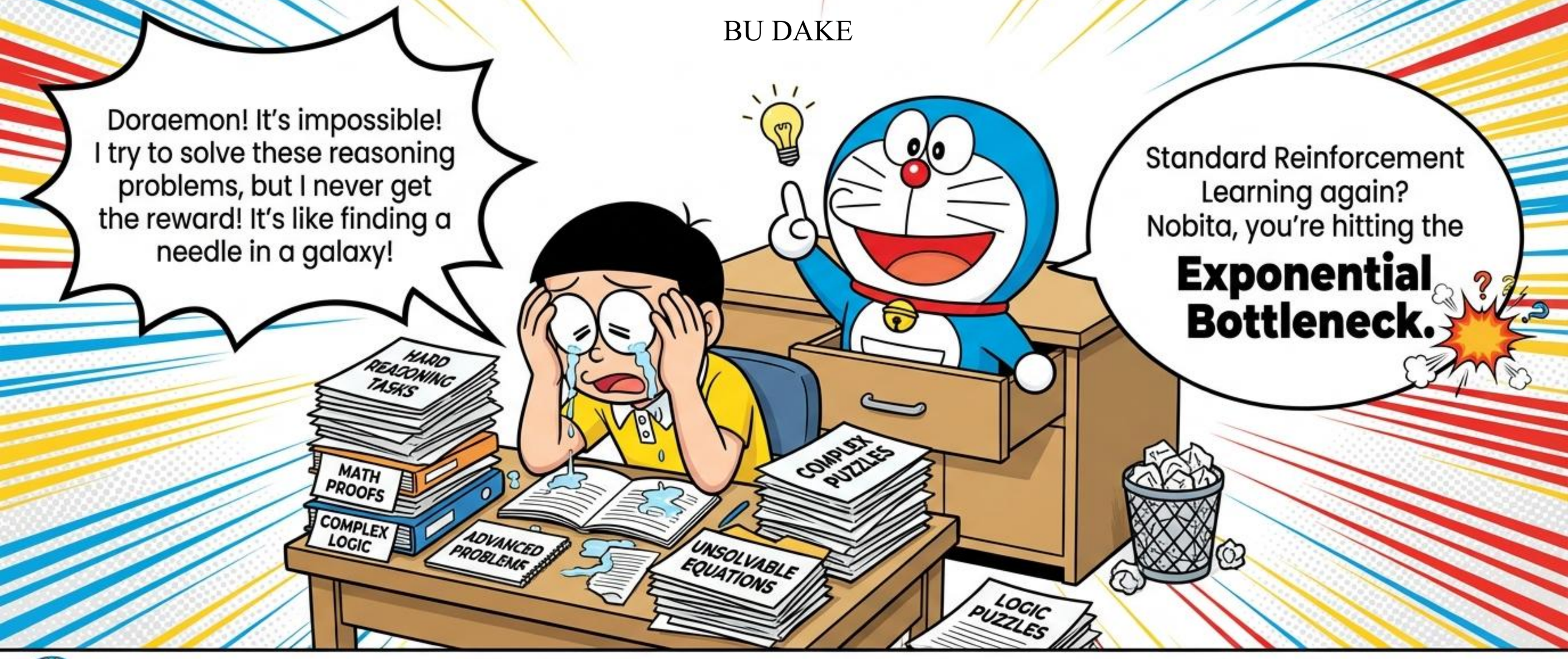
NOBITA'S REASONING REVOLUTION

Provable Sample Efficiency of Curriculum Post-Training

BU DAKE



The Problem: Direct Reinforcement Learning

Answer

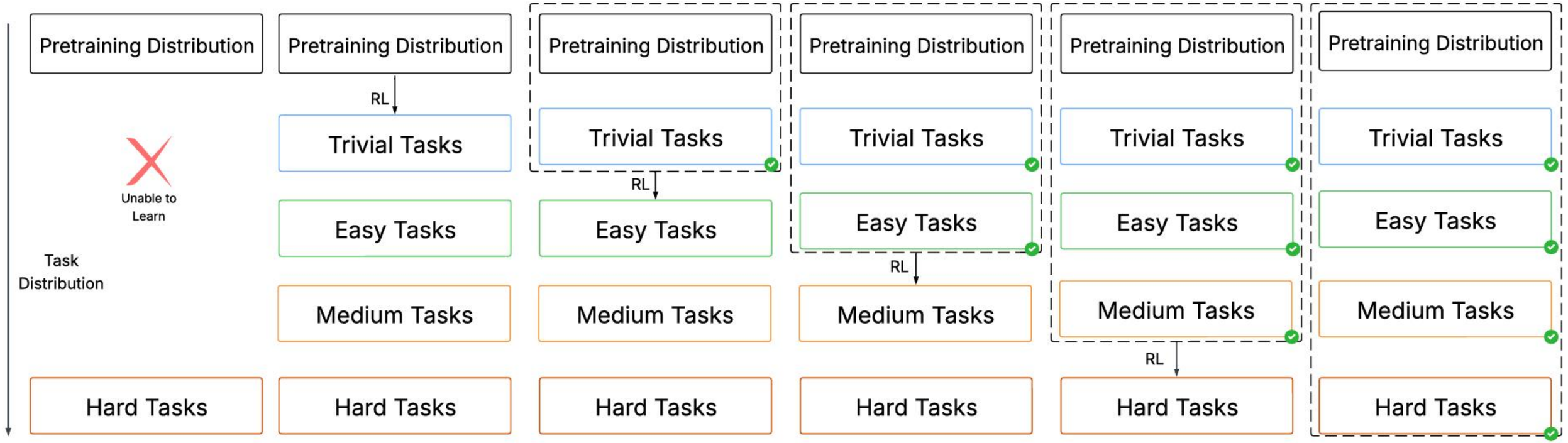
I keep guessing, but the 'Correct Chain-of-Thought' is too rare! I never get any feedback!

Depth K = Exponential Paths $(C^*)^K$

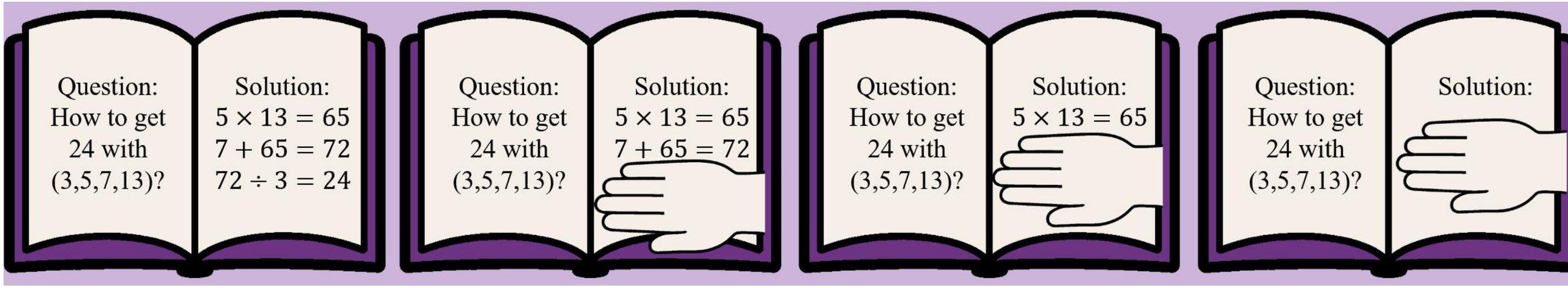
Answer

That is **Sample Complexity**. Under Direct Post-Training, if the correct path is rare, you need millions of tries. You'll be an old man before you learn this way!

Exponential Bottleneck: The rarity of correct reasoning trajectories under the base model drives the sampling cost.



[1] Parashar et al (2025) Curriculum Reinforcement Learning from Easy to Hard Tasks Improves LLM Reasoning



[2] Liu et al. (2025) UFT: Unifying Supervised and Reinforcement Fine-Tuning

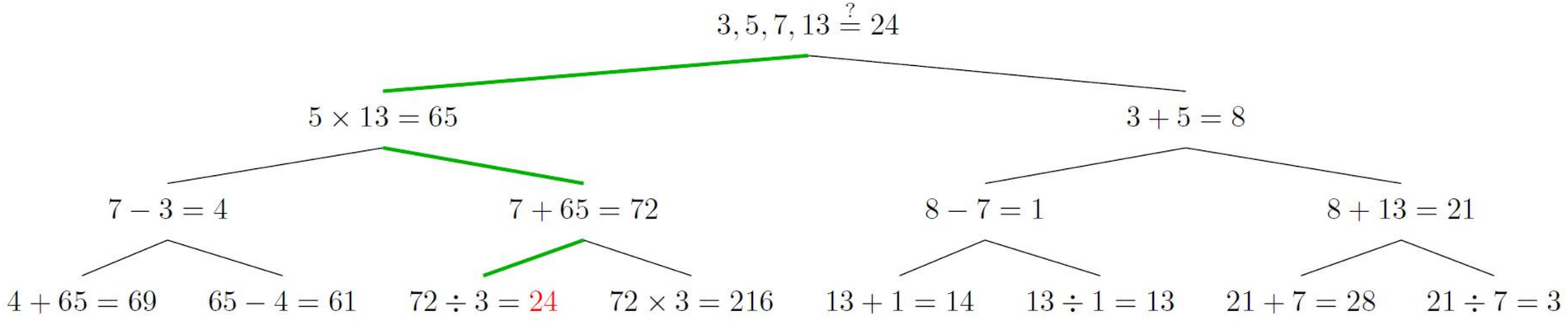


Figure 1. An illustration in Liu et al. (2025b) of the Chain-of-Thought for Countdown game, where the goal is to obtain 24 by applying basic arithmetic operations $(+, -, \times, \div)$ ($\Phi_l(\cdot, \cdot)$ in our Def. 2) between the current step's number (e.g., 13, 65 or 72) and some unused number (e.g., 5, 7 or 3) in $\{3, 5, 7, 13\}$, targeting 24 as the final outcome. Per (Parashar et al., 2025), the difficulty measure of Countdown is the number of arithmetic operations required to solve an instance.

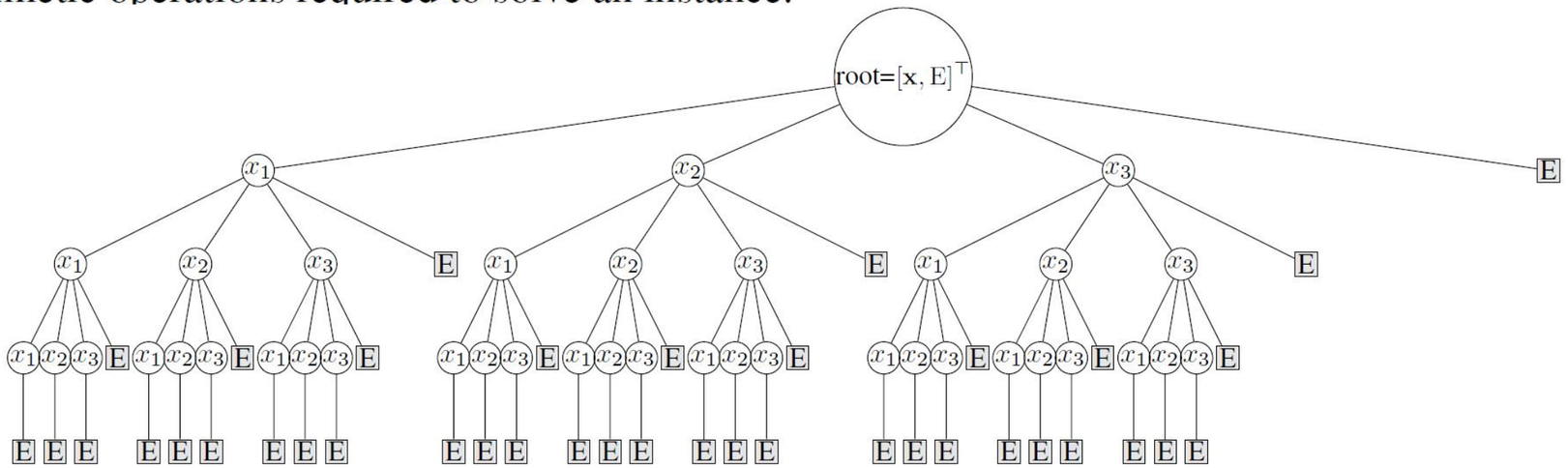
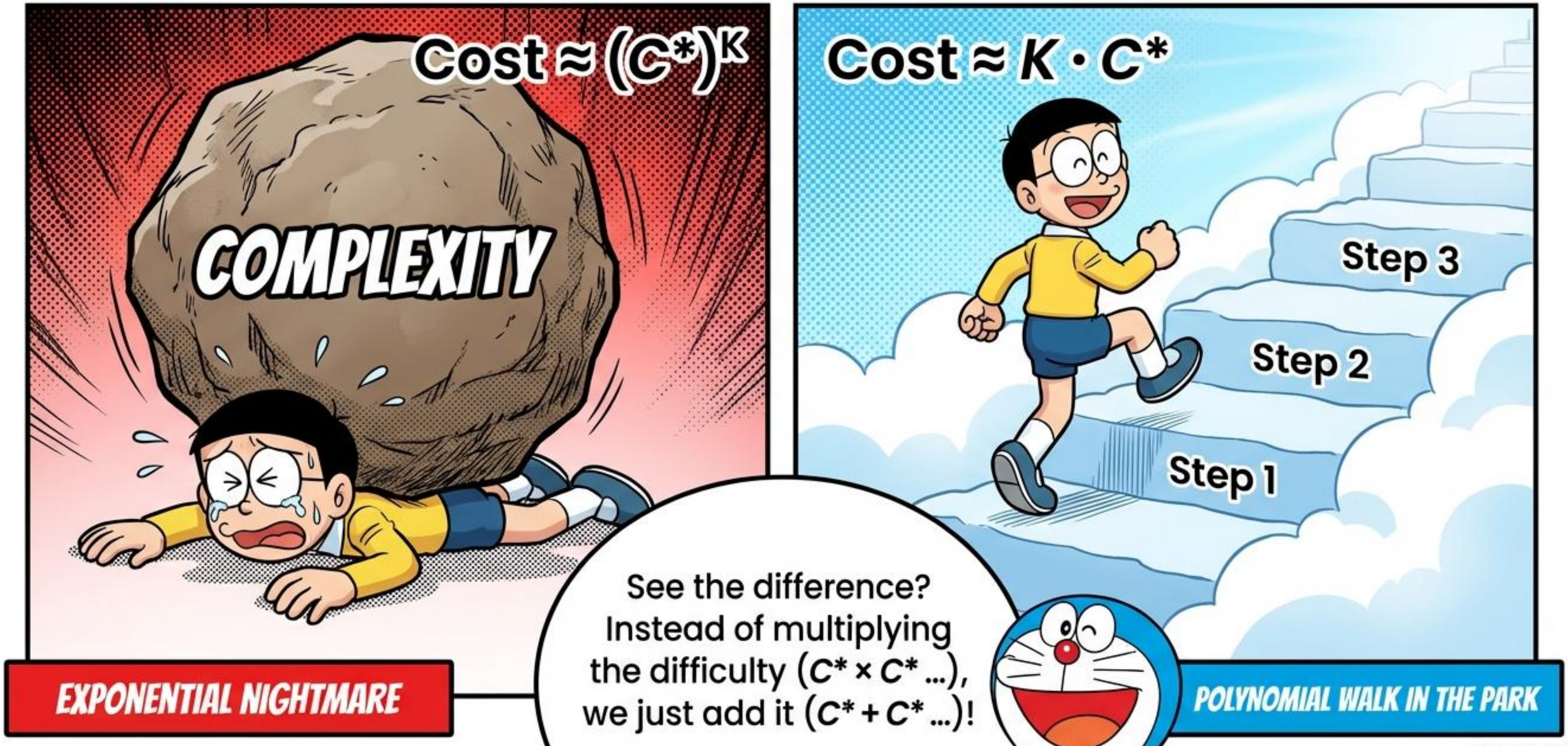


Figure 2. Reasoning tree for parity problems with $d = K = 3$ and input x_1, x_2, x_3, EOS. The nodes on the 2nd–4th levels denote hypotheses about which index is the current secret index, corresponding to x_{i_1}, x_{i_2}, and x_{i_3}, respectively. In our parity CoT class $f \in \mathcal{F}^{\text{parity}}_{\text{2S-ART}}$, each step actually consists of two actions: (i) choose the next secret index i_t; (ii) after the choice, apply an XOR over z_{t-1} and x_{i_t} as $z_t = \Phi_l(z_{t-1}, x_{i_t}) := z_{t-1} \oplus x_{i_t}$, as formalized in Eq. (6). For visual clarity, the tree only displays the index-selection branches and omits the explicit XOR updates. E denotes the EOS token. For parity tasks, there are illegal children $\notin \mathcal{I}_l$ for each parent that violate the "legal" criteria: non-repeating indices and strictly increasing order ($i_1 < i_2 < i_3 < \cdots$); thus, for any parity problem, CoTs with duplicate variables or decreasing index order are illegal.

