Reasoning LLMs

AI with Deep Learning

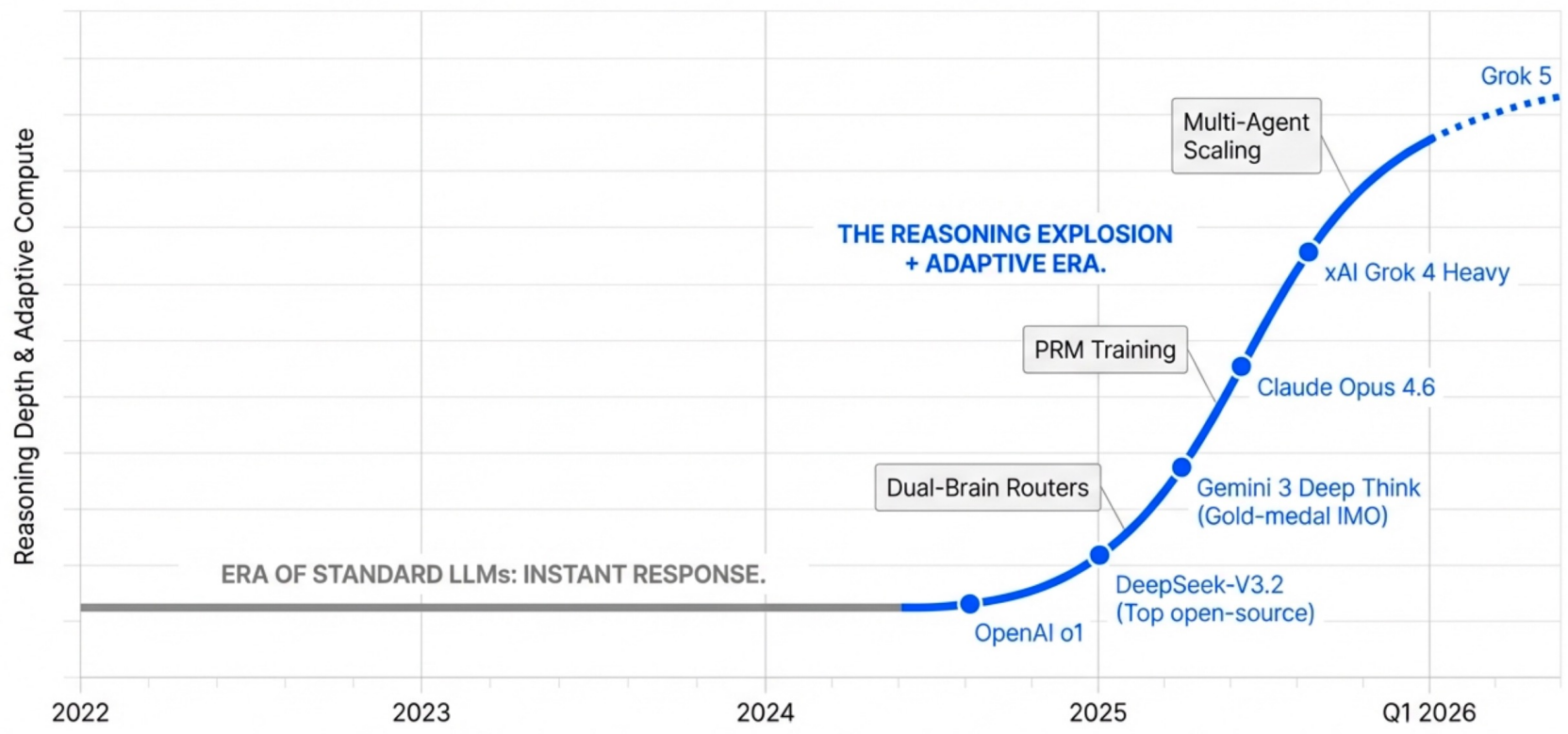
Prof. Lai-Man Po

Department of Electrical Engineering

City University of Hong Kong

The 2026 LLM Landscape: The Reasoning Explosion Accelerates

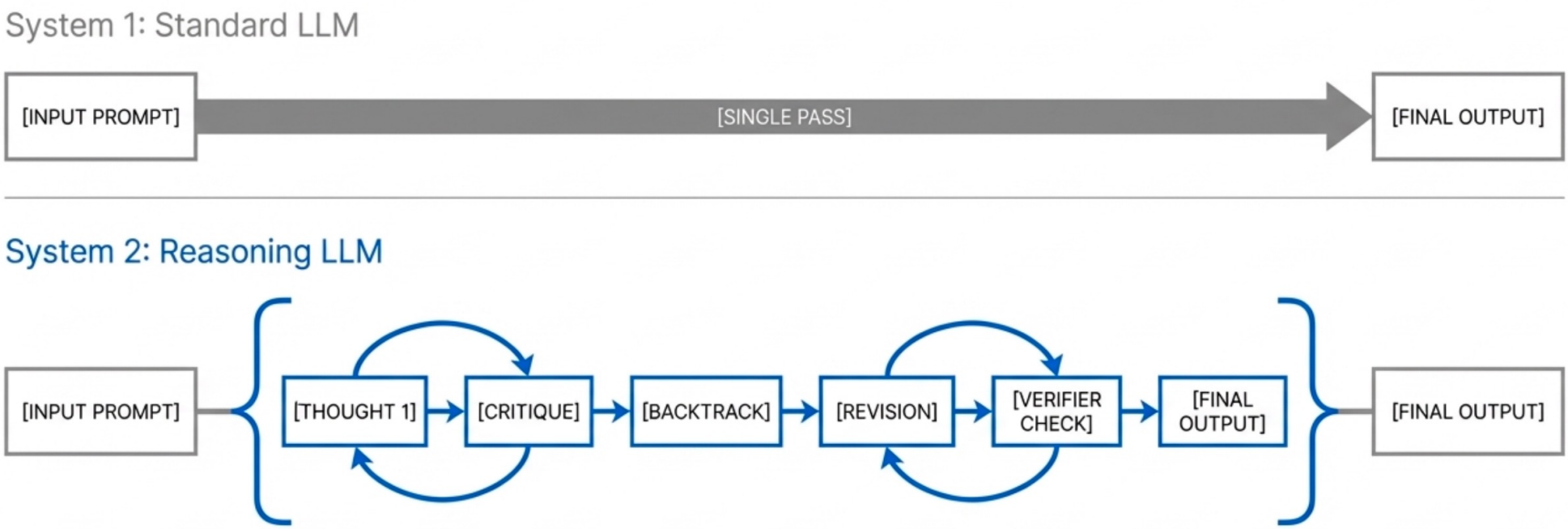
The industry has moved from next-token prediction to compute-heavy deliberation.



