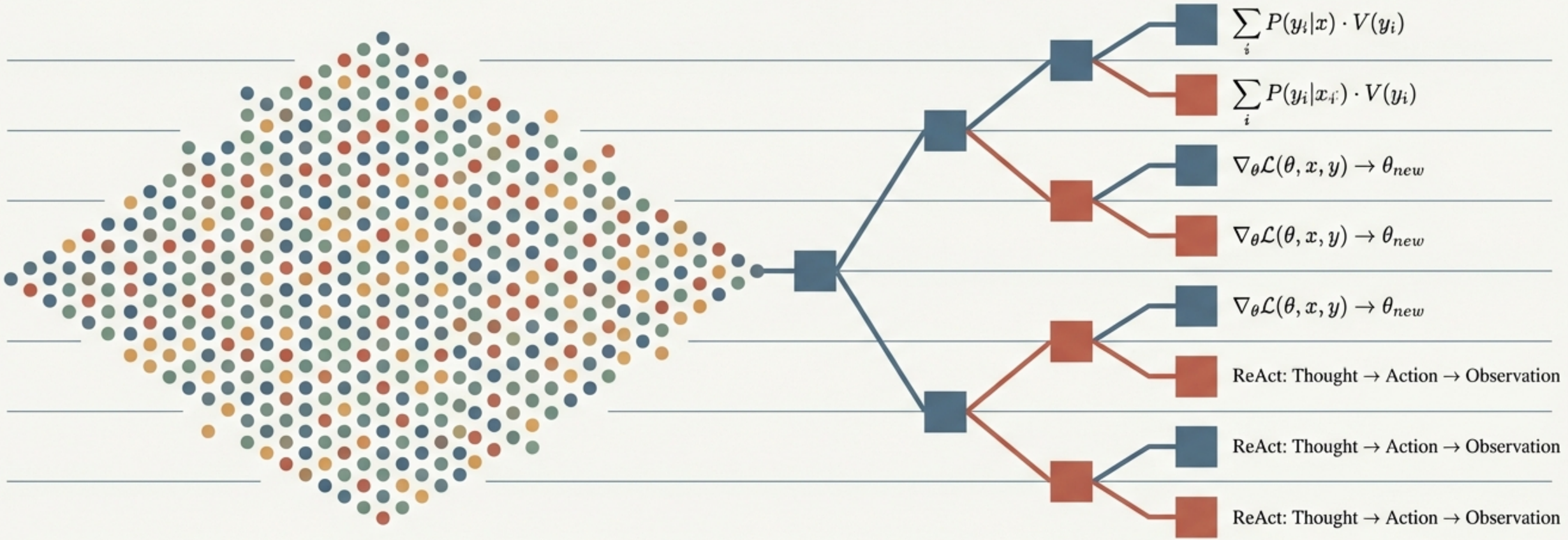
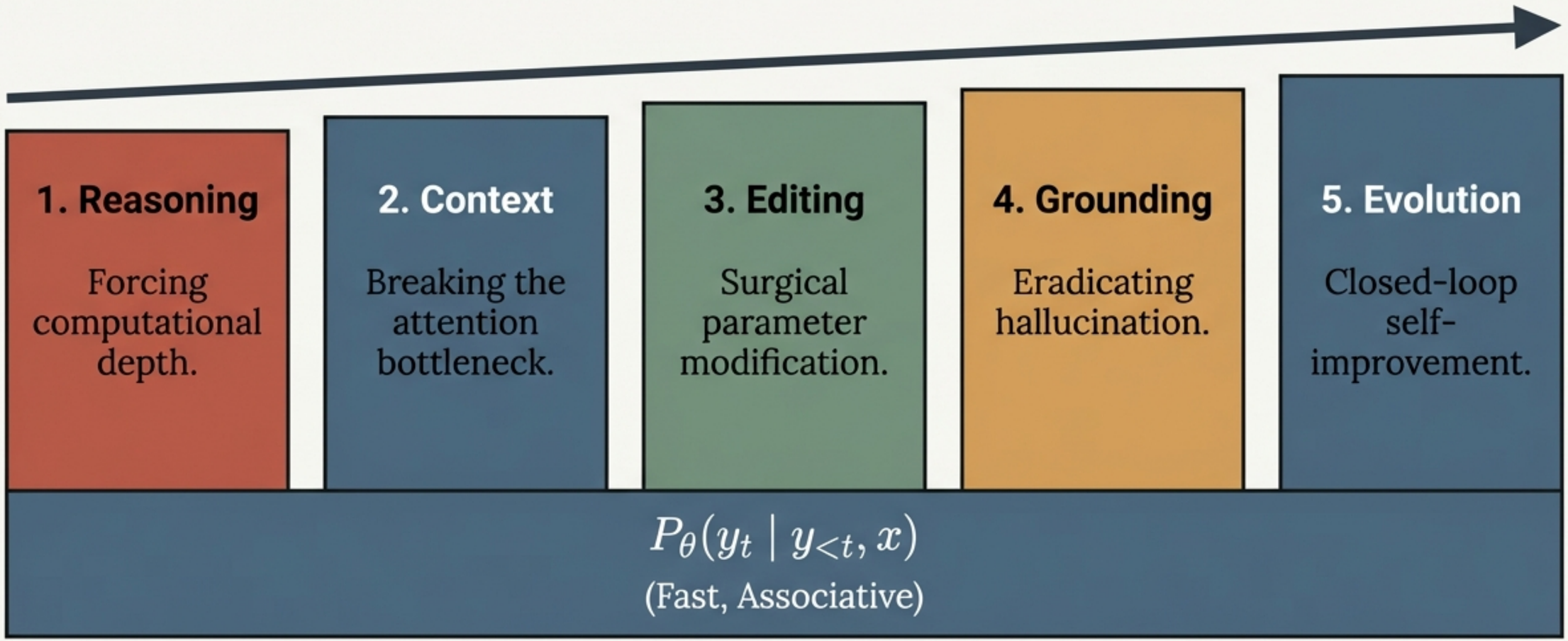
Engineering System 2 in Large Languagse Models