EXPONENTIAL VS. POLYNOMIAL



Contribution: A High-level Sketch

➢ **High-level Intuition of the cost under direct and curriculum scenario.**

To sample π_k from π_{ref} with prob $1-\delta$, standard rejection sampling needs $\Theta(\|\frac{\pi_k^\star}{\pi_{\text{ref}}}\|_\infty \log(\delta^{-1}))$ trails.

Assumptions:

(A1: Realizability): $\|\frac{\pi_{k+1}^\star}{\pi_k^\star}\|_\infty, \|\frac{\pi_K^\star}{\pi_{\text{ref}}}\|_\infty < \infty$)

(A2: Suitable Step-wise Difficulty): $\|\frac{\pi_{k+1}^\star}{\pi_k^\star}\|_\infty \le \Theta(C^p), \forall k$ for some $p \ll K$ and constant $C > 1$

(A3: Intensive Direct Difficulty): $\|\frac{\pi_K^\star}{\pi_{\text{ref}}}\|_\infty \ge \Theta(C^{bK-c})$ for the same constant $C > 1$ $1 \le b \ll K, c \ll bK$

Conclusions: Direct Learning requires $\Theta(C^{bK-c})$ trials Curriculum Learning requires $\Theta(KC^p)$ trials

For any $\varepsilon > 0$, let $N_\varepsilon(\pi_2 \mid \pi_1)$ denote the ε-precision sample complexity[1] required to reach the optimal policy π^2 starting from a policy π_1. Consider learning the optimal policy $\pi^\star$ from a base policy π_{ref} via a step-wise curriculum consisting of K intermediate subtask policies

$$\pi_0^\star = \pi_{\text{ref}}, \quad \pi_1^\star, \quad \ldots, \quad \pi_K^\star := \pi^\star.$$

By definition, curriculum-style post-training is more sample-efficient than direct post-training if and only if the cumulative sample complexity of the intermediate stages is smaller than that of directly learning $\pi^\star$ from π_{ref}, namely,

$$\sum_{k\in[K]} N_\varepsilon(\pi_k^\star \mid \pi_{k-1}^\star) < N_\varepsilon(\pi^\star \mid \pi_{\text{ref}}).$$

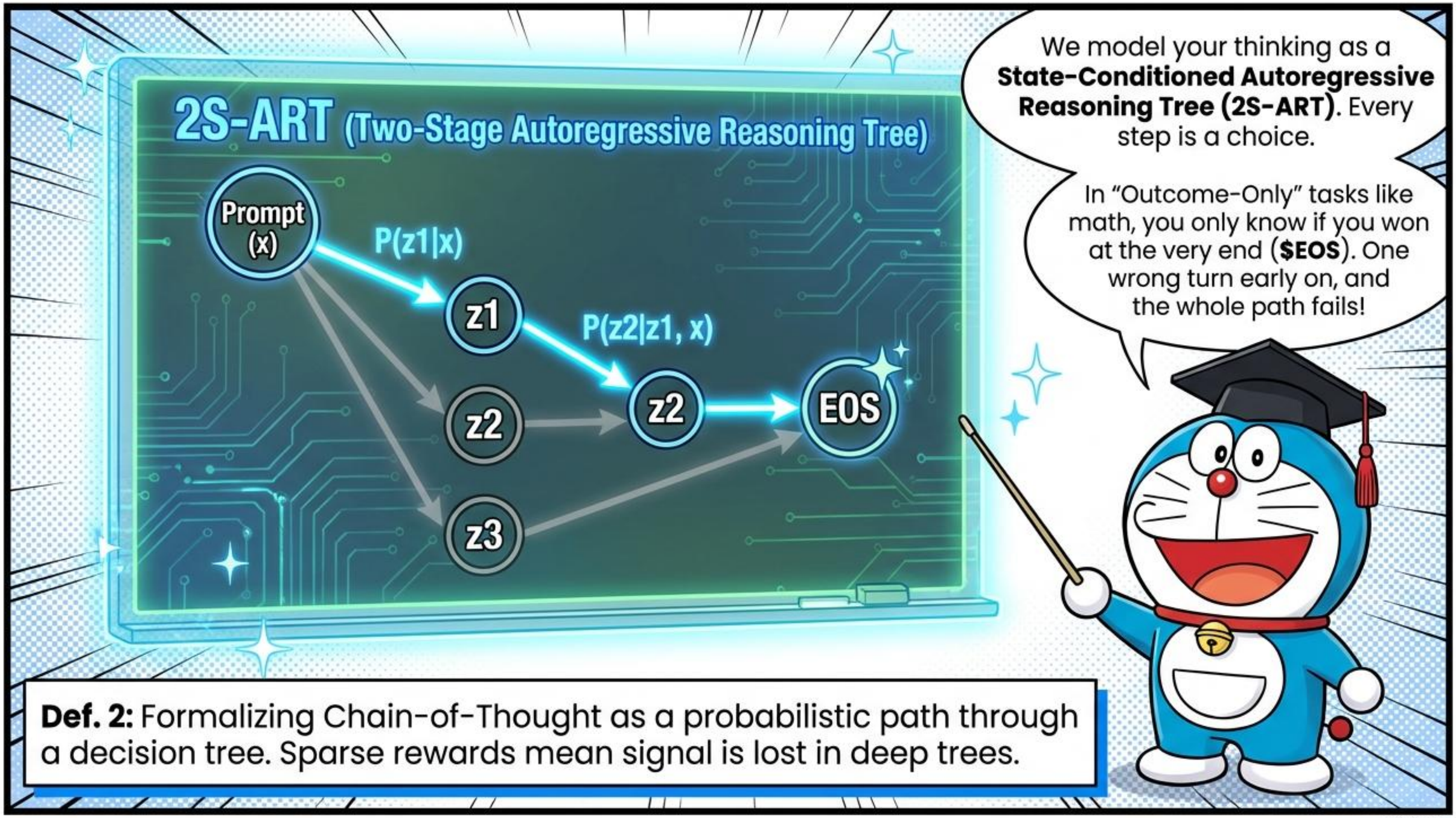
Corollary 1 (Sufficient Condition for Curriculum Exponential Improvement). *For any $\varepsilon > 0$, consider a base policy $\pi_{\text{ref}} = \pi_0^\star$ and a target optimal policy $\pi^\star \neq \pi_{\text{ref}}$. Suppose there exists a curriculum of K subtask optimal policies $\pi_1^\star, \ldots, \pi_K^\star := \pi^\star$ such that*

$$N_\varepsilon(\pi_k^\star \mid \pi_{k-1}^\star) = \Theta\big(\sqrt[K]{N_\varepsilon(\pi^\star \mid \pi_{\text{ref}})}\big), \qquad (1)$$

for all $k \in [K]$. Then, the ratio between the sample complexity of direct estimation $N_\varepsilon^{\text{direct}}$ and that of step-wise curriculum estimation $N_\varepsilon^{\text{curriculum}}$ satisfies

$$\frac{N_\varepsilon^{\text{direct}}}{N_\varepsilon^{\text{curriculum}}} = \frac{N_\varepsilon(\pi^\star \mid \pi_{\text{ref}})}{\sum_{k\in[K]} N_\varepsilon(\pi_k^\star \mid \pi_{k-1}^\star)} = \Theta\Big(\frac{(C^\star)^K}{KC^\star}\Big),$$

where $C^\star := \sqrt[K]{N_\varepsilon(\pi^\star \mid \pi_{\text{ref}})} > 1$.

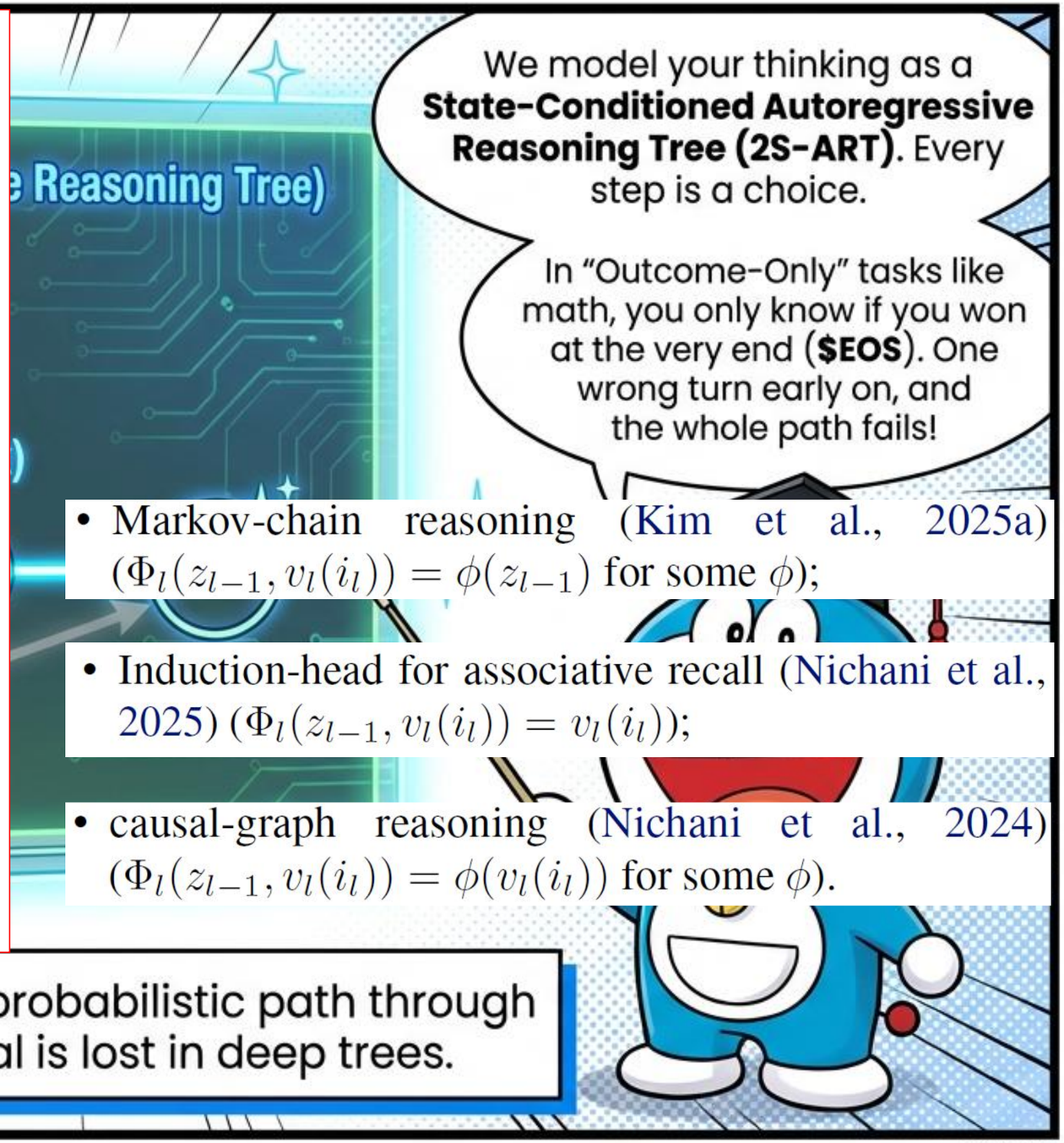

2S-ART (Two-Stage Autoregressive Reasoning Tree)
Prompt (x)
P(z1|x)
z1
z2
z3
P(z2|z1, x)
z2
EOS
We model your thinking as a **State-Conditioned Autoregressive Reasoning Tree (2S-ART)**. Every step is a choice.
In "Outcome-Only" tasks like math, you only know if you won at the very end (**$EOS**). One wrong turn early on, and the whole path fails!
Def. 2: Formalizing Chain-of-Thought as a probabilistic path through a decision tree. Sparse rewards mean signal is lost in deep trees.

Definition 1 (2-States Conditioned Autoregressive Reasoning Tree (2S-ART)). *A 2S-ART, denoted as* $\mathcal{F}_{\text{2S-ART}}(\{\Phi_l\}_{l\le L}, \{\mathcal{I}_l\}_{l\le L})$ *is a class of autoregressive tasks. Given input* $(\mathbf{x}, \text{EOS})$ *where* $\mathbf{x} = (x_1, \ldots, x_d) \in [K]^d$ *is followed by a* EOS *token, a task* $f_{S^k} \in \mathcal{F}_{\text{2S-ART}}$ *generates a CoT* $(z_1, \ldots, z_k, \text{EOS})$ *deterministically according to its* $(k+1)$*-size index path* $S^k = (i_1, \ldots, i_k, d+1)$ *where* $k < L$*: at step* l*,* f_{S^k} *chooses an index* $i_l \le d + l$ *from a legal set* $\mathcal{I}_l$ *with size* $\Theta(d)$*, reads the corresponding element* v_{i_l} *from the current sequence, and updates its reasoning state by a two-states map* $z_l = \Phi_l(z_{l-1}, v_{i_l})$*, with* $z_0 := \text{EOS}$*,* $i_{k+1} = d + 1$ *and* $\Phi_{k+1}(z_k, \text{EOS}) = \text{EOS}$.

The ***function value*** $f_{S^k}(\mathbf{x})$ *of task* f_{S^k} *is defined by the pre-*EOS *token, i.e.*

$$f_{S^k}(\mathbf{x}) := \Phi_k(z_{k-1}, v_{i_k}).$$

The associated curriculum subtask family is then characterized by the EOS*-ended prefix paths* $S^l = (i_1, \ldots, i_l, d{+}1)$ *with function values* $f_{S^l}(\mathbf{x}) := \Phi_k(z_{l-1}, v_{i_l}), l < k$.



Definition 1 (2-States Conditioned Autoregressive Reasoning Tree (2S-ART)). *A 2S-ART, denoted as* $\mathcal{F}_{\text{2S-ART}}(\{\Phi_l\}_{l\le L}, \{\mathcal{I}_l\}_{l\le L})$ *is a class of autoregressive tasks. Given input* $(\mathbf{x}, \text{EOS})$ *where* $\mathbf{x} = (x_1, \ldots, x_d) \in [K]^d$ *is followed by a* EOS *token, a task* $f_{S^k} \in \mathcal{F}_{\text{2S-ART}}$ *generates a CoT* $(z_1, \ldots, z_k, \text{EOS})$ *deterministically according to its* $(k+1)$*-size index path* $S^k = (i_1, \ldots, i_k, d+1)$ *where* $k < L$*: at step* l*,* f_{S^k} *chooses an index* $i_l \le d + l$ *from a legal set* $\mathcal{I}_l$ *with size* $\Theta(d)$*, reads the corresponding element* v_{i_l} *from the current sequence, and updates its reasoning state by a two-states map* $z_l = \Phi_l(z_{l-1}, v_{i_l})$*, with* $z_0 := \text{EOS}$*,* $i_{k+1} = d + 1$ *and* $\Phi_{k+1}(z_k, \text{EOS}) = \text{EOS}$.

The ***function value*** $f_{S^k}(\mathbf{x})$ *of task* f_{S^k} *is defined by the pre-*EOS *token, i.e.*

$$f_{S^k}(\mathbf{x}) := \Phi_k(z_{k-1}, v_{i_k}).$$

The associated curriculum subtask family is then characterized by the EOS*-ended prefix paths* $S^l = (i_1, \ldots, i_l, d{+}1)$ *with function values* $f_{S^l}(\mathbf{x}) := \Phi_k(z_{l-1}, v_{i_l}), l < k$.