The Evolution of Intelligence

	Standard LLMs (Pre-2024)	**Reasoning LLMs (2025-2026+)**
Cognitive Metaphor	System 1 (Fast, intuitive pattern matching)	**System 2 (Slow, analytical deliberation)**
Compute Allocation	Heavy Pre-training	**Heavy Inference / Dynamic**
Output Speed	Instant	**Adaptive / Delayed**
Core Mechanism	Next-token prediction	**Intermediate token deliberation**
Primary Value	Fluency & Speed	**Reliability & Truth**

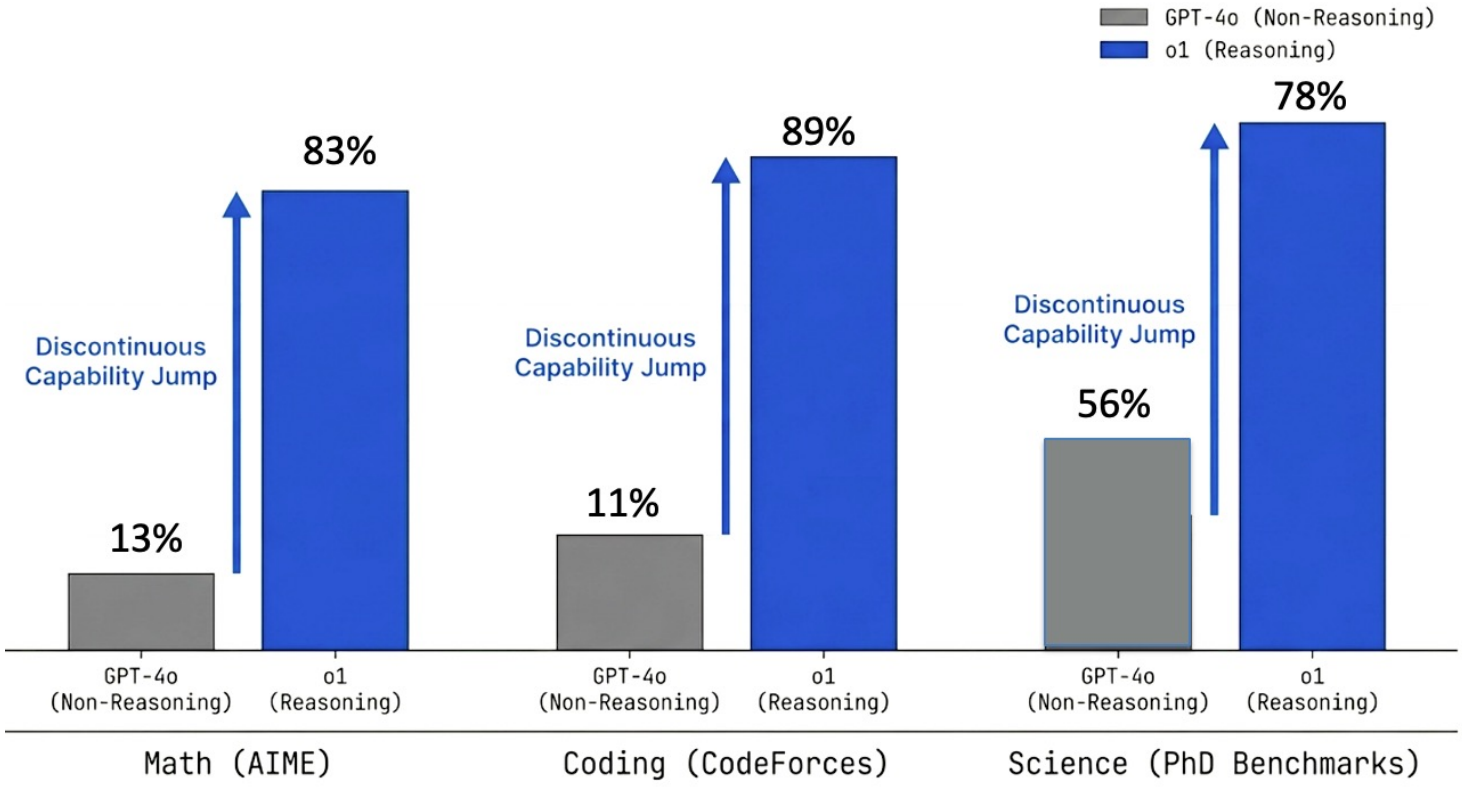
Defining Reasoning: The Hidden Computation