A SOTA treatment of Reasoning, Context Window Extension, Parameter Editing, Factuality, and Autonomous Self-Evolution

The Frontier of System 2 Thinking

Autoregressive generation is fundamentally a System 1 process. Engineering an AGI-trajectory model requires forcing deliberate, sequential, and verifiable System 2 computation.



Breaking the O(1) Reasoning Bound with Chain-of-Thought

$$P_\theta(y \mid x) = \sum_r P_\theta(r \mid x) \cdot P_\theta(y \mid x, r)$$

Intermediate scratchpad expands computational depth from O(1) to O(m).

FORMAL DEFINITION

Constant-depth transformers cannot solve TC^0-hard problems (e.g., unbounded integer arithmetic). CoT circumvents this by chaining serial forward passes.

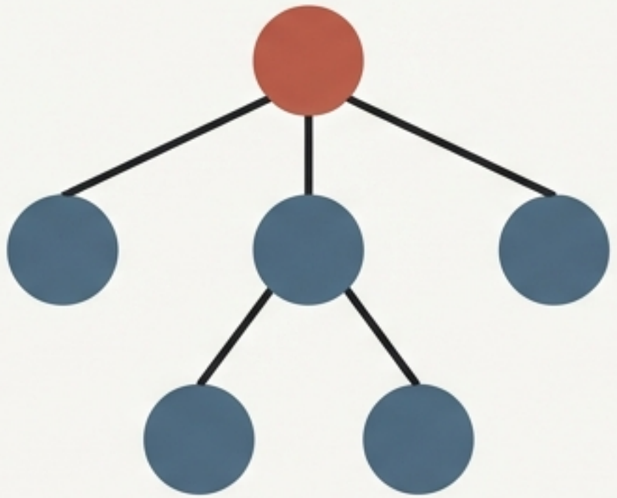
EFFECTIVE DEPTH:

$$L \times m$$

Computer Modern (where L is layers and m is reasoning steps)