Def. 2: Formalizing Chain-of-Thought as a prob
a decision tree. Sparse rewards mean signal is

Transformer ($\text{TF}(\cdot; \mathbf{W})$) **as Learning Model.** Let $\mathbf{U} = [\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\mu}_{\text{EOS}}] \in \mathbb{R}^{d_{\text{X}} \times (K+1)}$ denote the embedding vocabulary, where the k-th column corresponds to the embedding of token k, and $\boldsymbol{\mu}_{\text{EOS}}$ denotes the embedding of EOS. Following Li et al. (2025), we adopt mutually orthogonal positional encodings $\mathbf{P} := [\mathbf{p}_1, \ldots, \mathbf{p}_{d+1}]$ satisfying $\mathbf{p}_i \perp \mathbf{U}$. Token embeddings and positional encodings are concatenated as $\mathbf{E}[x_m] = [\boldsymbol{\mu}^{x_m\top}, \mathbf{p}_m{}^\top]^\top$, $\mathbf{E}[\text{EOS}] = [\boldsymbol{\mu}_{\text{EOS}}{}^\top, \mathbf{p}_{d+1}{}^\top]^\top$, and $\mathbf{E}[z_l] = [\boldsymbol{\mu}^{z_l\top}, \mathbf{p}_{i_l}{}^\top]^\top$, for all $m \in [d]$, $l \in [L{+}1]$, and $i_l \in [d{+}1]$. Given the embedded inputs $\mathbf{E}[x_1], \ldots, \mathbf{E}[x_d], \mathbf{E}[\text{EOS}] \in \mathbb{R}^{d_{\text{E}}}$. For notational convenience, we define $\mathbf{E}[z_{-d}] := \mathbf{E}[x_1], \ldots, \mathbf{E}[z_{-1}] := \mathbf{E}[x_d]$, and $\mathbf{E}[z_0] := \mathbf{E}[\text{EOS}]$. The transformer performs autoregressive next-token prediction via

$$\text{TF}(\mathbf{E}[z_{-d}], \ldots, \mathbf{E}[z_{l-1}]; \mathbf{W}) = \mathbf{E}[z_l], \tag{4}$$

where $\mathbf{W}$ denotes the model parameters. Throughout the reasoning process, the original input embeddings remain fixed, while the CoT states $\mathbf{E}[z_l]$ are generated autoregressively. Specifically, the forward computation at step l is

$$\begin{aligned}
\hat{\mathbf{p}}_l &= \sum_{j=-d}^{l-2} \mathbf{V}\Big(\mathbf{E}[z_j]\,\text{softmax}\big(\mathbf{E}[z_j]^\top \mathbf{K}^\top \mathbf{Q}\mathbf{E}[z_{l-1}]\big)\Big) \in \mathbb{R}^{d_{\text{X}}}, \\
\mathbf{p}_{i_l} &\sim \text{softmax}(\hat{\mathbf{p}}_l^\top \mathbf{P}/\beta), \\
\hat{\boldsymbol{\mu}}^{z_l} &= \text{FFN}_l(\boldsymbol{\mu}^{x_{i_l}}, \hat{\boldsymbol{\mu}}^{z_{l-1}}), \quad \mathbf{E}[z_l] = [\hat{\boldsymbol{\mu}}^{z_l}, \mathbf{p}_{i_l}]^\top \in \mathbb{R}^{d_{\text{E}}}.
\end{aligned} \tag{5}$$

Definition 1 (2-States Conditioned Autoregressive Reasoning Tree (2S-ART)). *A 2S-ART, denoted as* $\mathcal{F}_{\text{2S-ART}}(\{\Phi_l\}_{l\le L}, \{\mathcal{I}_l\}_{l\le L})$ *is a class of autoregressive tasks. Given input* $(\mathbf{x}, \text{EOS})$ *where* $\mathbf{x} = (x_1, \ldots, x_d) \in [K]^d$ *is followed by a* EOS *token, a task* $f_{S^k} \in \mathcal{F}_{\text{2S-ART}}$ *generates a CoT* $(z_1, \ldots, z_k, \text{EOS})$ *deterministically according to its* $(k+1)$*-size index path* $S^k = (i_1, \ldots, i_k, d+1)$ *where* $k < L$*: at step* l*,* f_{S^k} *chooses an index* $i_l \le d + l$ *from a legal set* $\mathcal{I}_l$ *with size* $\Theta(d)$*, reads the corresponding element* v_{i_l} *from the current sequence, and updates its reasoning state by a two-states map* $z_l = \Phi_l(z_{l-1}, v_{i_l})$*, with* $z_0 := \text{EOS}$*,* $i_{k+1} = d+1$ *and* $\Phi_{k+1}(z_k, \text{EOS}) = \text{EOS}$.

The ***function value*** $f_{S^k}(\mathbf{x})$ *of task* f_{S^k} *is defined by the pre-*EOS *token, i.e.*

$$f_{S^k}(\mathbf{x}) := \Phi_k(z_{k-1}, v_{i_k}).$$

The associated curriculum subtask family is then characterized by the EOS*-ended prefix paths* $S^l = (i_1, \ldots, i_l, d{+}1)$ *with function values* $f_{S^l}(\mathbf{x}) := \Phi_k(z_{l-1}, v_{i_l}), l < k$.

Def. 2: Formalizing Chain-of-Thought as a prob
a decision tree. Sparse rewards mean signal is

CoT and Curriculum Subtask Family for Parity. Per Def. 1, the resulting XOR-based CoT (Wen et al., 2025a; Abedsoltan et al., 2025) is:

$$z_1 = x_{i_1},\ z_2 = x_1 \oplus x_{i_2},\ \ldots,\ z_k = z_{k-1} \oplus x_{i_k} = f_S(\mathbf{x}), \quad (6)$$

where $i_1 < i_2 < \ldots < i_k$. In this case, the curriculum subtask family is the prefix-index family $\mathcal{F}_S = \{f_{S^1}, \ldots, f_{S^{k-1}}\}$ with $S^{k'} = (i_1, \ldots, i_{k'}, d+1)$ and $f_{S^{k'}}(\mathbf{x}) = z_{k'}$ for $k' \in [k]$; early termination by EOS follows Def. 1. In particular, the legal sets $\{\mathcal{I}_l\}$ is defined to satisfy $i_1 < i_2 < \cdots < i_k$.

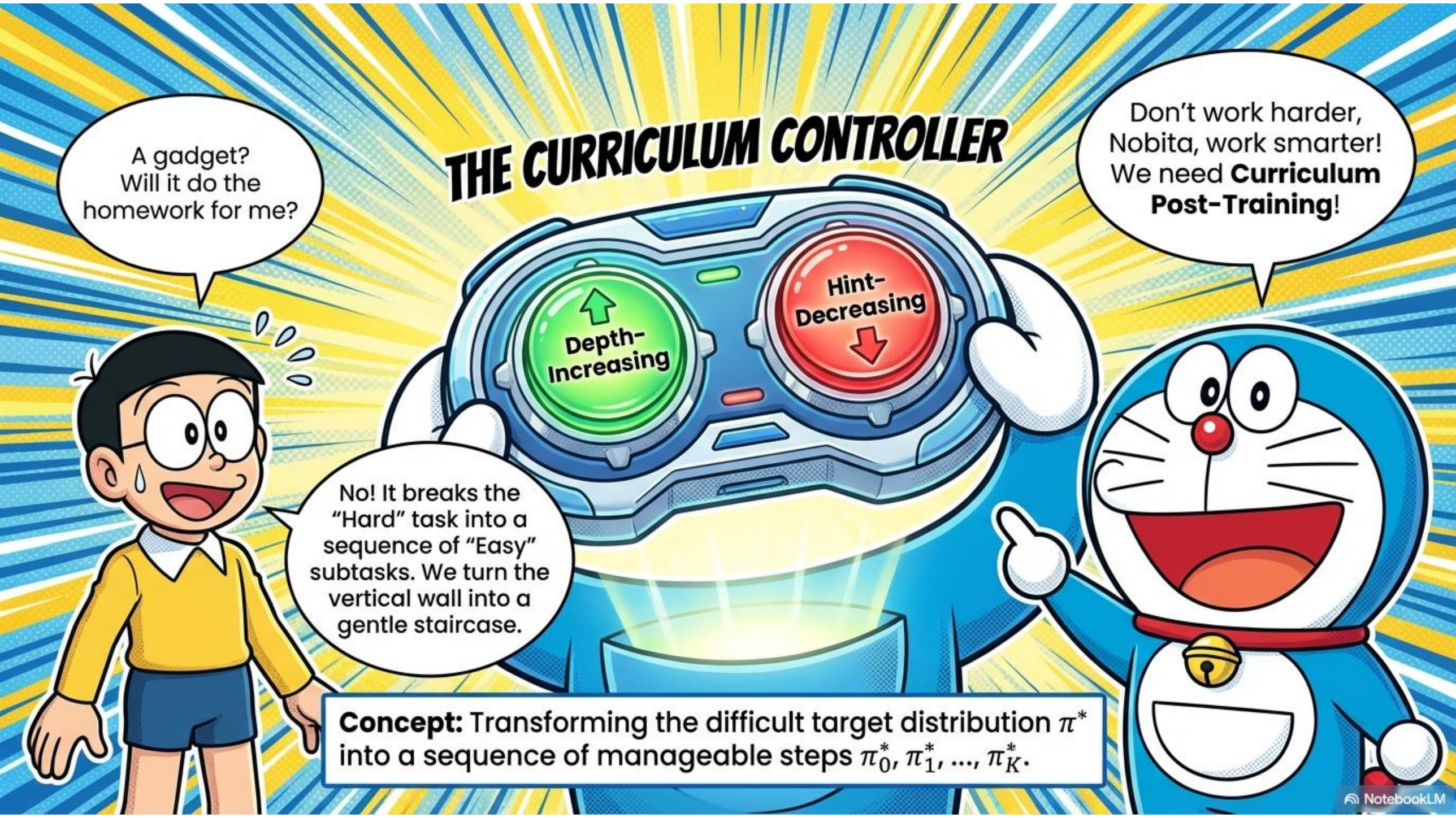
Transformer $(\text{TF}(\cdot; \mathbf{W}))$ **as Learning Model**. We follow Eq. (4), Eq. (5), set the vocabulary as $\mathbf{U} = [\boldsymbol{\mu}^0, \boldsymbol{\mu}^1, \boldsymbol{\mu}_{\text{EOS}}] = [\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3]$ for $0, 1, \text{EOS}$, and the FFN_l is instantiated for XOR operation:

$$\text{FFN}_l = \mathbf{W}_2\, \text{ReLU}[\mathbf{W}_1(\hat{\boldsymbol{\mu}}^{x_{i_l}} + \mathbf{E}[z_{l-1}]_{:d_X})] \quad (7)$$

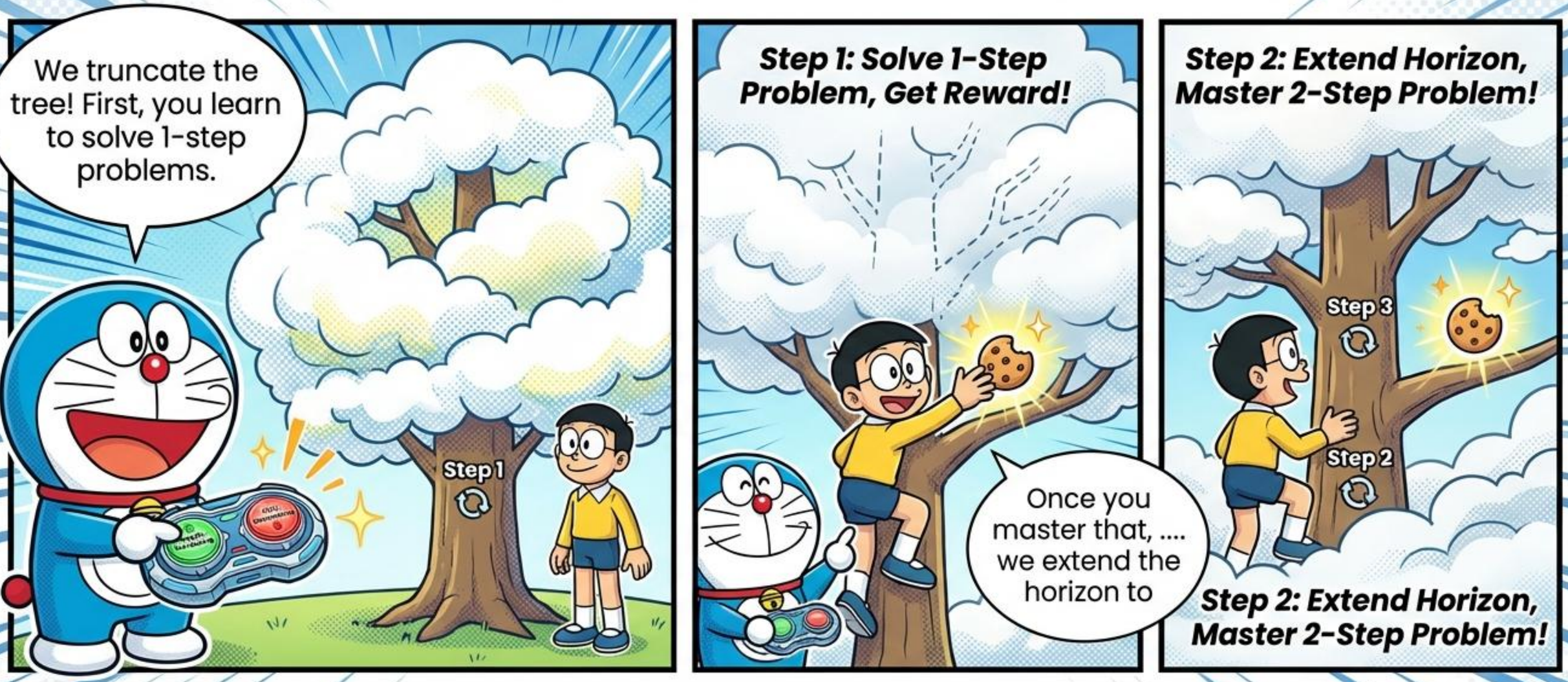
where

$$\mathbf{W}_1 = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ 0 & 1 & 0 \\ -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \end{bmatrix}, \quad \mathbf{W}_2 = \begin{bmatrix} 1 & -1 & 2 \\ 0 & 1 & -2 \\ 0 & 0 & 0 \end{bmatrix}. \quad (8)$$

It can be checked directly that this ensure our FFN_l replicates the $\Phi_l(z_{l-1}, v_l(i_l)) = z_{l-1} \oplus v_l(i_l)$.

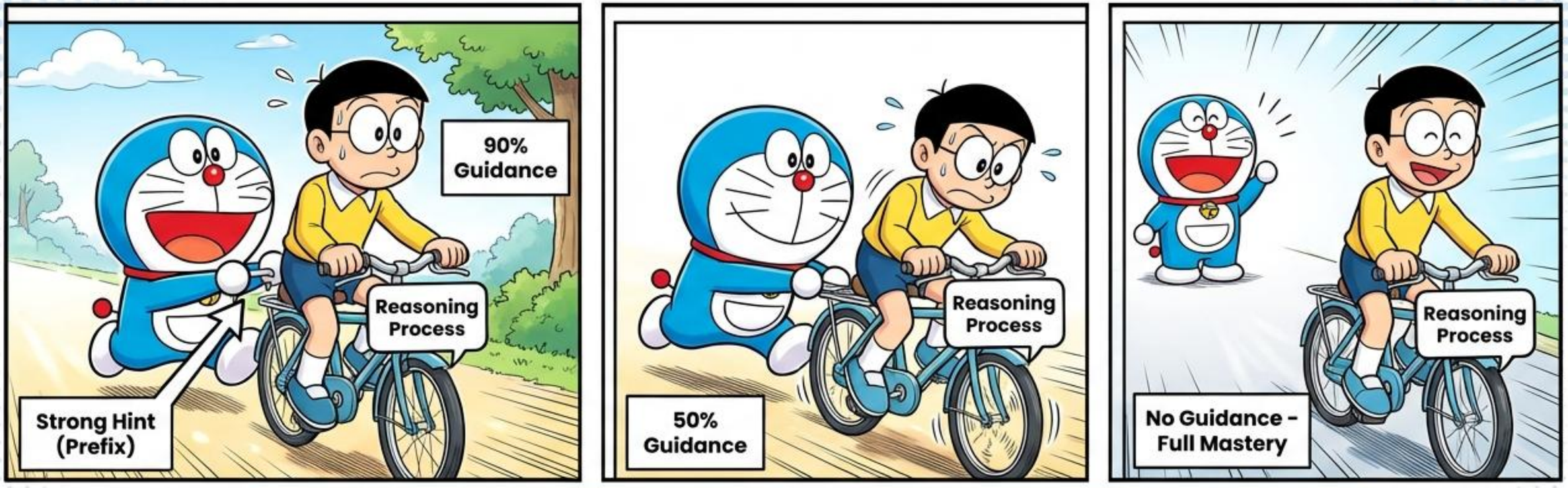

THE CURRICULUM CONTROLLER
A gadget? Will it do the homework for me?
Don't work harder, Nobita, work smarter! We need **Curriculum Post-Training**!
Depth-Increasing
Hint-Decreasing
No! It breaks the "Hard" task into a sequence of "Easy" subtasks. We turn the vertical wall into a gentle staircase.
Concept: Transforming the difficult target distribution π^* into a sequence of manageable steps $\pi_0^*, \pi_1^*, \ldots, \pi_K^*$.
NotebookLM

GADGET MODE 1: DEPTH-INCREASING



Depth-Increasing Curriculum: Progressively extending the reasoning horizon. The model learns short chains before attempting long ones.