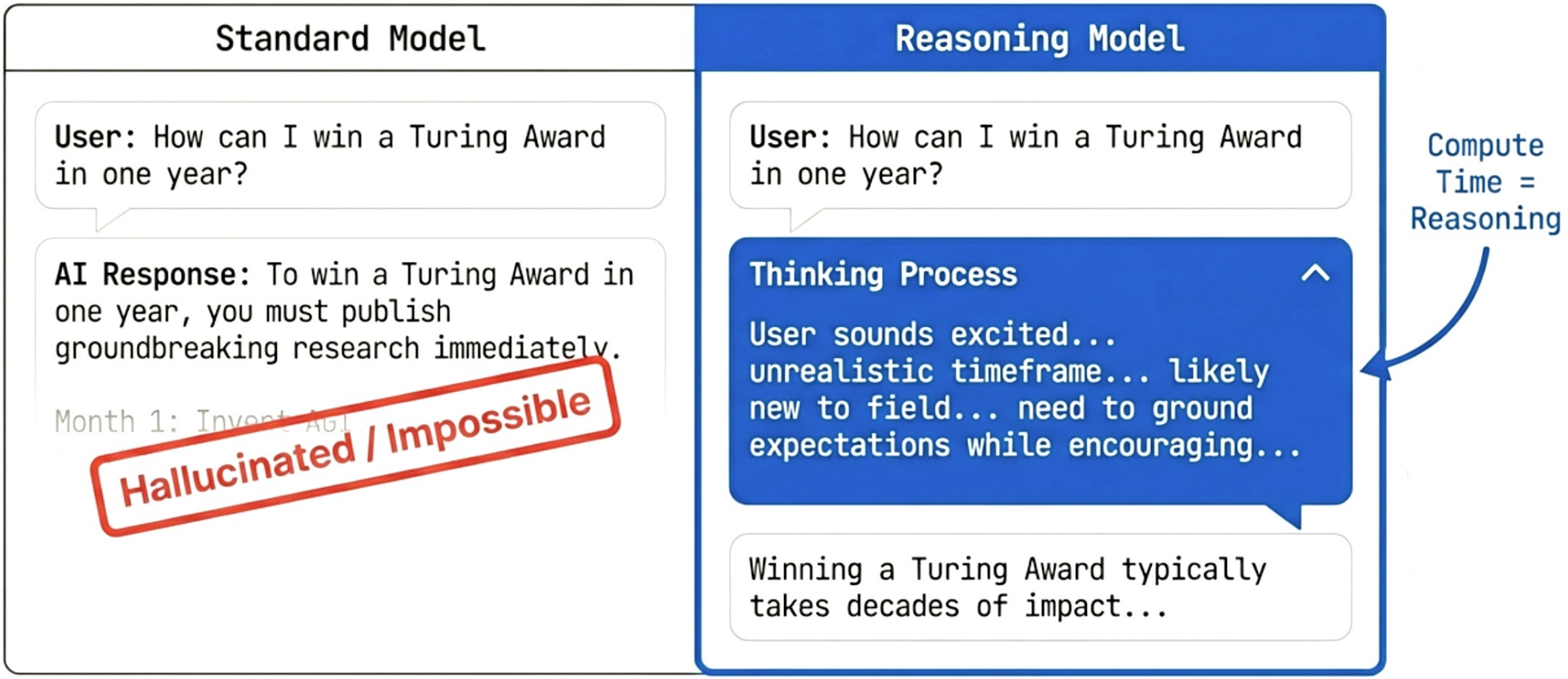
Reasoning in LLMs is defined as **the presence of intermediate tokens** between the question and the final answer." - Denny Zhou, Google DeepMind

The Performance Gap: GPT-4o and o1

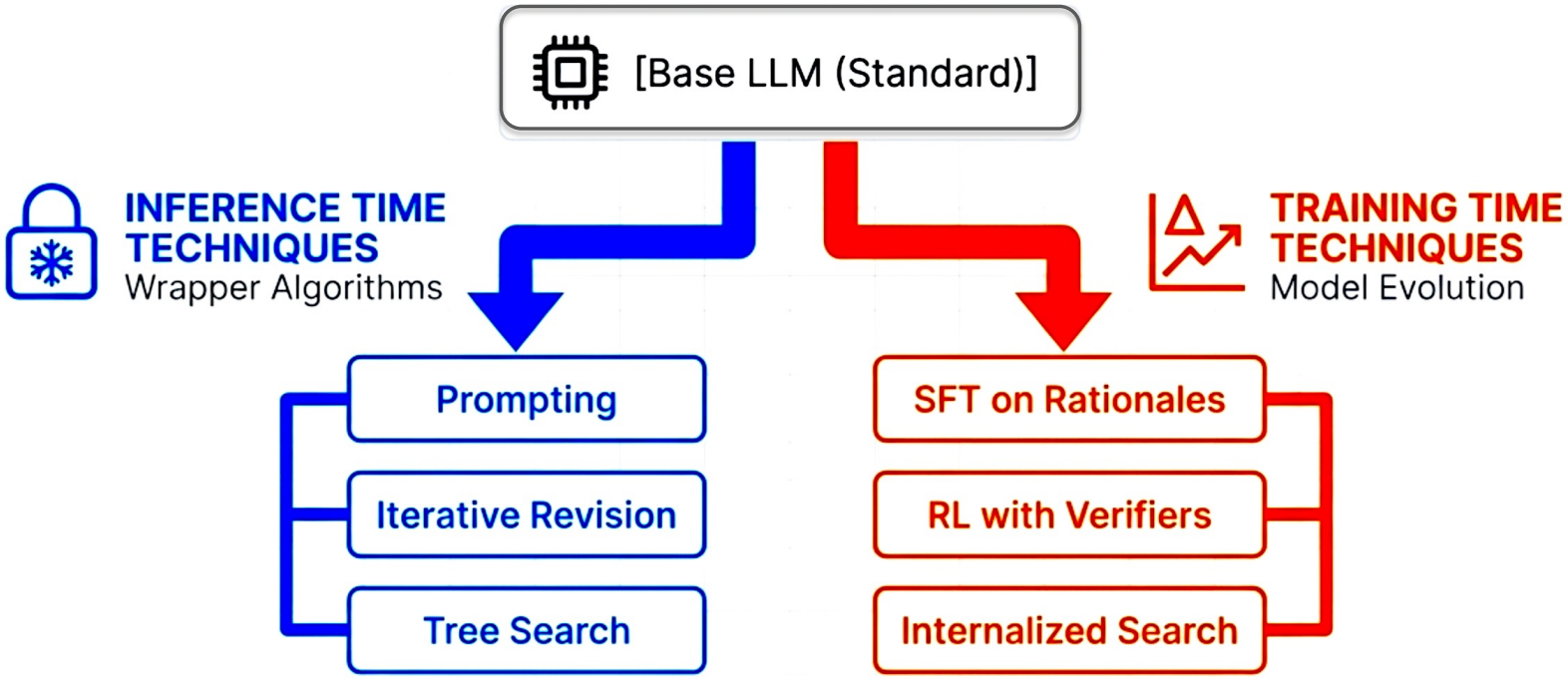
Why compute costs are justified: Massive accuracy gains in STEM domains.