Tree-Based Search and Step-Level Verification

Tree of Thought (MCTS)



$$\mathrm{UCB1}(s, s') = \underbrace{Q(s, s')}_{\text{Exploitation}} + \underbrace{c \cdot \sqrt{\ln \frac{N(s)}{N(s')}}}_{\text{Exploration}}$$

Reward Models: ORM vs PRM

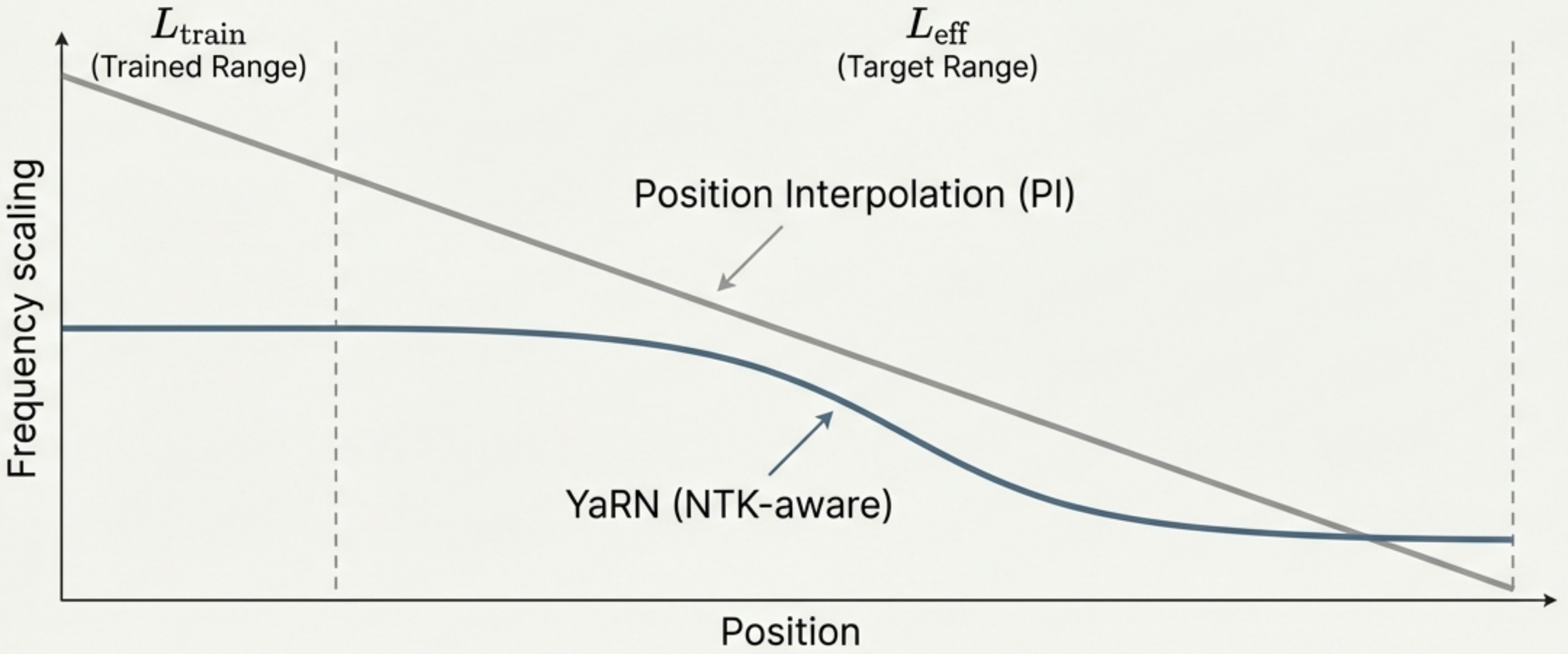
Outcome Reward Model (ORM)

Process Reward Model (PRM)

$$\mathrm{Var}[\nabla_\theta L_{\mathrm{PRM}}] \ll \mathrm{Var}[\nabla_\theta L_{\mathrm{ORM}}]$$

Dense, step-level supervision signals drastically reduce gradient variance and enable accurate credit assignment.