GADGET MODE 2: HINT-DECREASING



Mode 2: **Hint-Decreasing!** I give you the correct prefix (a hint) and you just finish the last step.

I get it! It's like reverse-engineering! You give me fewer hints until I can do the whole chain myself!

Hint-Decreasing Curriculum: Completing the Chain-of-Thought from increasingly shorter correct prefixes.

Algorithm 1 No-Curriculum REINFORCE Finetuning (A)

Inputs: input distribution $\mathbf{x} \sim \mathrm{Unif}([K]^d)$, error budget $\varepsilon > 0$, target length $k^\star$, terminal oracle $\mathbf{R}_{\mathbf{x}}^{f_{S_\star}}(\cdot)$, sample complexity $n_0 = \tilde{\Omega}\big(d^{2k^\star+4}\,(1-\rho_{\mathrm{spur}}^{\mathrm{sup},>1})^{-2}\big)$, learning rate $\eta^{\mathsf{A}} = \Theta(\beta \log(d) d^{k^\star+1}(1-\rho_{\mathrm{spur}}^{\mathrm{sup},>1})^{-1})$.
Initialize transformer parameters $\mathbf{W}^{(0)}$.
Draw a single batch of n_0 i.i.d. samples $(\mathbf{x}^{(s)}, \tau^{(s)})_{s=1}^{n_0}$ where $\mathbf{x}^{(s)} \sim \mathrm{Unif}([K]^d)$ and $\tau^{(s)} \sim p_{\mathbf{W}^{(0)}}(\cdot \mid \mathbf{x}^{(s)})$.
Compute empirical REINFORCE gradient:
    $\widehat{\mathbf{g}} = \frac{1}{n_0}\sum_{s=1}^{n_0} \mathbf{R}_{\mathbf{x}^{(s)}}^{f_{S_\star}}(\hat{\mathbf{p}}^{z_{k^\star}}(\tau_{-2}^{(s)})) \sum_{\ell=1}^{k^\star+1} \nabla_{\mathbf{W}} \log \pi_{\mathbf{W}^{(0)}}(i_\ell^{(s)} \mid \mathbf{x}^{(s)}, \hat{\mathbf{p}}^{z_{1:\ell}})$.
Update parameters: $\mathbf{W}^{(1)} = \mathbf{W}^{(0)} + \eta^{\mathsf{A}} \cdot \widehat{\mathbf{g}}$.
Return $\mathbf{W}^{(1)}$.

Algorithm 2 Depth-Increasing Curriculum REINFORCE Finetuning (B)

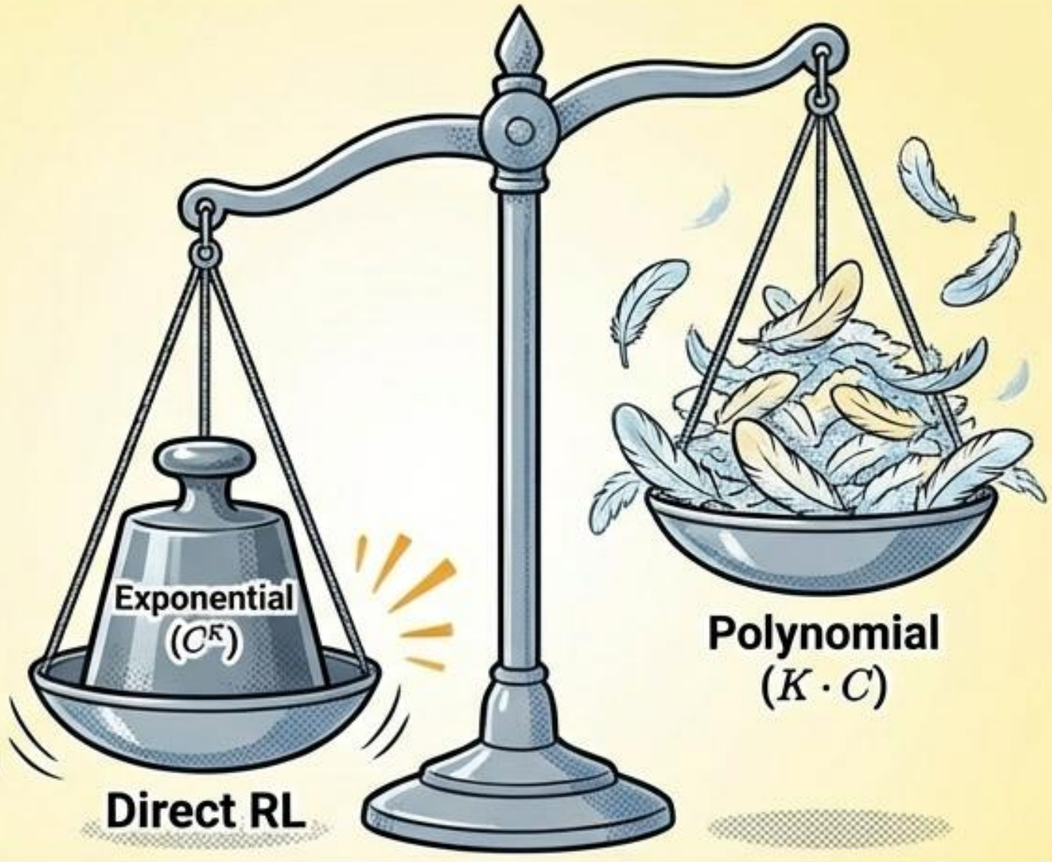
Inputs: input distribution $\mathbf{x} \sim \mathrm{Unif}([K]^d)$, error budget $\varepsilon > 0$, target length $k^\star$, family oracle $\mathbf{R}_{\mathbf{x}}^{\mathcal{F}_{S_\star}}(\cdot)$, per-step sample complexity $n_\ell = \tilde{\Theta}\big(d^2(1-\rho_{\mathrm{spur}}^{\mathrm{sup},\ell})^{-2}\big)$, learning rates $\eta_\ell^{\mathsf{B}} = \Omega(\beta \log(d) d^2 (1-\rho_{\mathrm{spur}}^{\mathrm{sup},\ell})^{-1})$.
Initialize transformer parameters $\mathbf{W}^{(0)}$.
for $\ell = 1$ to $k^\star + 1$ **do**
    Draw fresh samples $(\mathbf{x}^{(s)}, \tau^{(s)})_{s=1}^{n_\ell}$ where $\mathbf{x}^{(s)} \sim \mathrm{Unif}([K]^d)$ and $\tau^{(s)} \sim p_{\mathbf{W}^{(\ell-1)}}(\cdot \mid \mathbf{x}^{(s)})$.
    If $\ell < k^\star + 1$, for each sample s, truncate trajectory at length $\ell + 1$:
        If $\tau^{(s)}$ contains EOS at position $\le \ell$, discard sample.
        Otherwise, append EOS to create truncated sequence $\tau_{\mathrm{trunc}}^{(s)}$.
    If $\ell = k^\star + 1$, $\tau_{\mathrm{trunc}}^{(s)} = \tau^{(s)}$.
    Compute empirical REINFORCE gradient for step ℓ:
        $\widehat{\mathbf{g}}_\ell = \frac{1}{n_\ell}\sum_{s=1}^{n_\ell} \mathbf{R}_{\mathbf{x}^{(s)}}^{\mathcal{F}_{S_\star}}(\tau_{\mathrm{trunc}}^{(s)}) \nabla_{\mathbf{W}} \log \pi_{\mathbf{W}^{(\ell-1)}}(i_\ell^{(s)} \mid \mathbf{x}^{(s)}, \hat{\mathbf{p}}^{z_{1:\ell}})$.
    Update parameters: $\mathbf{W}^{(\ell)} = \mathbf{W}^{(\ell-1)} + \eta_\ell^{\mathsf{B}} \cdot \widehat{\mathbf{g}}_\ell$.
end for
Return $\mathbf{W}^{(k^\star+1)}$.

THE SECRET SAUCE: COVERAGE COEFFICIENT.

Here is the secret sauce: The **Coverage Coefficient** (C_{cov}).

$$N_{\text{curriculum}} \ll N_{\text{direct}}$$

It measures how 'surprised' you are by the answer. Direct learning has infinite surprise. But with the Curriculum Gadget, the 'surprise' at each step is small and bounded.



Theorem 2: Curriculum strategies achieve high accuracy with **Polynomial** sample complexity, avoiding the exponential bottleneck of direct training.

THE SECRET SAUCE: COVERAGE COEFFICIENT.

Theorem 3 (Curriculum RL Finetuning Avoid Exponential Bottleneck)**.** *Let* $\mathbf{U}$ *be an orthogonal matrix with* $d_{\mathrm{X}} = \Theta(K), d_{\mathrm{E}} = \Theta(d + L)$, $\mathrm{TF}(\cdot; \mathbf{W}_{base})$ *per in Thm.* 2 *with trainable* $\mathbf{W}$*, and a target* $f_{S_\star} \in \mathcal{F}_{\text{2S-ART}}$ *with* $|S_\star| = k^\star + 1$*. For any* $\varepsilon \in (0, 1)$*, using the RL objective in Eq.* (9) *and one gradient step per stage (a single step for no-curriculum;* $k^\star{+}1$ *online steps for curricula), with probability no less than* $1 - \delta$*, there exist learning-rate choices* η *such that*

1. *No-curriculum (*$R_{\mathbf{x}}^{f_{S_\star}}(\cdot)$ *as oracle, one-shot update with* $\eta = \tilde{\Theta}(\beta d^{k^\star+1})$*): the sample complexity to achieve* $\mathbb{E}_{\mathbf{x}\sim\mathrm{Unif}([K]^d)}\left[R_{\mathbf{x}}^{f_{S_\star}}(\mathrm{TF}(\mathbf{x}; \mathbf{W}))\right] \geq 1{-}\varepsilon$ *is at least* $n \geq \tilde{\Omega}(d^{2k^\star+2})$.

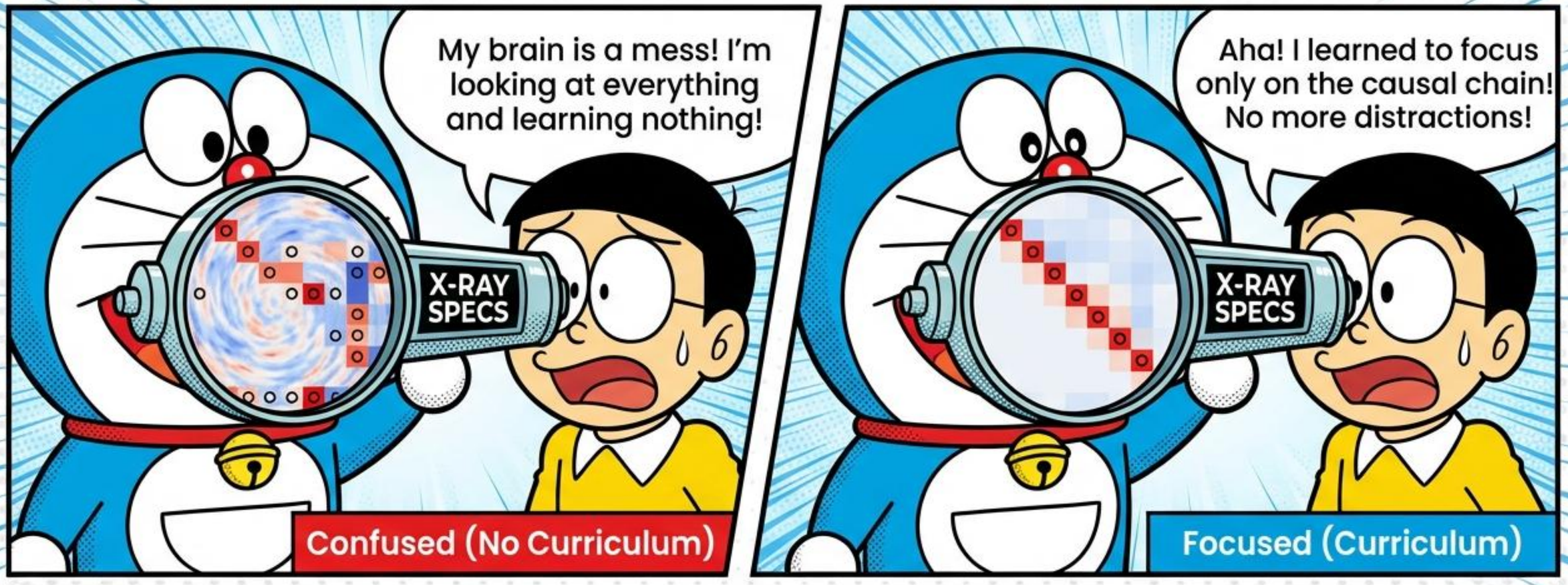
2. *Depth-increasing Curriculum and Hint-decreasing Curriculum (*$\mathbf{R}_{\mathbf{x}}^{\mathcal{F}_{S_\star}}(\cdot)$ *as oracle,* $k^\star{+}1$ *online updates with* $\eta = \tilde{\Theta}(\beta d^2)$*): the sample complexity to achieve* $\mathbb{E}_{\mathbf{x}\sim\mathrm{Unif}([K]^d)}\left[R_{\mathbf{x}}^{f_{S_\star}}(\mathrm{TF}(\mathbf{x}; \mathbf{W}))\right] \geq 1 - \varepsilon$ *is at most* $n \leq \tilde{O}((k^\star + 1)d^2)$.

Here $\tilde{\Omega}(\cdot)$ *and* $\tilde{O}(\cdot)$ *hide polylogarithmic factors in* $d, 1/\delta$ *and* $1/\varepsilon$*, and absolute constants (e.g., temperature* β*) as well as task-dependent spurious-acceptance parameters.*

surprise' at each step is small and bounded.

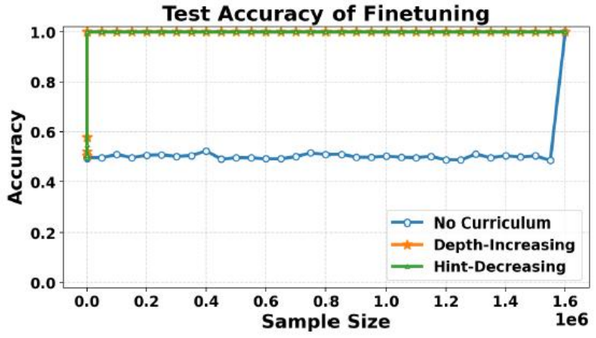
complexity, avoiding the exponential bottleneck of direct training.

