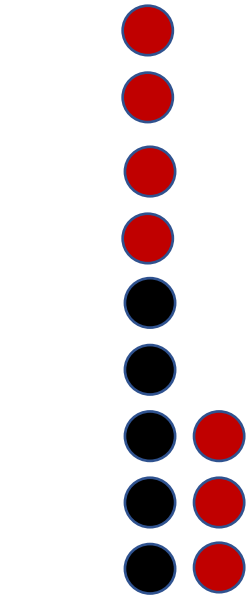
The Path to Artificial General Intelligence Vol. II

Edward Y. Chang

ang@cs.stanford.edu

Gaps between Sysrem-1 and System-2

Gap	System-1 (LLM)	System-2 (AGI)
Patterns vs. Semantics	Statistical regularity	Meaning, reference, truth
Individual vs. Collective	Single-agent	Multi-agent, societal
Self-verification	External oracle	Internal consistency checking
Observation vs. Counterfactual	What happened	What would have happened
Association vs. Causation	$P(Y\|X)$	$P(Y\|\text{do}(X))$
Perspective-taking	Context-free	Situated (spatial, temporal, social)
Retrospection / Prospection	Stateless prediction	Regret, anticipation, planning
Metacognition	Implicit confidence	Explicit uncertainty, "I don't know"
Abstraction / Analogy	Instance-bound	Relational structure transfer
Grounding	Symbolic only	Embodied (partial requirement)



Computer Vision


System 1:
Unconscious Pattern Associaton
(High-Res Hallucination)
System 2:
Conscious Semantic Anchoring
(Deliberate Control)
Causal Reasoning
Planning
The path to AGI is not about adding physical sensors to the dream; it is about the dreamer becoming lucid.
We must build the cognitive architecture (System 2) that allows the agent to recognize, regulate, and plan within its own

Causation vs. Correlation

These two variables are strongly correlated—as ice cream sales increase, drowning deaths increase proportionally. A naive statistical model would conclude that ice cream consumption causes drowning (or perhaps that drownings drive people to eat ice cream as comfort food).

The actual structure: both variables are caused by a confounding variable—summer heat. Hot weather independently increases ice cream purchases AND increases swimming activity, which increases drowning risk. The ice cream-drowning correlation is entirely spurious; intervening on ice cream sales (banning ice cream) would have zero effect on drowning rates.

Today's Lecture Outline

- Can LLMs Alone Lead to AGI?
- Syllabus
  - Content & Textbook
  - Course Project and Assignments
  - KDD (February 9), NeurIPS (May)
  - Final Presentation (March 11, 13)

CA & Project Introduction

Stanford CS372 Winter 2026 Course Project
Scaling Up T^3 Benchmark for Causal Reasoning

Edward Y. Chang (Instructor), Longling Geng (CA)
Stanford University