The Positional Dilemma in Extended Contexts



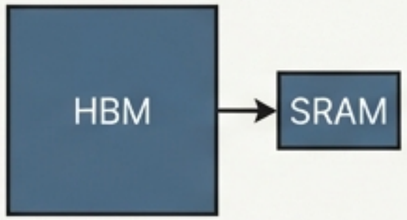
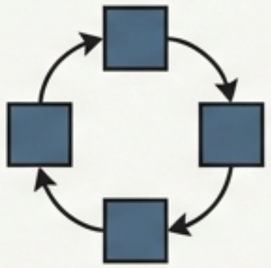
YaRN Attention Temperature Correction:

$$t = 0.1 \ln(\alpha) + 1$$

Goal:

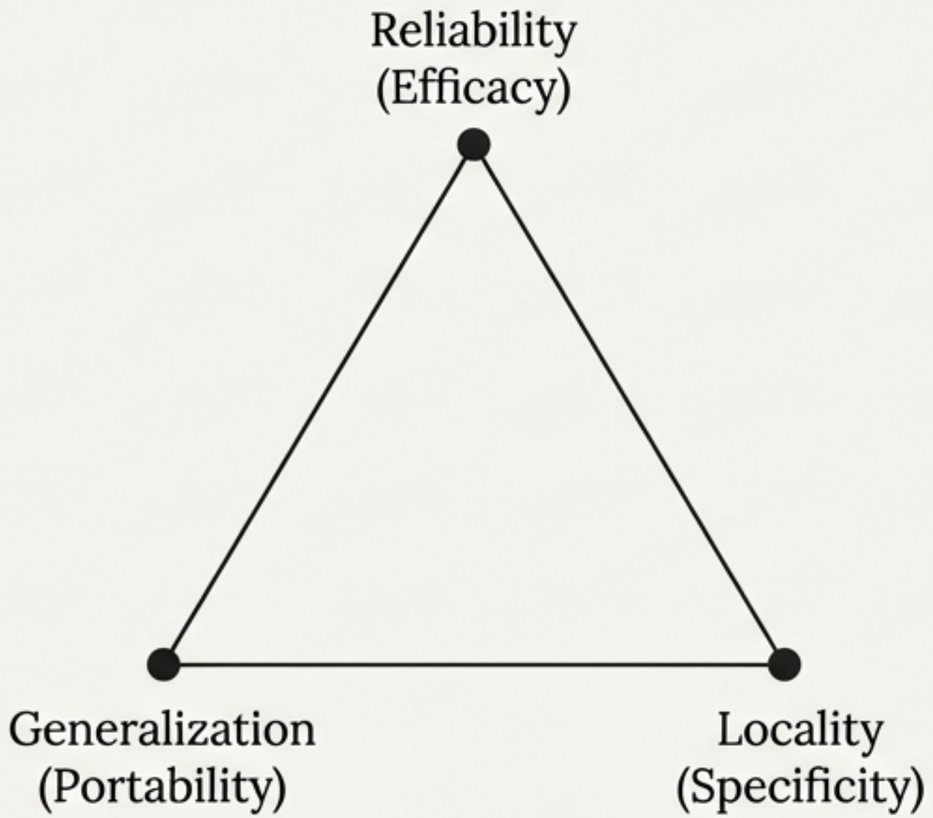
$$PPL(x_{1:L_{eff}}) \approx PPL(x_{1:L_{train}})$$