SIMULATION: INSIDE THE ELECTRONIC BRAIN

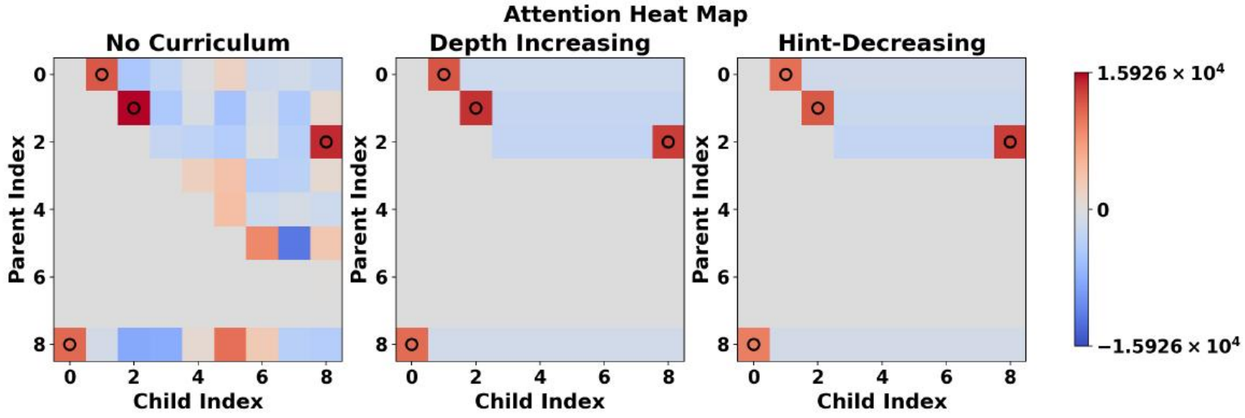


Empirical Evidence: Curriculum training leads to clean attention heads that track the causal chain, while direct training gets lost in noise.

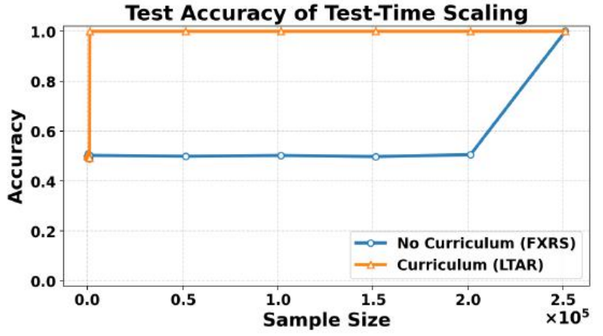
SIMULATION: INSIDE THE ELECTRONIC BRAIN



(a) Test Accuracy of Finetuning.



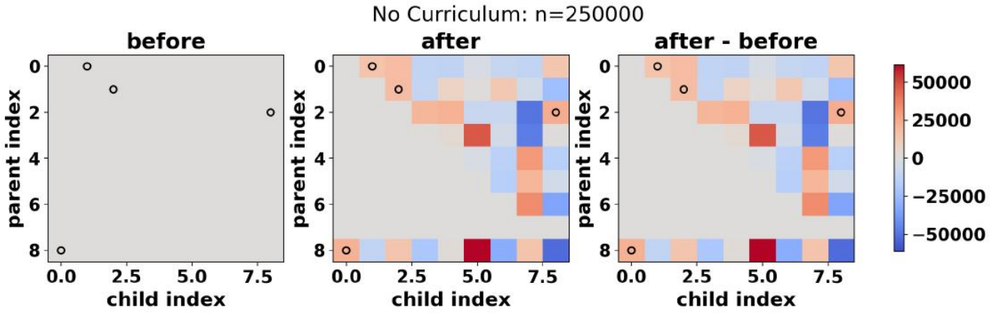
(b) Attention Heat Map after Finetuning.



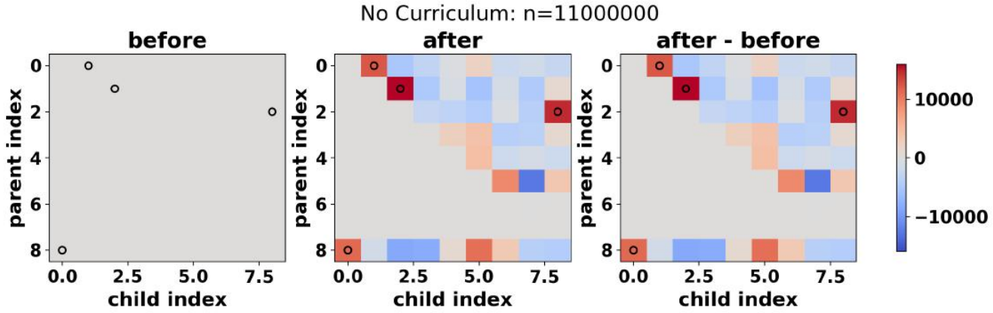
(c) Test Accuracy of Test-Time Scaling.

Figure 3. Dynamics of Finetuning and Test-time Scaling on the parity task ($d = 8$, $k^\star = 3$, $S^\star = [0, 1, 2]$). (a) Test accuracy evolution of finetuning along sample size; (b) attention heatmap after finetuning; (c) test accuracy evolution of test-time scaling along reward oracle query complexity. For finetuning, both curriculum algorithms (Alg. 1 & Alg. 2) converge at sample size $n = 76$, where no-curriculum counterpart (Alg. 1) converges at $n = 1600076$; Attention learning of curriculum algorithms focus on crucial parent-children maps, while no curriculum counterpart is contaminated with spurious corelations; For test-time scaling, curriculum algorithm (LTAR in Alg. 6) converges at $n = 1410$ while its no curriculum counterpart (Alg. 4) converges at $n = 251410$.

Empirical Evidence: Curriculum training leads to clean attention heads that track the causal chain, while direct training gets lost in noise.

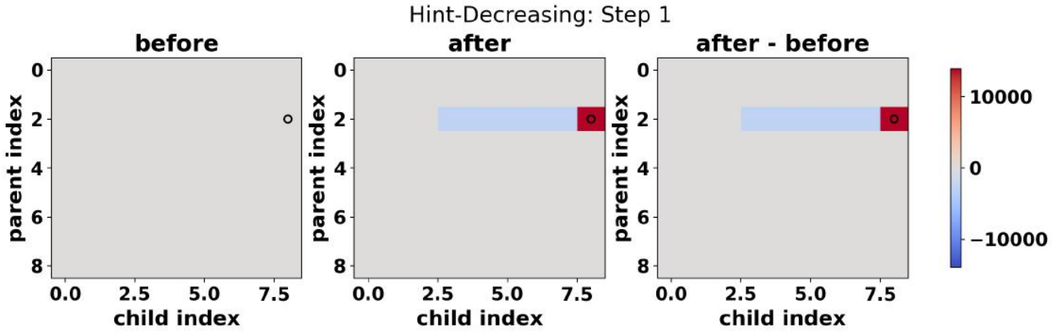


(a) $n = 250000$.

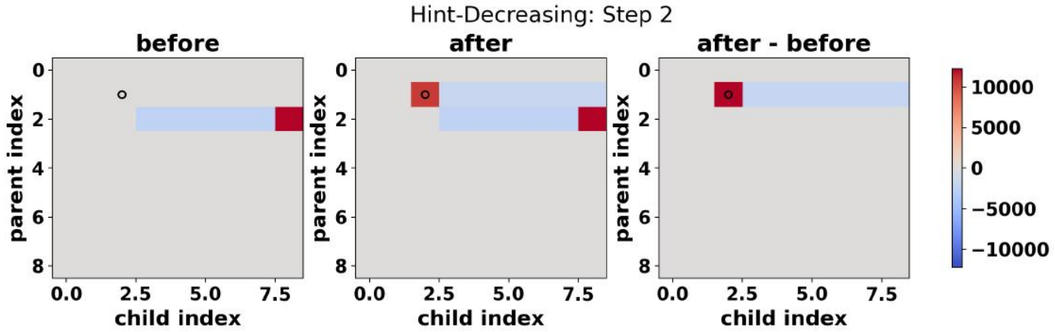


(b) $n = 11000000$.

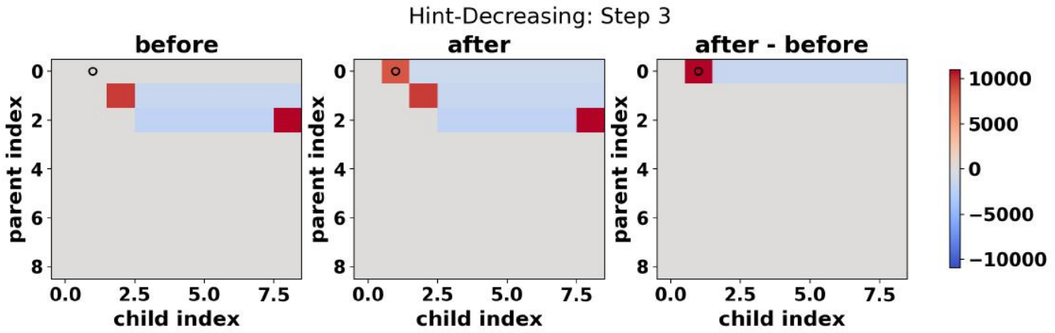
Figure 4. Attention Heatmaps of No-Curriculum (Direct Learning) with different sample complexity.



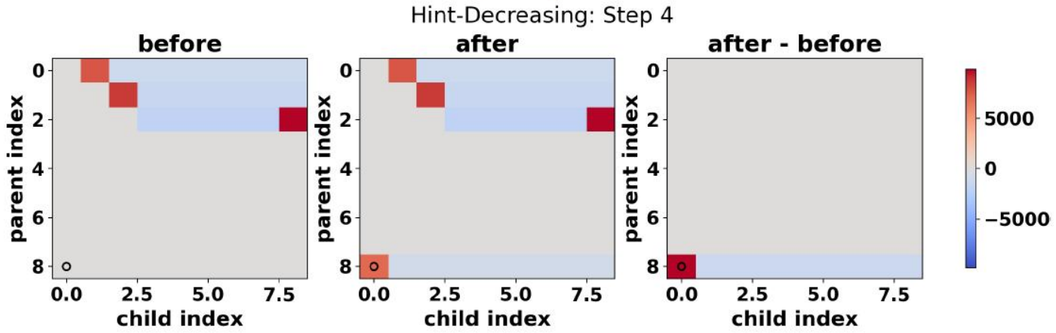
(a) Step 1.



(b) Step 2.

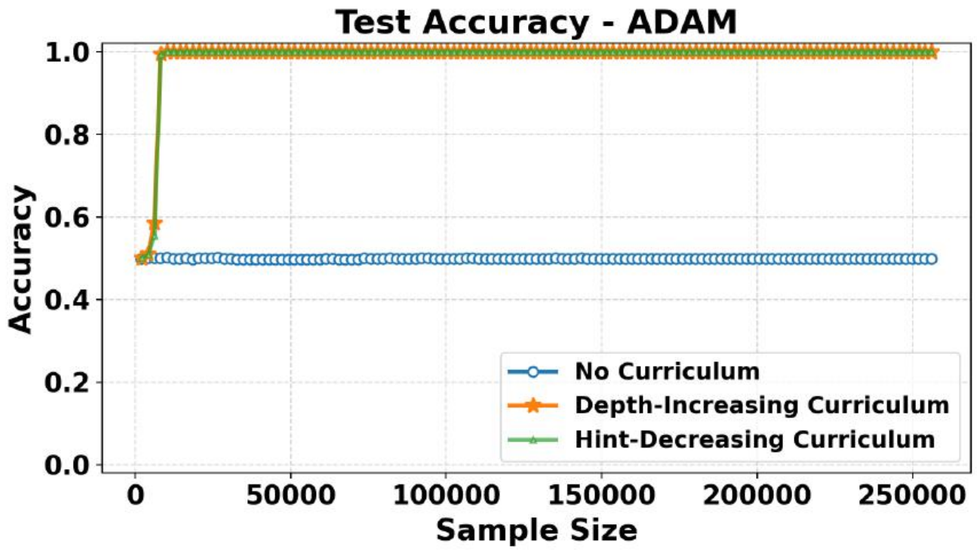


(c) Step 3.

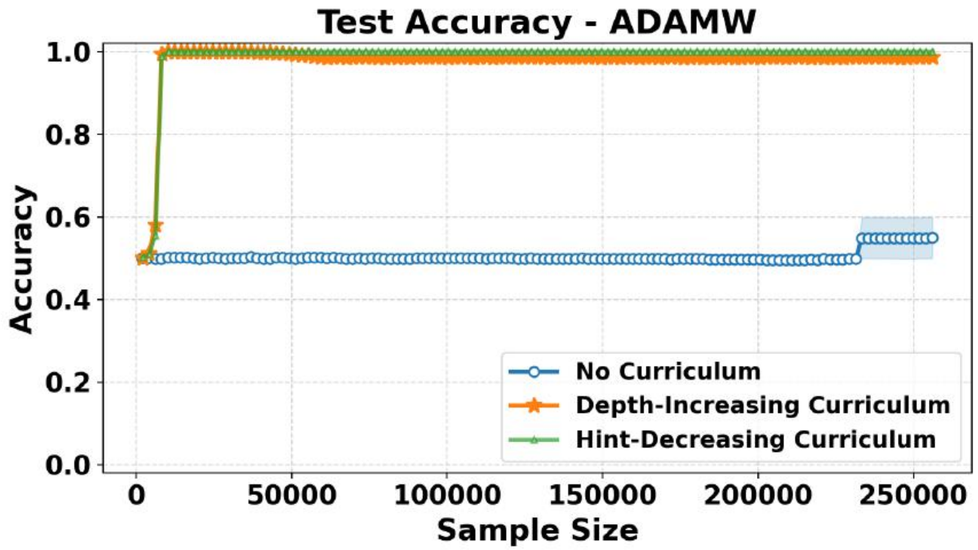


(d) Step 4.

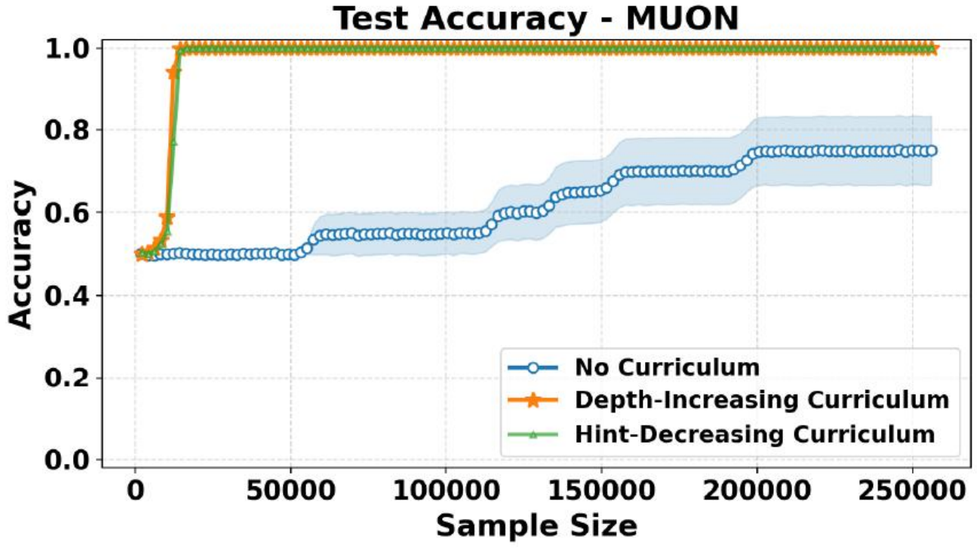
Figure 7. Step-wise Attention Heatmaps for Hint-Decreasing Curriculum ($d = 8, k^\star = 3, S^\star = [0, 1, 2], n = 100000$).