Case Study: The "Tuning Award" Test



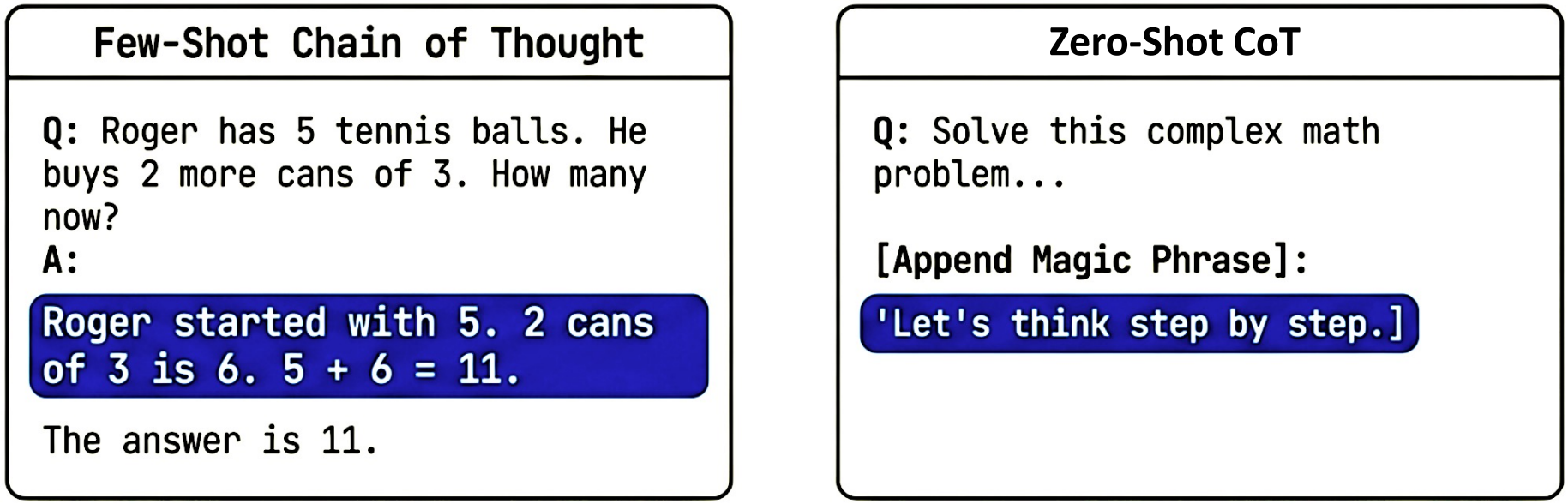
Two Pathways to Reasoning



Pathway 1: Eliciting logic at runtime from frozen models via prompting and scaffolding.

Pathway 2: Baking reasoning natively into model weights via self-correction and internal search.

Pathway 1: Advanced Prompting

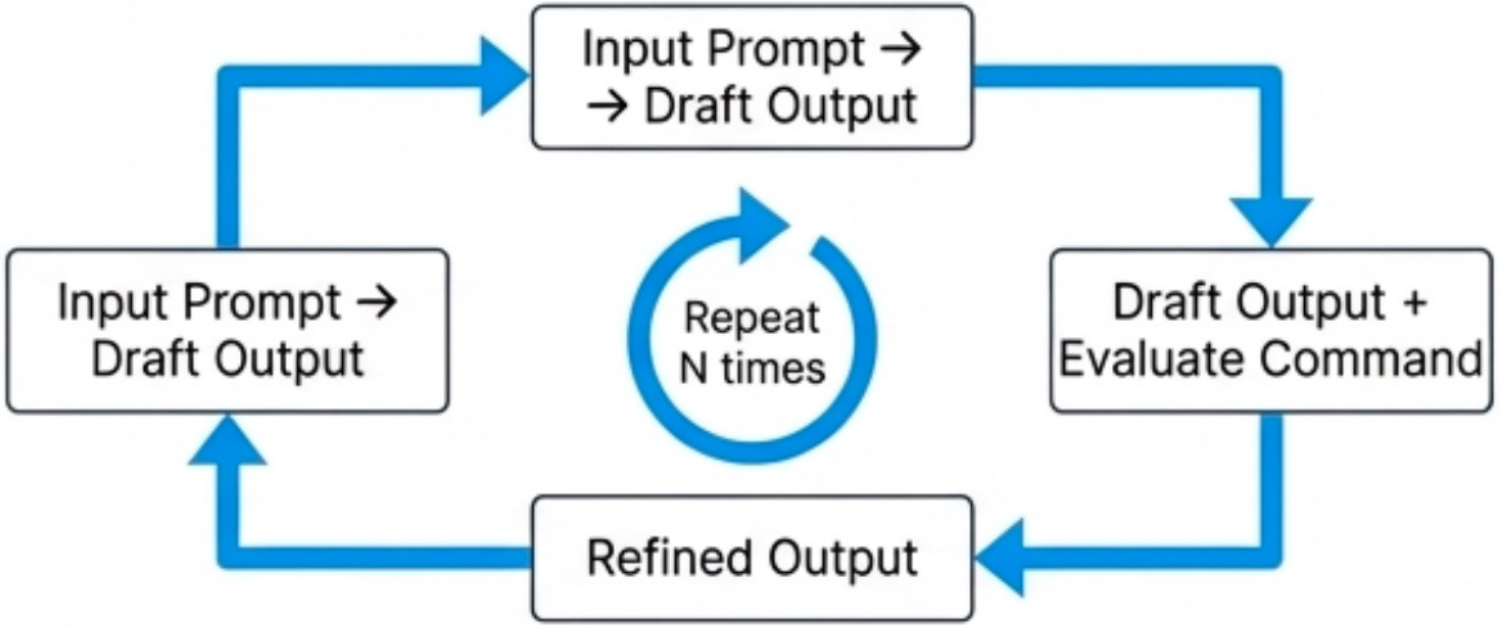


Forcing reasoning traces through input context.

Zero-Shot Chain of Thought (CoT). A single phrase triggers the generation of intermediate reasoning tokens.