Cheating the Quadratic Physics of Attention

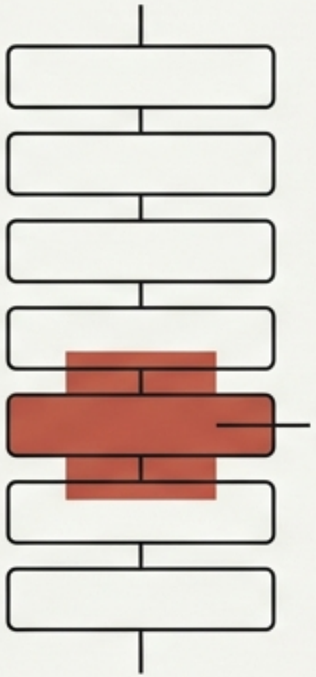
Hardware / Exact	Algorithmic / Exact
Flash Attention IO-aware tiling minimizing HBM reads to $\mathrm{O}(n^2 d \,/\, M)$.  **Ring Attention** Sequence parallelism via blockwise communication. 	**Multi-Scale Attention** Hierarchical token grouping.
Algorithmic / Approx	**Context Compression**
Linear Attention $\mathrm{O}(n \cdot d^2)$ kernel approximation.	**Attention Sinks** Preserving initial tokens. $\mathrm{KV\text{-}Cache} = \mathrm{Sink}(x_{1:\ell}) \oplus \mathrm{Window}(x_{n-w:n})$

Surgical Targeted Modification of Model Weights

The Desiderata of Editing



The Knowledge Localization Hypothesis



$$\text{MLP}^{(\ell)}(\boldsymbol{h}) = W_{proj} \cdot \sigma(W_{fc} \cdot \boldsymbol{h} + b_{fc}) + b_{proj}$$

Causal Tracing via Average Indirect Effect (AIE) reveals that factual recall is highly localized in early-site MLP layers, which function as key-value memories.

Rank-One Model Editing (ROME)

$$\hat{W} = W + \Lambda \cdot k_e^T C^{-1}$$

Rank-1 value shift
(inserts new fact v_e^{new})

Localized to key direction
(preserves key covariance C)

ROME treats MLP weights as a linear associative memory, surgically inserting new mappings ($k_e \rightarrow v_e^{new}$) while minimizing disruption to the rest of the network.

Deconstructing the Root Causes of Hallucination

Data-Level

Source-reference divergence and unfaithful abstractive pairs.

Architecture-Level

Attention dilution and the Softmax Bottleneck ($d \ll |V|$), forces approximation errors.

Decoding-Level

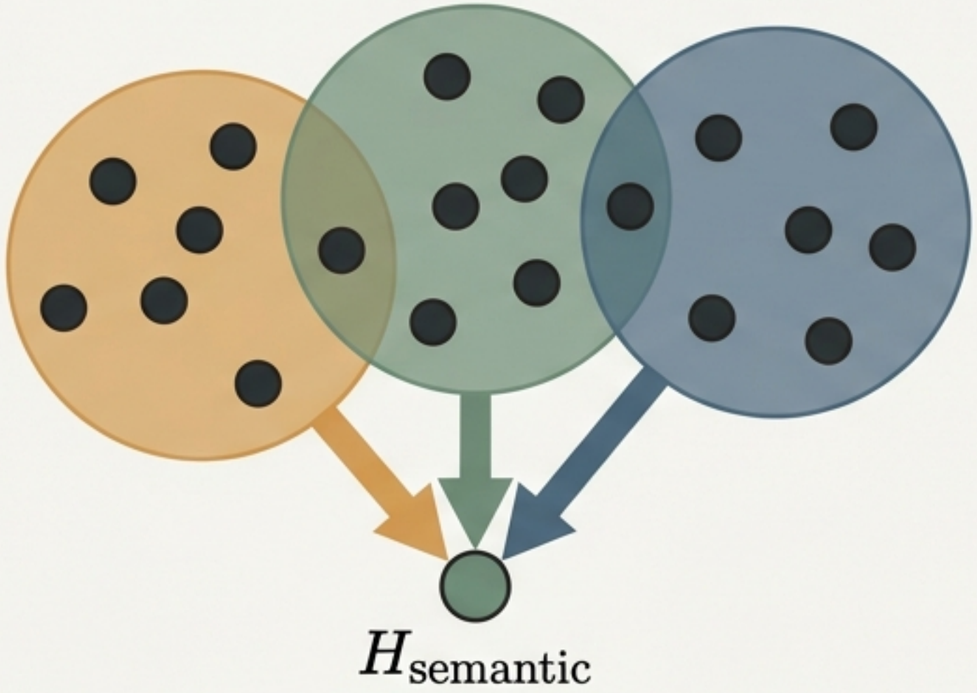
Exposure bias compounding prefix errors ($Error_T = O(\sum \epsilon_t)$), and high-entropy amplification.