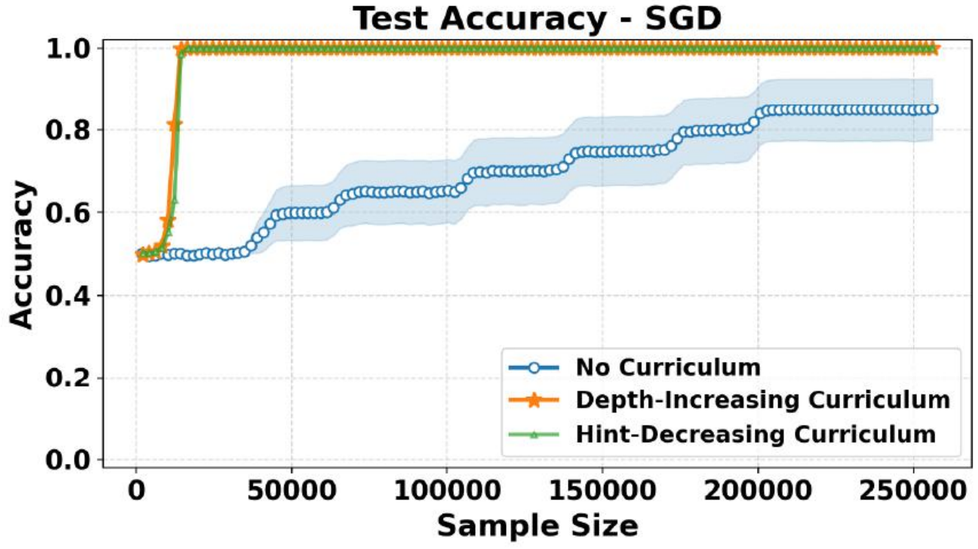
(a) Adam



(b) AdamW

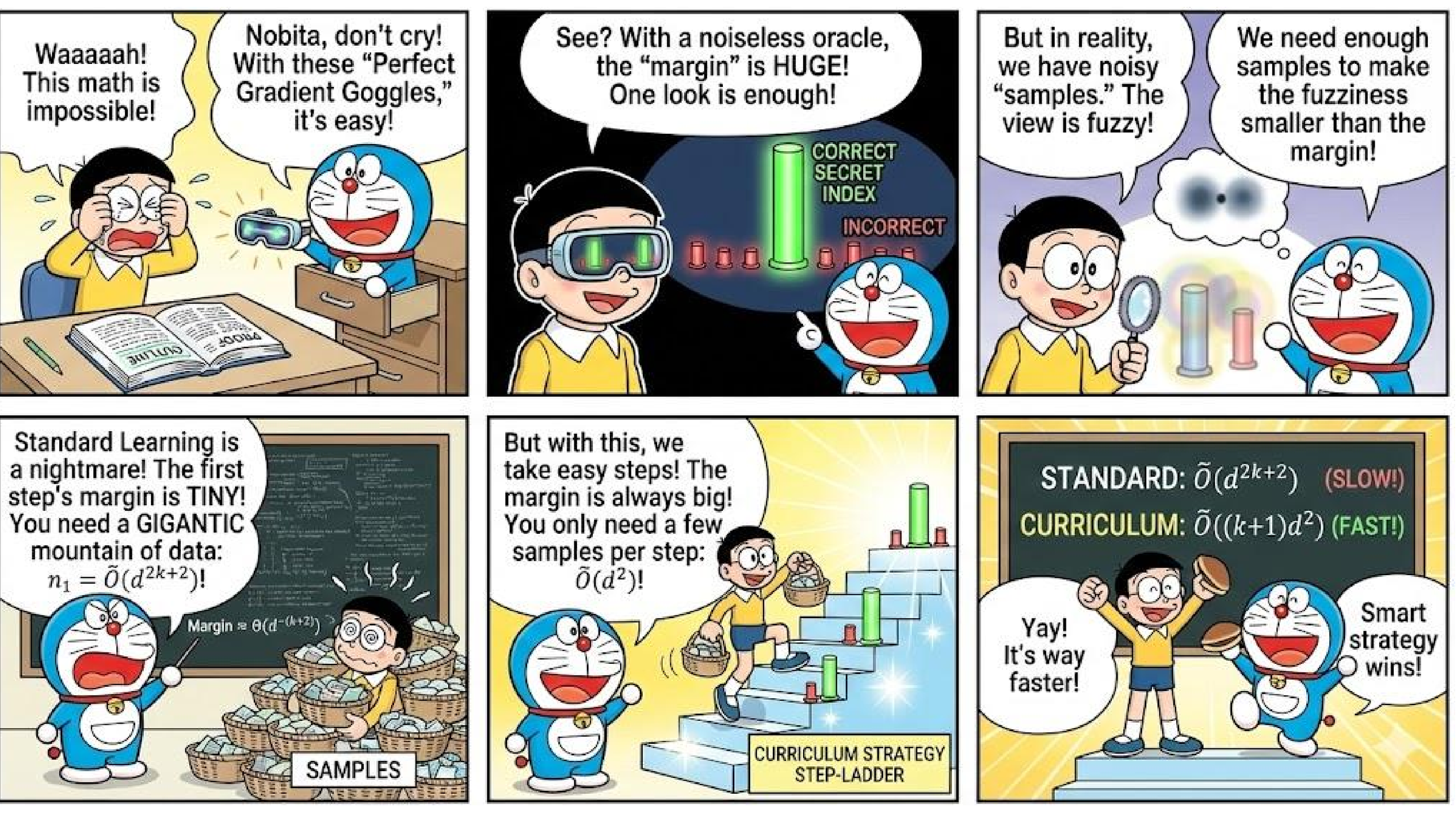


(c) Muon

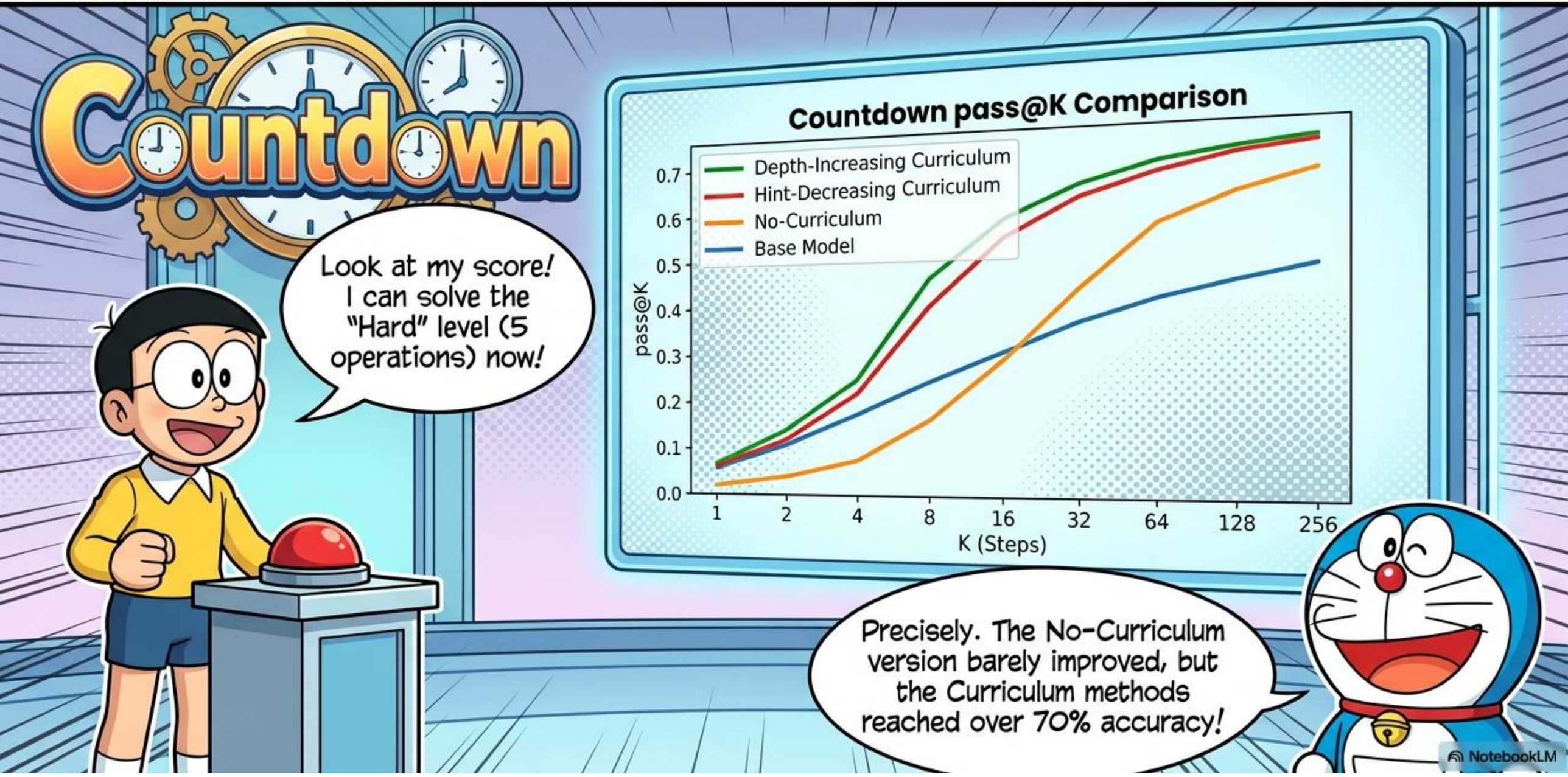


(d) SGD

Figure 11. Small Learning rate regime ($d = 8, k^\star = 3, S^\star = [0, 1, 2], \eta = 0.5, B = 256$).


Waaaaah! This math is impossible!
Nobita, don't cry! With these "Perfect Gradient Goggles," it's easy!
See? With a noiseless oracle, the "margin" is HUGE! One look is enough!
CORRECT SECRET INDEX
INCORRECT
But in reality, we have noisy "samples." The view is fuzzy!
We need enough samples to make the fuzziness smaller than the margin!
Standard Learning is a nightmare! The first step's margin is TINY! You need a GIGANTIC mountain of data: $n_1 = \tilde{O}(d^{2k+2})$!
Margin $\approx \Theta(d^{-(k+2)})$
SAMPLES
But with this, we take easy steps! The margin is always big! You only need a few samples per step: $\tilde{O}(d^2)$!
CURRICULUM STRATEGY STEP-LADDER
STANDARD: $\tilde{O}(d^{2k+2})$ (SLOW!)
CURRICULUM: $\tilde{O}((k+1)d^2)$ (FAST!)
Yay! It's way faster!
Smart strategy wins!

THE COUNTDOWN GAME CHALLENGE



THE COUNTDOWN GAME CHALLENGE

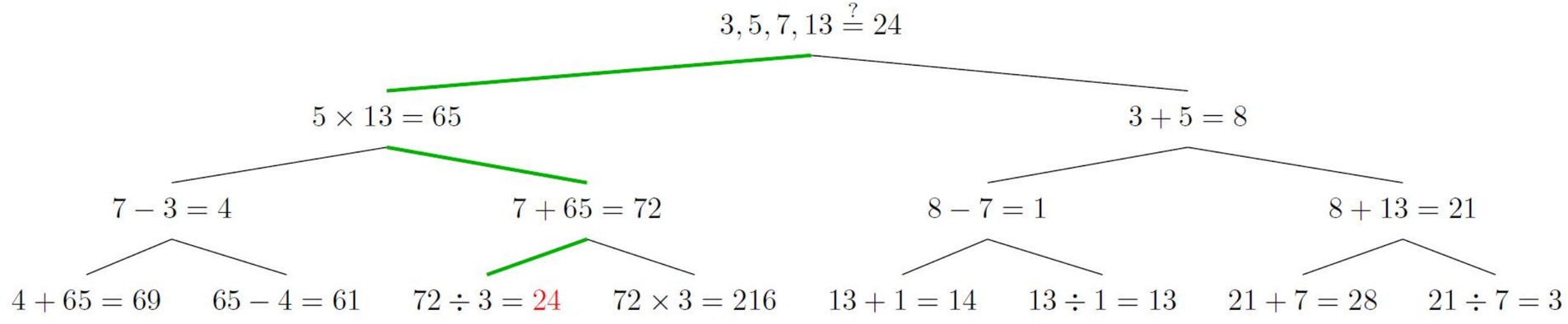
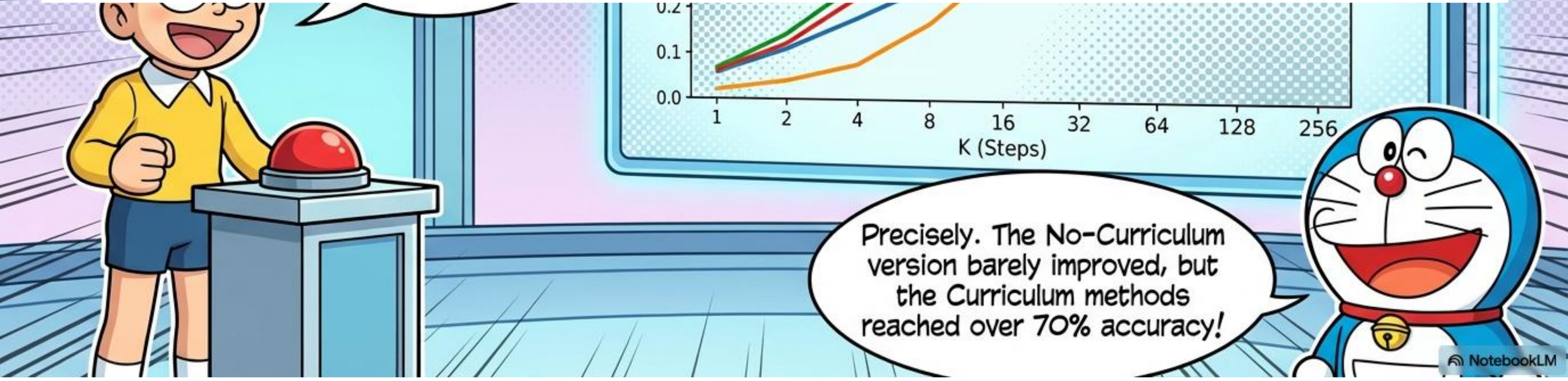
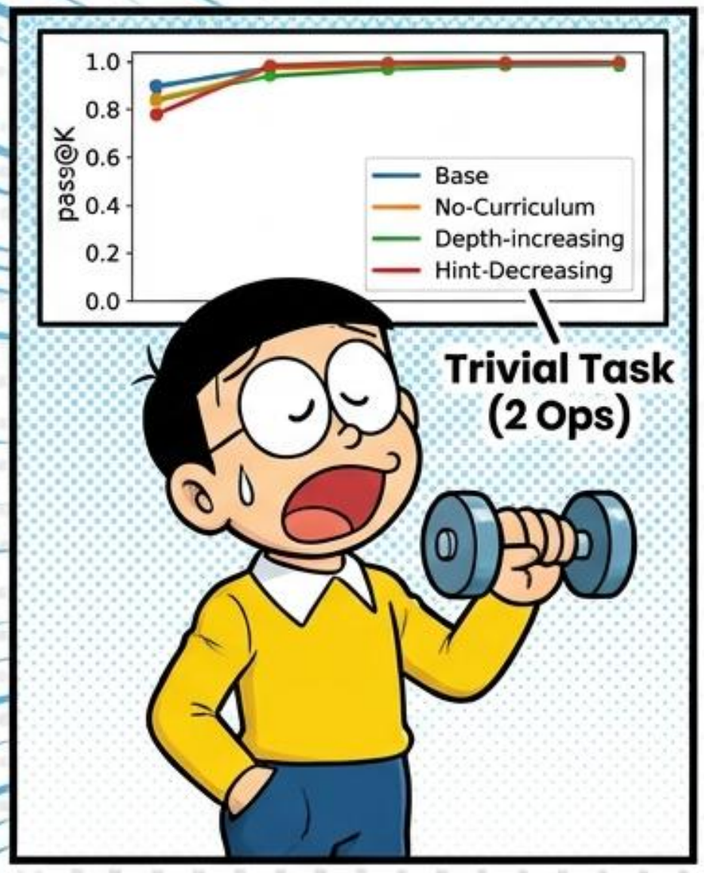
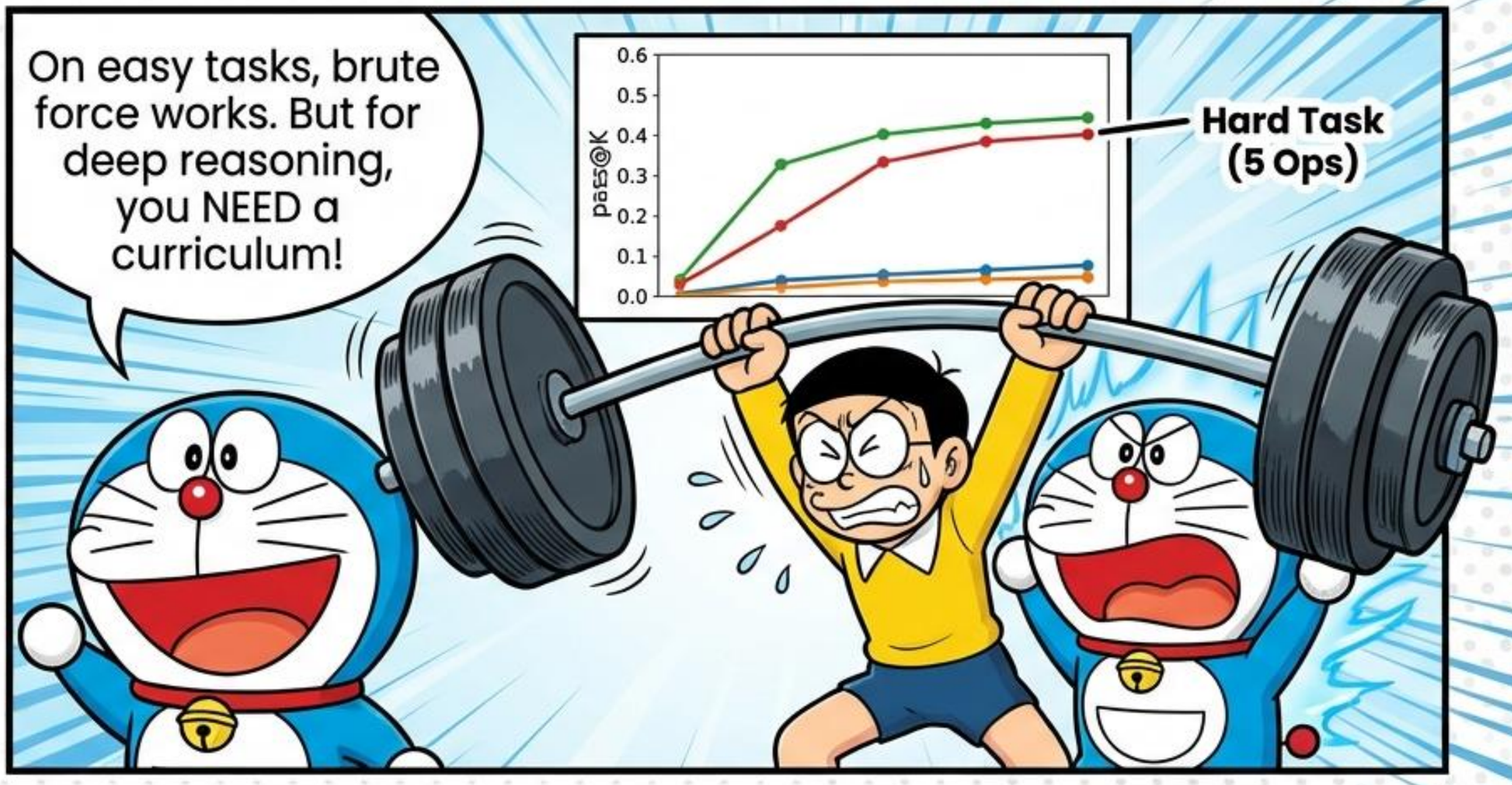