Pathway 1: Sequential Revision

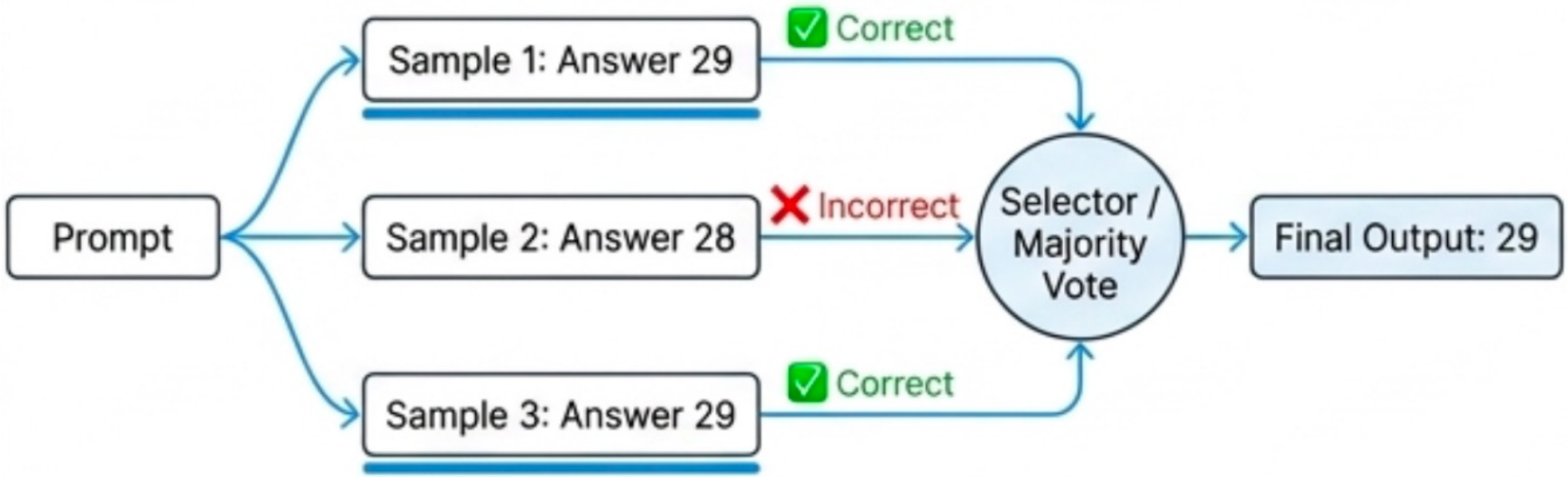
The model acts as its own editor, critiquing and refining drafts in a loop.



A self-critique loop where the model refines its draft N times

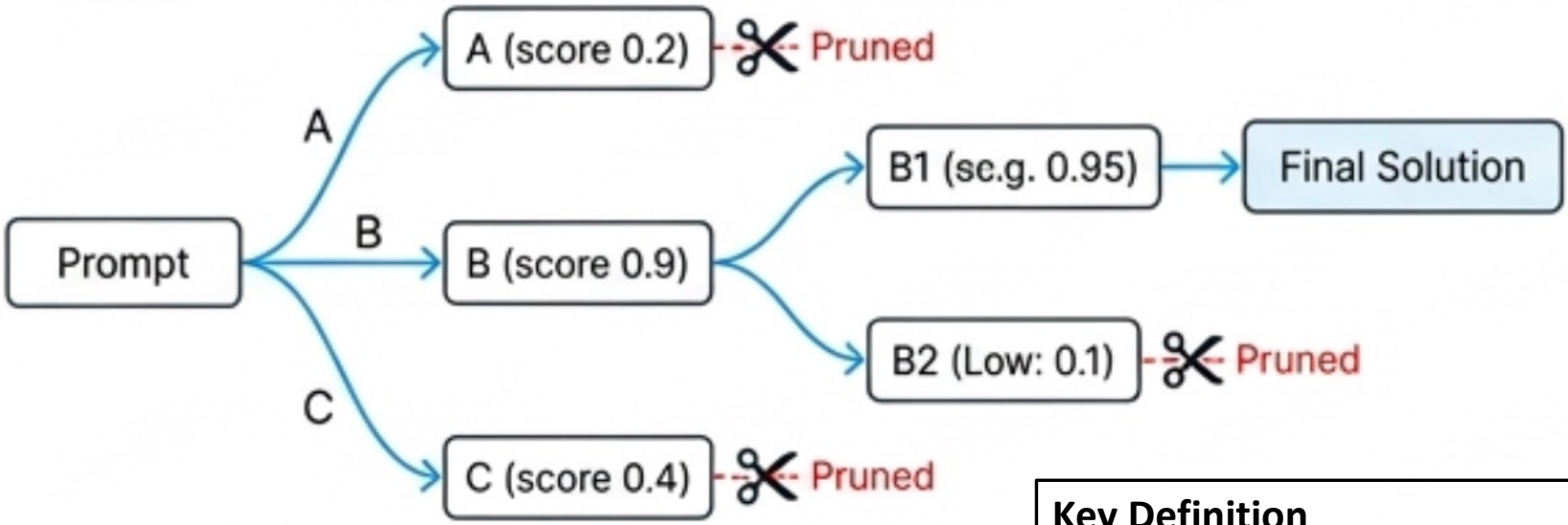
Pathway 1: Best of N Sampling

Parallel generation followed by a selection mechanism (Voting or Reward Model).



Pathway 1: Search Against a Verifier

Using Tree Search (MCTS/Beam) to explore solution spaces. The **Verifier** prunes bad logic branches



Key Definition

The Verifier: A separate model trained specifically to score partial solutions.

Algorithms: Beam Search, Lookahead Search, Monte Carlo Tree Search.

Pathway 2: Internalizing Thought (STaR)

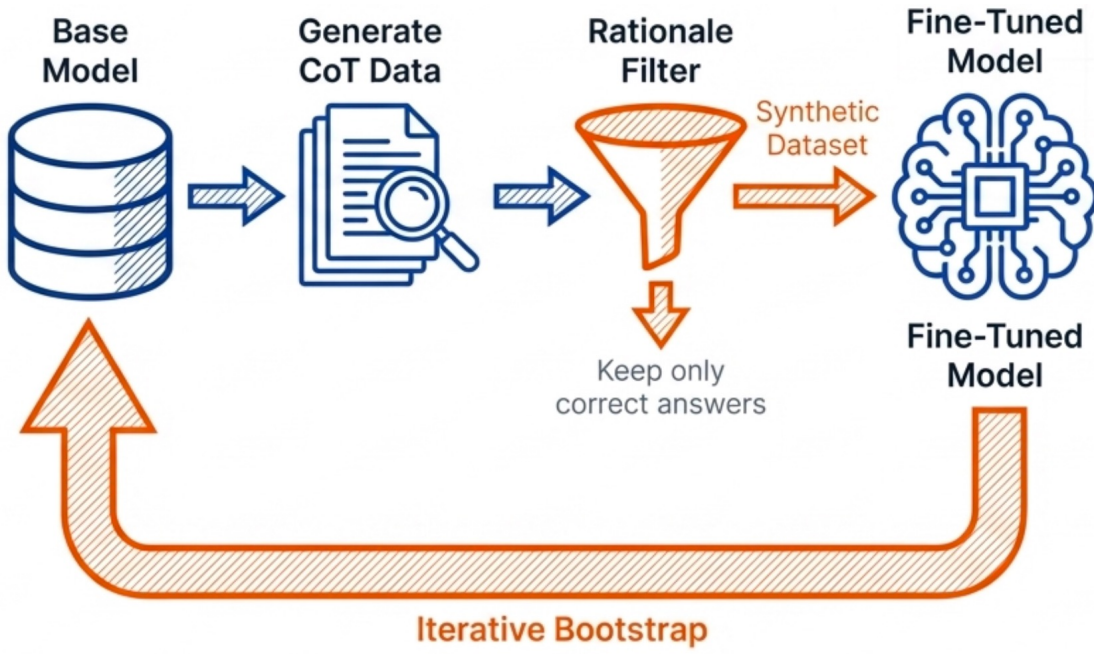
Supervised Fine-Tuning via Self-Taught Reasoners