Training-Level

RLHF-induced sycophancy. Reward models scoring confident lies higher than honest uncertainty.

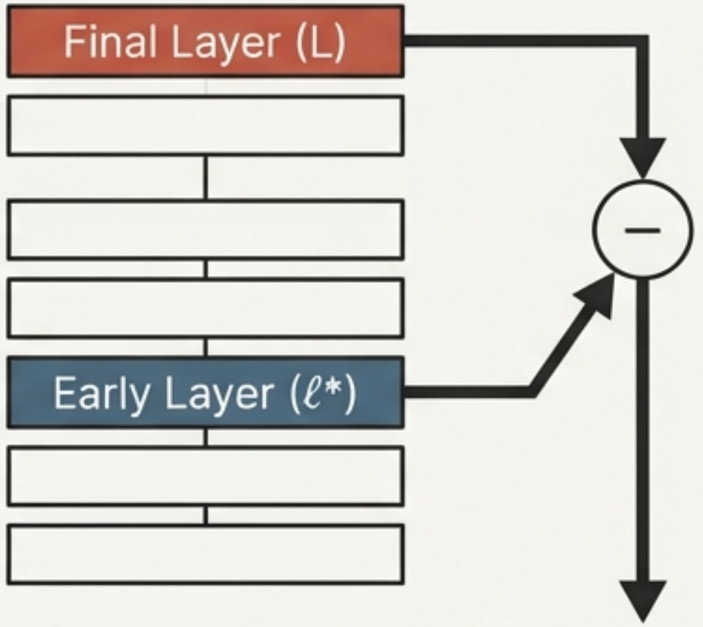
Real-Time Hallucination Detection and Mitigation

Semantic Entropy



H_{semantic} groups equivalent responses to find true confidence, avoiding surface-level lexical mismatches.