Figure 1. An illustration in Liu et al. (2025b) of the Chain-of-Thought for Countdown game, where the goal is to obtain 24 by applying basic arithmetic operations $(+, -, \times, \div)$ ($\Phi_l(\cdot, \cdot)$ in our Def. 2) between the current step's number (e.g., 13, 65 or 72) and some unused number (e.g., 5, 7 or 3) in $\{3, 5, 7, 13\}$, targeting 24 as the final outcome. Per (Parashar et al., 2025), the difficulty measure of Countdown is the number of arithmetic operations required to solve an instance.

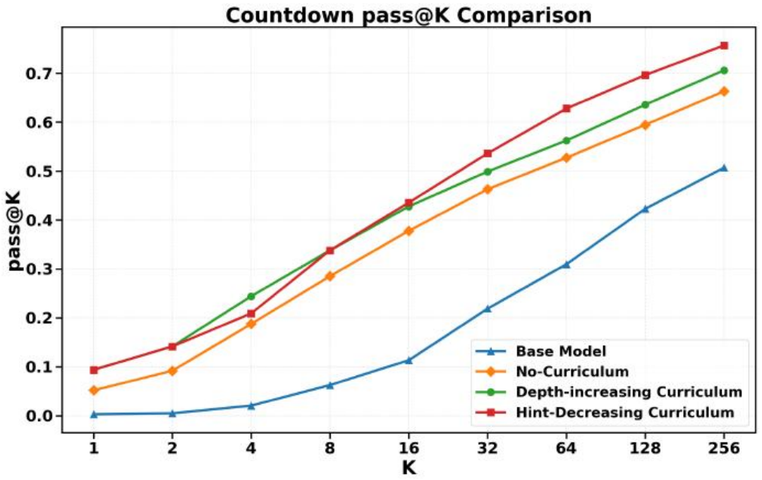


OUT-OF-DISTRIBUTION (OOD) GENERALIZATION

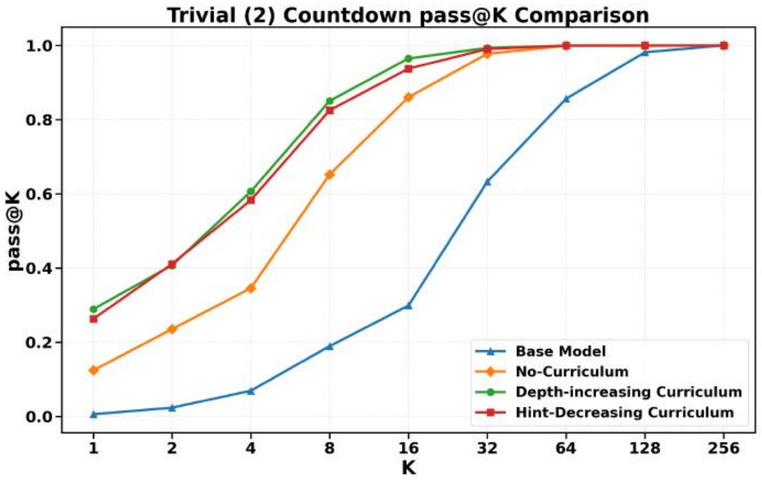




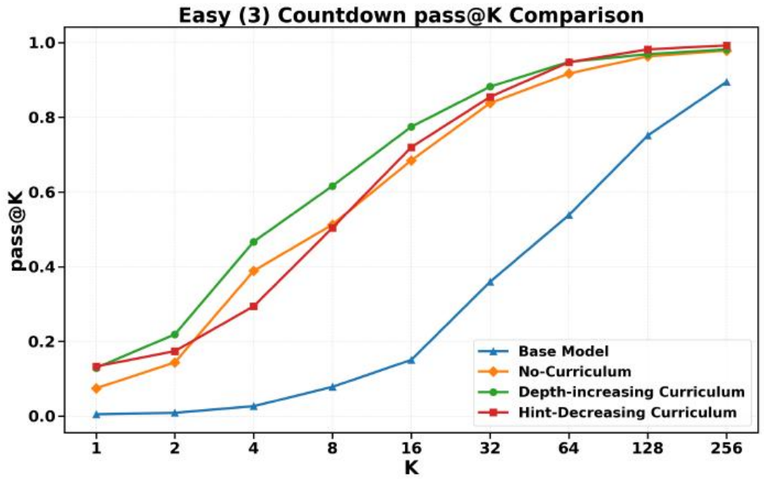
Insight: Curriculum methods show the largest relative gains on Out-Of-Distribution (deeper) tasks.



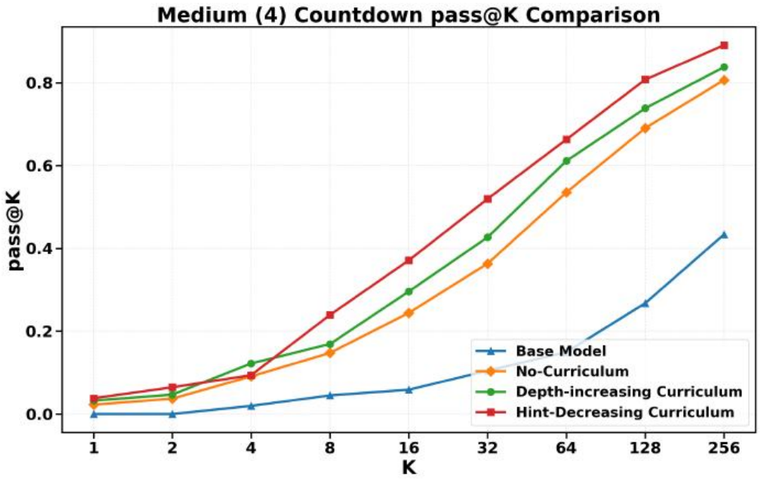
(a) Total validation set.



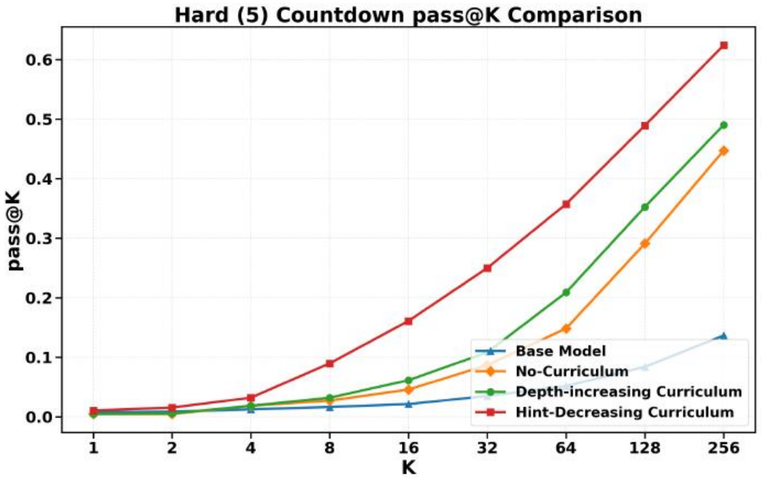
(b) Trivial (2 ops).



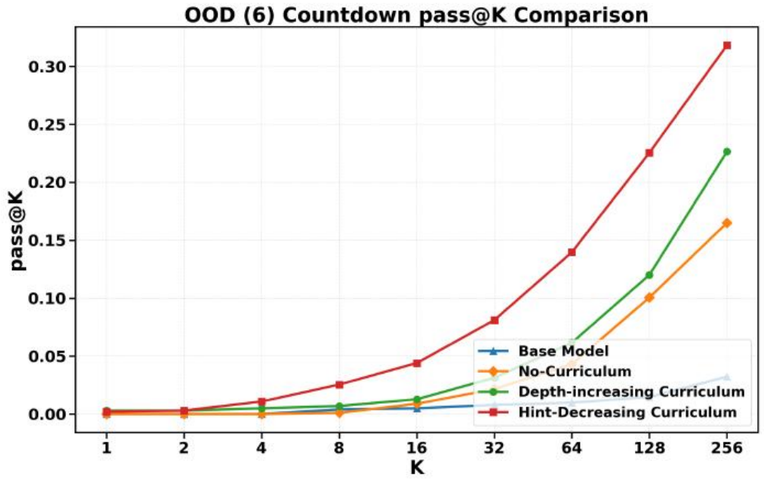
(c) Easy (3 ops).



(d) Medium (4 ops).



(e) Hard (5 ops).



(f) OOD (6 ops).

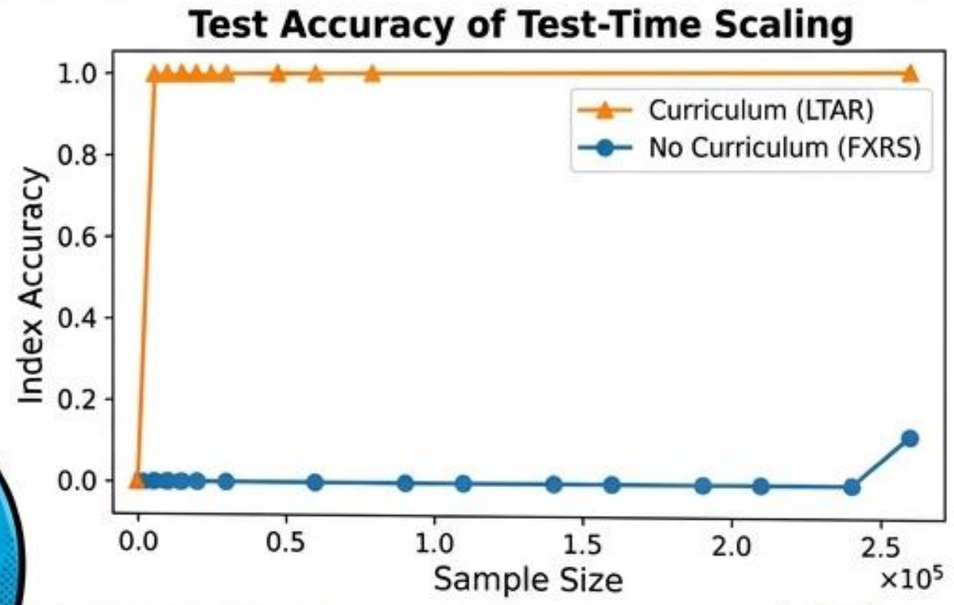
Figure 12. **Test performance on Countdown across difficulty strata.** Both Depth-Increasing and Hint-Decreasing curricula outperform the No-Curriculum baseline as reasoning depth increases. Performance gaps are modest on Trivial and Easy subsets but widen substantially on Medium and Hard subsets. On the OOD (6-operation) subset, curriculum methods achieve the largest relative gains, indicating improved generalization to deeper reasoning chains.

TEST-TIME SCALING: THE ANYWHERE DOOR

This works for testing too! We use a **Curriculum-Aware Oracle** to verify steps incrementally.

So I catch mistakes early? That saves so much time!

EXAM



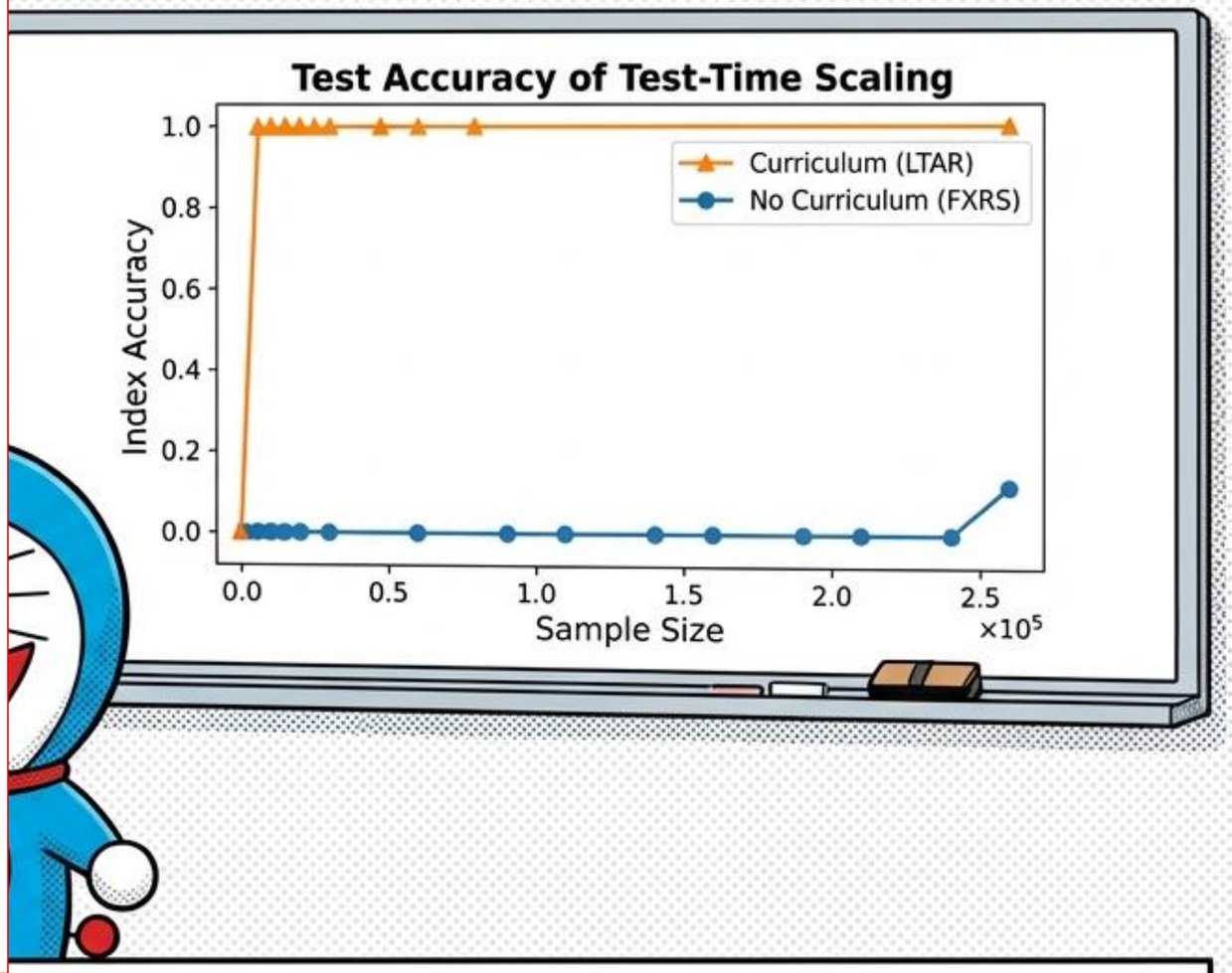
Theorem 3: Provable efficiency in Test-Time Scaling. Reduces query cost from Exponential to Polynomial.