Step 1. Generation: Base model creates Chain-of-Thought rationales.

Step 2. Verification: Filter retains only paths leading to the correct final answer.

Step 3. Fine-Tuning: Model is trained on this filtered synthetic dataset.

Step 4. Iteration: The smarter model generates even better rationales, repeating the loop natively.

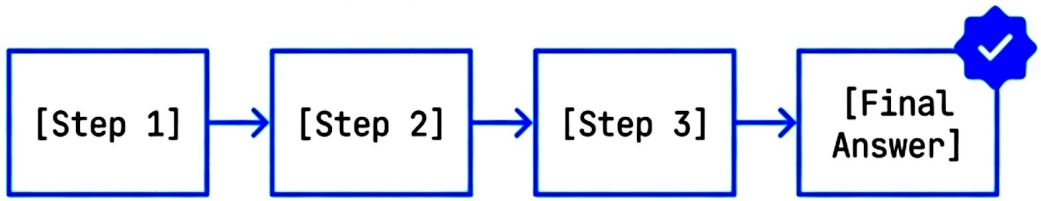


Pathway 2: Reinforcement Learning: Judging the Process

Outcome-Supervised / ORM

ORM (Outcome Reward Model):

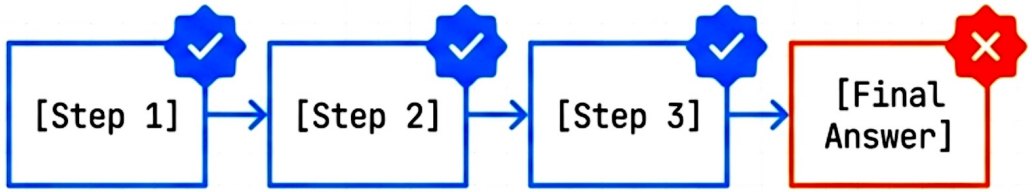
- Mechanism: **Grades only the final answer** (Sparse Reward).
- Flaw: Suffers from 'reward hacking' (guessing without valid logic).



Process-Supervised / PRM

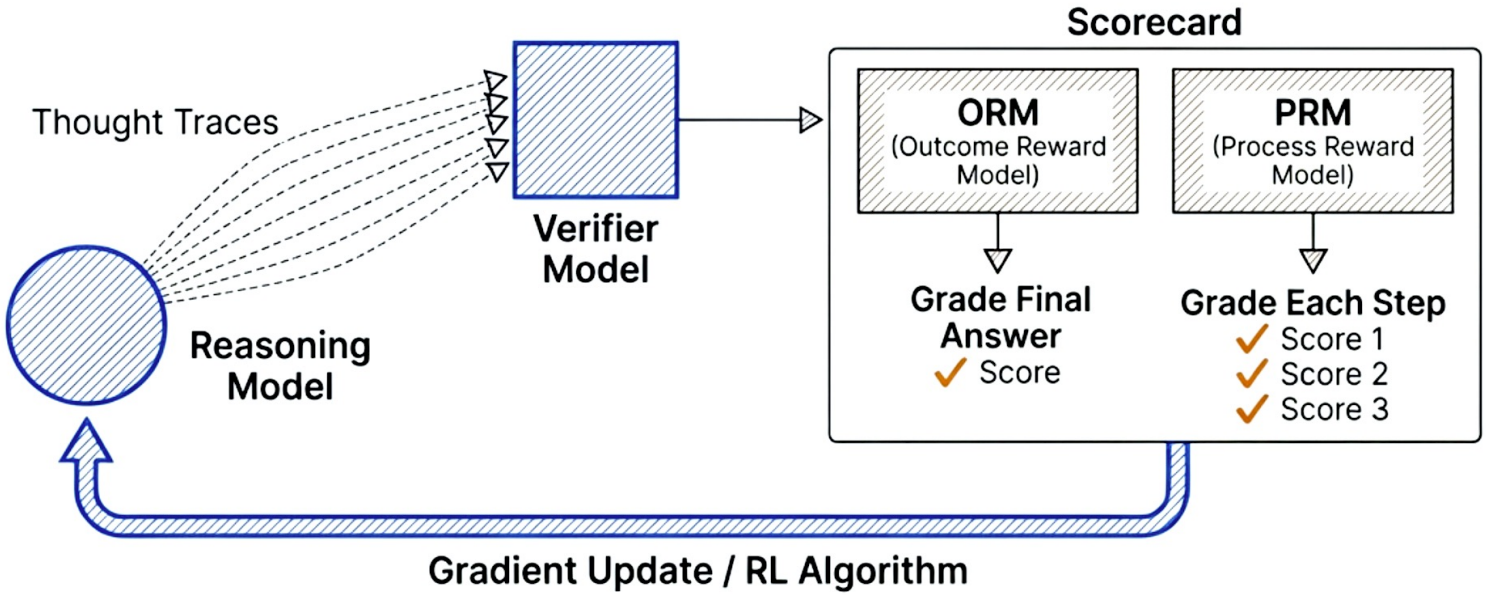
PRM (Process Reward Model):

- Mechanism: **Grades every intermediate step** (Dense Reward)
- Advantage: Penalizes early logical flaws. Teaches robust logical frameworks, not just answer memorization.