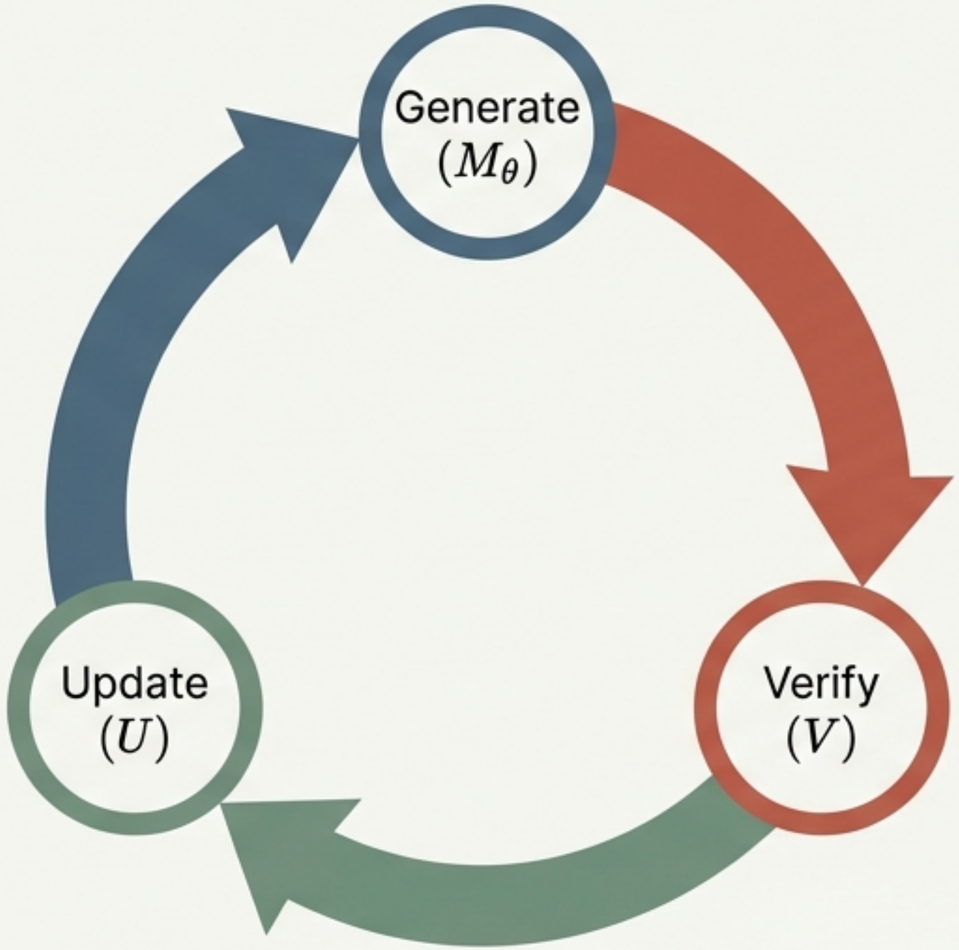
DoLa: Decoding by Contrasting Layers



$$P_{\text{DoLa}}(y_t) \propto \exp(\text{logit}^{(L)}(y_t) - \text{logit}^{(\ell^*)}(y_t))$$

Factual tokens change significantly between early and late layers. Generic hallucinations remain stable and cancel out.

The Closed-Loop Self-Evolution Paradigm



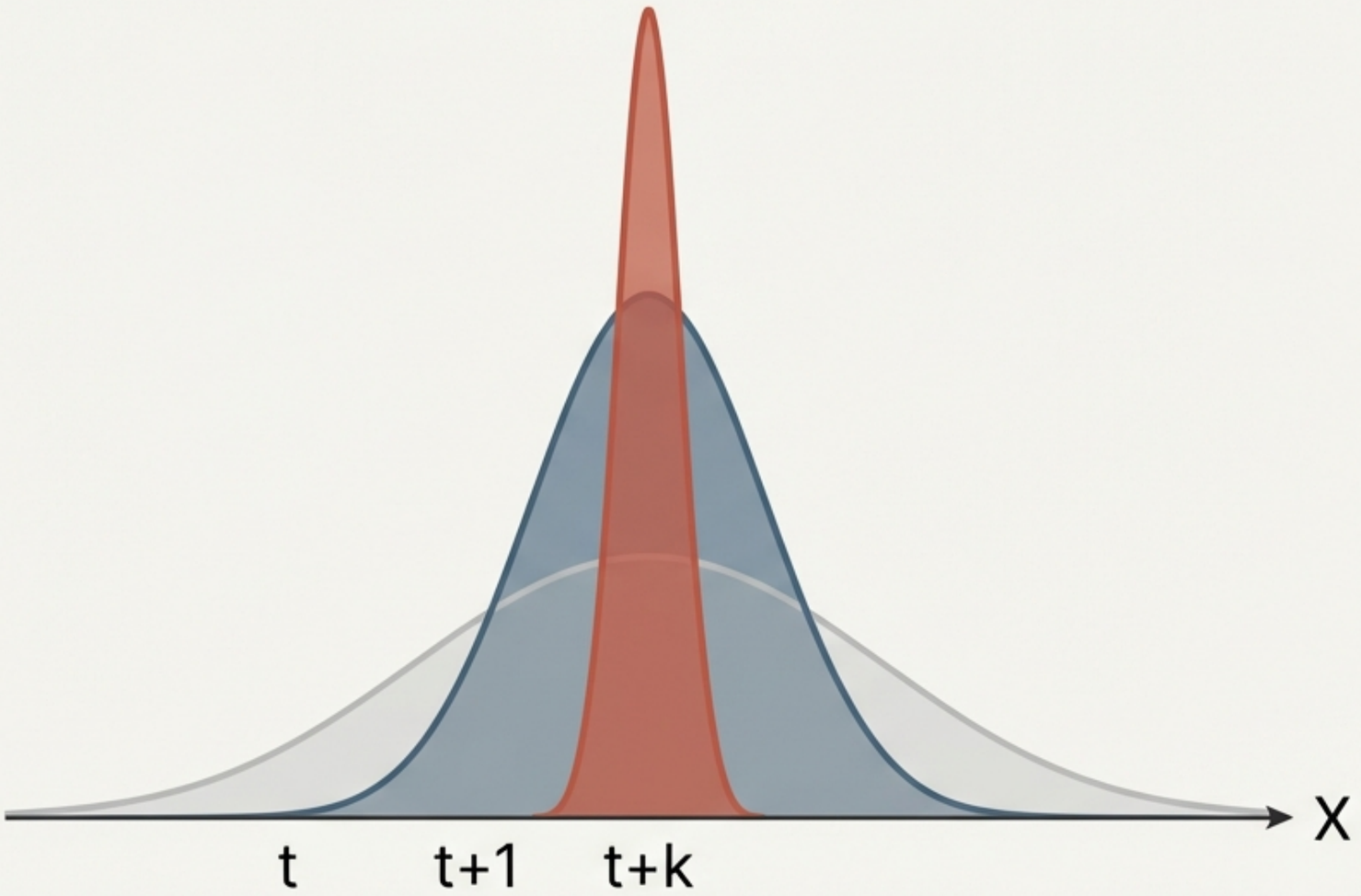
Reinforcement Learning from Verifiable Rewards (RLVR)

Automated verifiers (e.g., math, code logic) provide objective ground truth.

$$R(x, y) = 1 \text{ if } V(x,y) \text{ is CORRECT}$$

The Threat of Point-Mass Model Collapse

$$D_{KL}(P_{\theta_{t+k}} \parallel P_{true}) \to \infty$$



Without perfect verification, training on model-generated data converges to a degenerate point mass.

Mitigation Checklist

- Rejection sampling
- Diversity bonuses: $R' = R + \lambda \cdot H(y)$
- Curriculum scheduling

The Grand Unification of Capability Expansion

$$\Delta_{\text{improve}} = \text{Verify}(M_\theta) - \text{Generate}(M_\theta)$$

This single asymmetry dictates the frontier of AI. If verification is easier than generation ($\Delta_{\text{improve}} > 0$), CoT self-corrects, PRMs provide clean signal, DoLa detects falsehoods, and self-evolution escapes model collapse.

State of the Art Landscape: Executive Reference

	Reasoning	Context	Editing	Hallucination	Evolution
Formal Goal	$O(m)$ depth	$L_{\text{eff}} \gg L_{\text{train}}$	$\theta' = \mathbf{E}(\theta, e)$	$s \in \text{Entailed}(W)$	closed-loop update
Key Breakthroughs	PRMs, ToT	YaRN, Flash Attention	ROME, MEMIT	DoLa, Semantic Entropy	RLVR, GRPO, SPIN
Core Math Mechanism	CoT expansion sum	NTK-aware scaling	Rank-one updates	Logit subtraction	Verification reward bounding