Pathway 2: Reinforcement Learning with Verifiers

Optimizing for the Process, not just the Outcome.



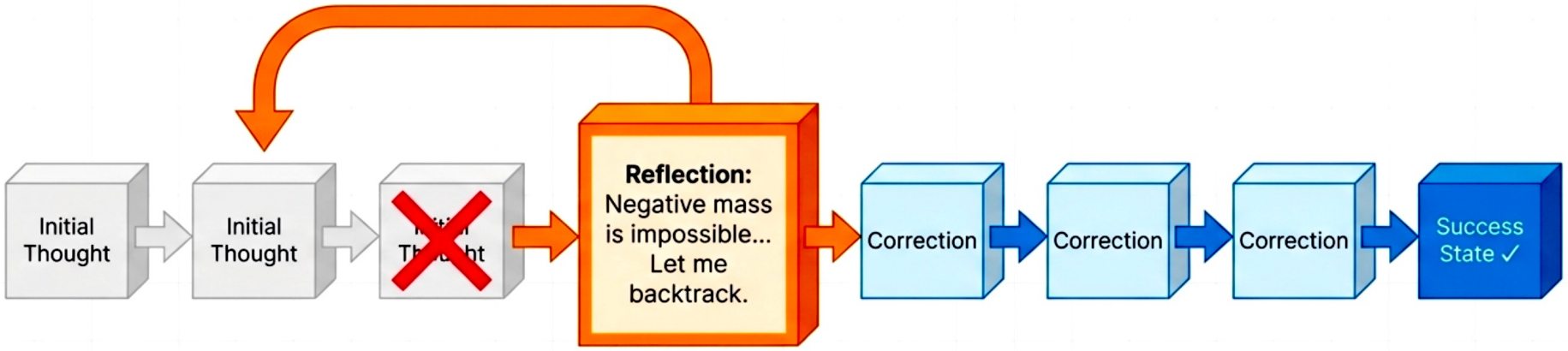
Process Supervision (PRM) prevents "hallucinated reasoning" by rewarding correct logic steps, not just the final answer.

Pathway 2: Latent Backtracking & Self-Correction

Treating reasoning as an internal search problem within the token stream.

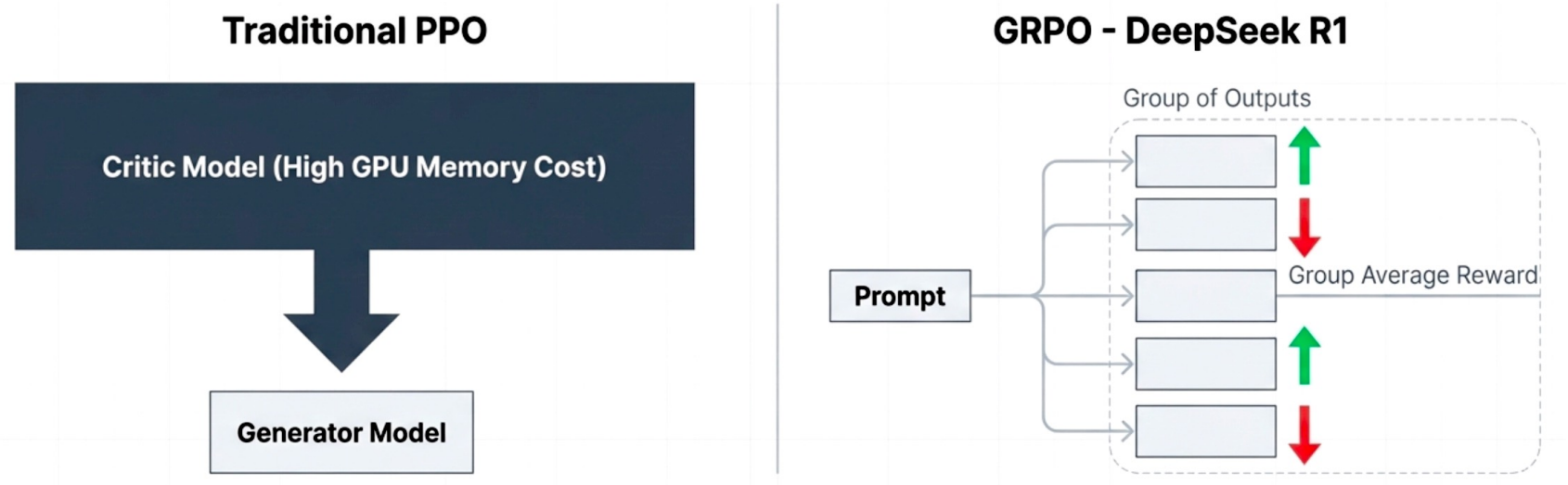
Mechanism:

1. Recognize high-entropy or low-confidence states.
2. Trigger a native "reflection" token.
3. Backtrack and correct mistakes internally without external API calls.



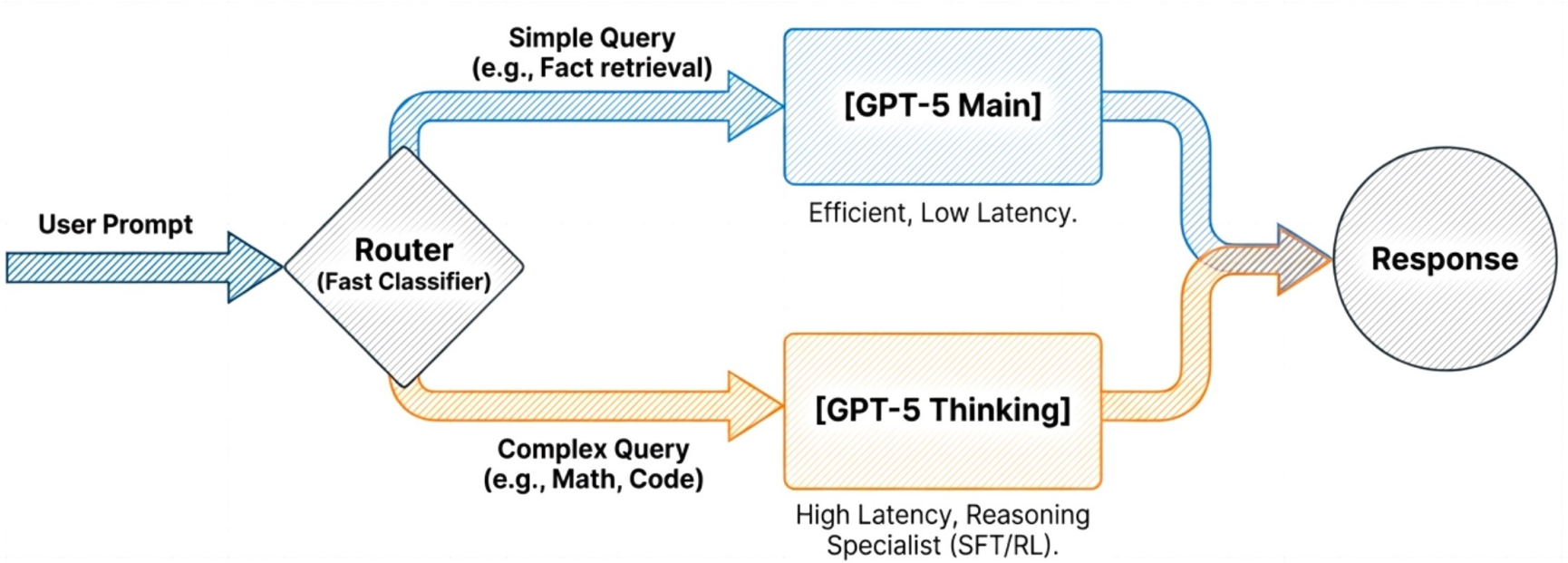
Pathway 2: The GRPO Breakthrough (DeepSeek-R1)

- **What it is:** Group Relative Policy Optimization—RL without the memory-heavy Critic Model.
- **How it works:** Samples a group of outputs, verifies the final outcome, and updates policy to favor outputs
- performing above the group average.
- **The Impact:** Eliminating the critic saves massive GPU memory. Enables scaling to incredibly long
- Chain-of-Thought contexts and allows the natural emergence of novel logic.

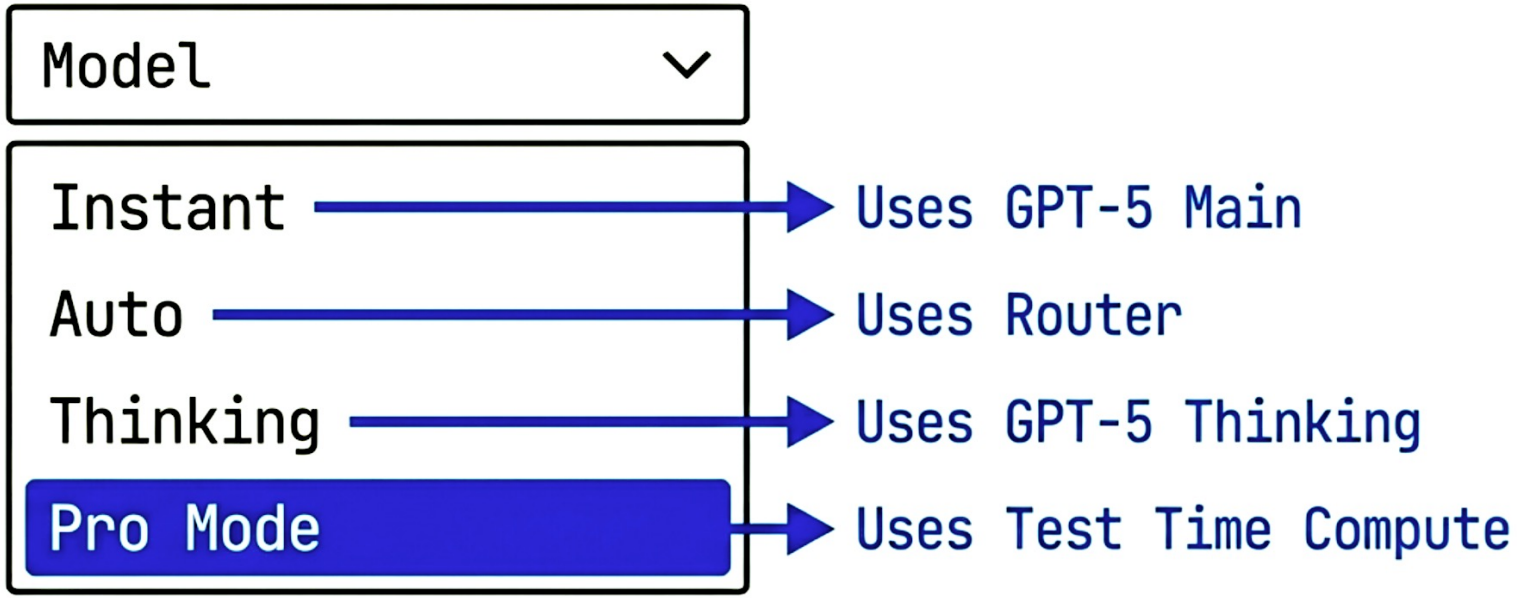


The "Dual-Brain" Frontier Architecture

- **The Adaptive Router:** Instantly classifies query complexity.
- **Fast Path**: Low latency, standard prediction for general tasks.
- **Heavy / Pro Path:** Stacked Monte Carlo, multi-agent swarms, and high-compute native reasoning for complex logic.
- **UX Result:** Instant response by default, with dynamic compute allocation based on task difficulty.



The User Experience of Architecture



The Mechanics of Modern Reasoning (2026)

Phase	Architectural Techniques	Core Outcome
Inference-Time (External)	Chain-of-Thought (CoT), Adaptive Effort, Best-of-N, Tree Search, Multi-Agent Routers.	Better algorithmic extraction from existing, frozen models.
Training-Time (Internal)	SFT / STaR, RL-PRM (Process Rewards), Meta-CoT (Backtracking), Group Relative Policy Optimization (GRPO).	Native reasoning, spontaneous self-correction, and agentic reliability encoded directly into model weights.

The Future of AI

- **Standard LLMs:** Predict the next word.
- **Reasoning LLMs:** Predict the next thought.
- We have moved from **'Fast Thinking'** pattern matching to **'Slow Thinking'** deliberate reasoning.
- The value of AI is shifting from **the size of the training set** to **the quality of the inference process.**