TEST-TIME SCALING: THE ANYWHERE DOOR

Theorem 3 (Curriculum Test-time Scaling Avoid Exponential Bottleneck)**.** *Let* $\mathrm{TF}_{base} := \mathrm{TF}(\cdot; \mathbf{W}_{base})$ *per definedbe as in Thm. 1, and consider a target* $f_{S_\star} \in \mathcal{F}_{2S\text{-}ART}$ *with* $|S_\star| = k^\star + 1$*. Then, for identifying the ground-truth path* $S_\star$ *with confidence* $1 - \delta$*:*

1. *Using only* $\mathbf{R}_{\mathbf{x}}^{f_{S_\star}}$*, any adaptive procedure (including ones that force visible* x*-tokens by rejection sampling and then roll out) requires* $T_{data},\ T_{comp} \geq \tilde{\Omega}(d^{2k^\star})$*.*
2. *Using curriculum queries adaptively to* $\mathbf{R}_{\mathbf{x}}^{\mathcal{F}_{S_\star}}$*, there exists a procedure with* $T_{data} \leq \tilde{O}((k^\star + 1)d^2)$ *and* $T_{comp} \leq \tilde{O}((k^\star + 1)d^3)$*.*

Here $\tilde{\Omega}(\cdot)$ *and* $\tilde{O}(\cdot)$ *hide polylogarithmic factors in* d*,* $1/\delta$ *and spurious-acceptance parameters.*



Theorem 3: Provable efficiency in Test-Time Scaling. Reduces query cost from Exponential to Polynomial.

Algorithm 4 Forced X-token Rejection Sampling (FXRS) under token-only observability

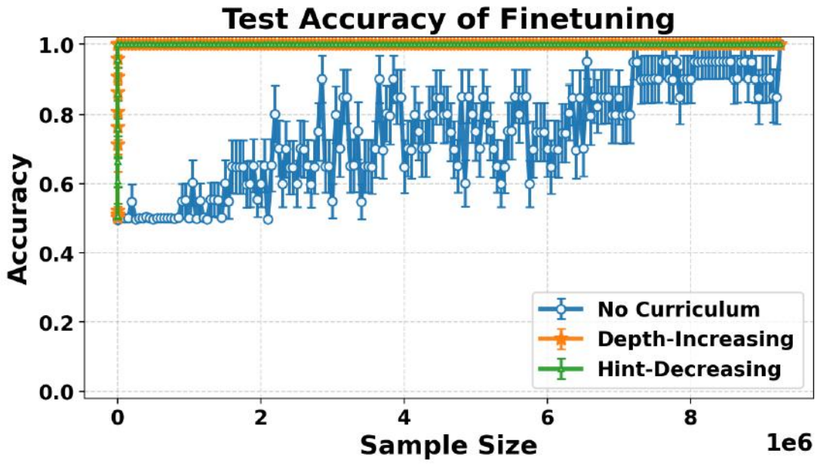
Inputs: depth ℓ, committed history $\mathbf{CoT}_{\ell-1}$, legal set $\mathcal{I}_\ell$, candidate $j \in \mathcal{I}_\ell$, model $\mathrm{TF}_{\mathrm{base}}$, oracle $\mathbf{R}_{\mathbf{x}}^{f_{S^\star}}$, max trials $T_{\max} = \Theta(d \log(\frac{d}{\delta}))$.
for $t = 1$ to $T_{\max}$ **do**
  From context $\mathbf{CoT}_{\ell-1}$, sample one step with $\mathrm{TF}_{\mathrm{base}}$ to emit $\hat{\boldsymbol{\mu}}^{x_{i_\ell}}$.
  if $\hat{\boldsymbol{\mu}}^{x_{i_\ell}} = \boldsymbol{\mu}^{x_j}$ **then**
    Roll out the remaining suffix using $\mathrm{TF}_{\mathrm{base}}$ until termination (EOS occurs at some depth by the copied PART behavior).
    Return the full rollout sequence and **stop**.
  end if
end for
If no match after $T_{\max}$, return **fail** (increase $T_{\max}$ if needed).

Algorithm 5 BAI across depths with FXRS and $\mathbf{R}_{\mathbf{x}}^{f_{S^\star}}$

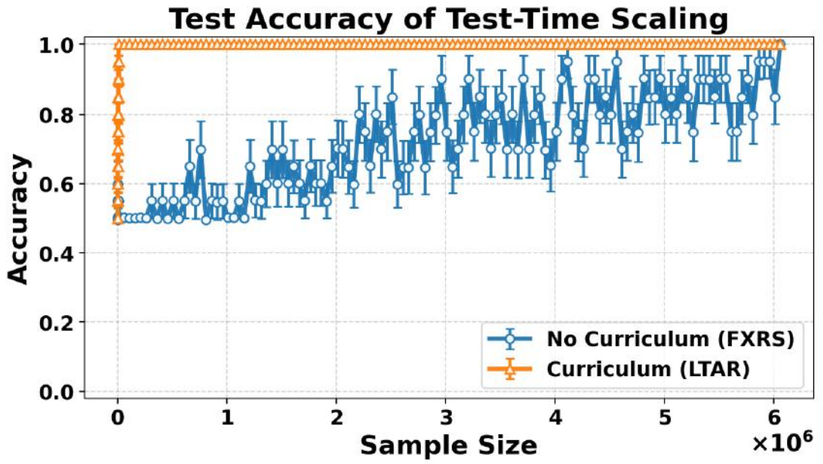
Inputs: legal sets $\{\mathcal{I}_\ell\}_{\ell=1}^{k^\star}$, model $\mathrm{TF}_{\mathrm{base}}$, oracle $\mathbf{R}_{\mathbf{x}}^{f_{S^\star}}$, confidence δ.
Initialize committed partial CoT $\mathcal{H} \leftarrow (\,)$.
for $\ell = 1$ to $k^\star + 1$ **do**
  Set per-arm repetitions $m \leftarrow \Theta\big(d^{2(k^\star+1-\ell)} \log(d/\delta)\big)$.
  for each $j \in \mathcal{I}_\ell$ **do**
    Initialize counter $c_j \leftarrow 0$.
    for $r = 1$ to m **do**
      Run FXRS (Alg. 4) with candidate j to obtain a full rollout (or retry until success). If FXRS fails, repeat until success.
      Query $\mathbf{R}_{\mathbf{x}}^{f_{S^\star}}$ on the final pre-EOS token of the rollout; record one Bernoulli outcome $Y_{j,r} \in \{0, 1\}$.
      Update $c_j \leftarrow c_j + Y_{j,r}$.
    end for
    Set empirical acceptance $\hat{p}_j \leftarrow c_j/m$.
  end for
  Let $\hat{j}_\ell \leftarrow \arg\max_{j \in \mathcal{I}_\ell} \hat{p}_j$.
  Commit depth-ℓ: resample until $\hat{\boldsymbol{\mu}}^{x_{i_\ell}} = \boldsymbol{\mu}^{x_{\hat{j}_\ell}}$, then compute $\hat{\boldsymbol{\mu}}^{z_\ell} = \mathrm{FFN}_\ell(\hat{\boldsymbol{\mu}}^{x_{i_\ell}}, \mathbf{E}[z_{\ell-1}]_{:d_{\mathrm{X}}})$; update $\mathcal{H} \leftarrow (\mathcal{H}, \hat{\boldsymbol{\mu}}^{x_{\hat{j}_\ell}}, \hat{\boldsymbol{\mu}}^{z_\ell})$.
end for
Return the committed path encoded by $\mathcal{H}$.

Algorithm 6 Layer-wise Truncated Accept-Reject (LTAR) with $\mathbf{R}_{\mathbf{x}}^{\mathcal{F}_{S_\star}}$ (BAI with inline forcing and truncation)

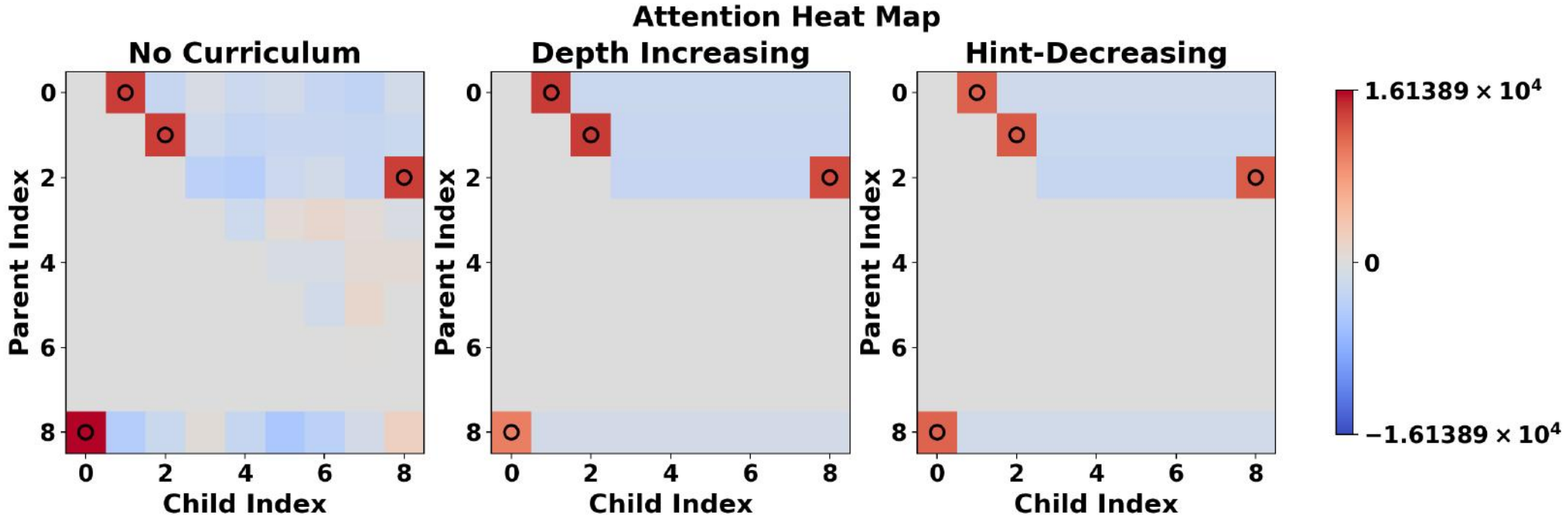
Inputs: legal sets $\{\mathcal{I}_\ell\}_{\ell=1}^{k^\star}$, confidence level δ, per-depth budgets $m_\ell = \Theta\big(d^2(1-\rho_{\text{spur}}^{\text{sup},\ell})^{-2}\log(\frac{d}{\delta})\big)$, max trials $T_{\max} = \Theta(d\log(\frac{d}{\delta}))$.
Initialize committed partial CoT $\mathcal{H} \leftarrow (\,)$.
for $\ell = 1$ to $k^\star + 1$ **do**
  For each $j \in \mathcal{I}_\ell$, estimate acceptance by m_ℓ repetitions with inline forcing:
    Initialize $c_j \leftarrow 0$.
  for $r = 1$ to m_ℓ **do**
    for $t = 1$ to $T_{\max}$ **do**
      From context $\mathcal{H}$, sample one step with $\mathrm{TF}_{\text{base}}$ to emit $\hat{\boldsymbol{\mu}}^{x_{i_\ell}}$.
      if $\hat{\boldsymbol{\mu}}^{x_{i_\ell}} = \boldsymbol{\mu}^{x_j}$ **then**
        Compute $\hat{\boldsymbol{\mu}}^{z_\ell} \leftarrow \mathrm{FFN}_\ell\big(\hat{\boldsymbol{\mu}}^{x_{i_\ell}},\ \mathbf{E}[z_{\ell-1}]_{:d_{\mathrm{X}}}\big)$.
        Externally append EOS if $l < k^\star + 1$; query $\mathbf{R}_{\mathbf{x}}^{\mathcal{F}_{S_\star}}$ at depth ℓ; record Bernoulli $Y_{j,r} \in \{0,1\}$.
        Update $c_j \leftarrow c_j + Y_{j,r}$ and **break** the retry loop.
      end if
    end for
    If no match within $T_{\max}$, optionally increase $T_{\max}$ and repeat this repetition.
  end for
  Set empirical acceptance $\hat{p}_j \leftarrow c_j/m_\ell$ for all $j \in \mathcal{I}_\ell$ and pick $\hat{j}_\ell = \arg\max_j \hat{p}_j$.
  Commit depth-ℓ: resample until $\hat{\boldsymbol{\mu}}^{x_{i_\ell}} = \boldsymbol{\mu}^{x_{\hat{j}_\ell}}$, then compute $\hat{\boldsymbol{\mu}}^{z_\ell} = \mathrm{FFN}_\ell(\hat{\boldsymbol{\mu}}^{x_{i_\ell}}, \mathbf{E}[z_{\ell-1}]_{:d_{\mathrm{X}}})$; update $\mathcal{H} \leftarrow (\mathcal{H}, \hat{\boldsymbol{\mu}}^{x_{\hat{j}_\ell}}, \hat{\boldsymbol{\mu}}^{z_\ell})$.
end for
Return the committed path encoded by $\mathcal{H}$.



(a) Test Accuracy of d.

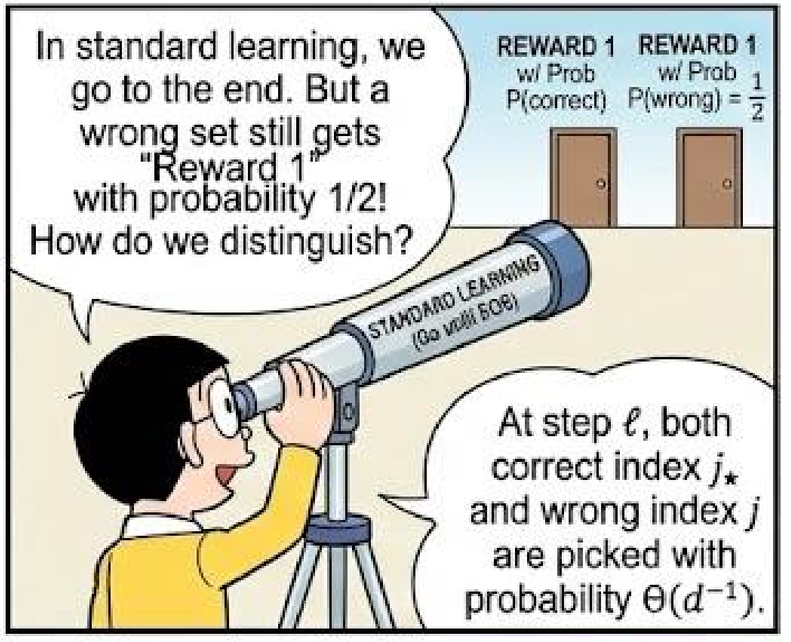


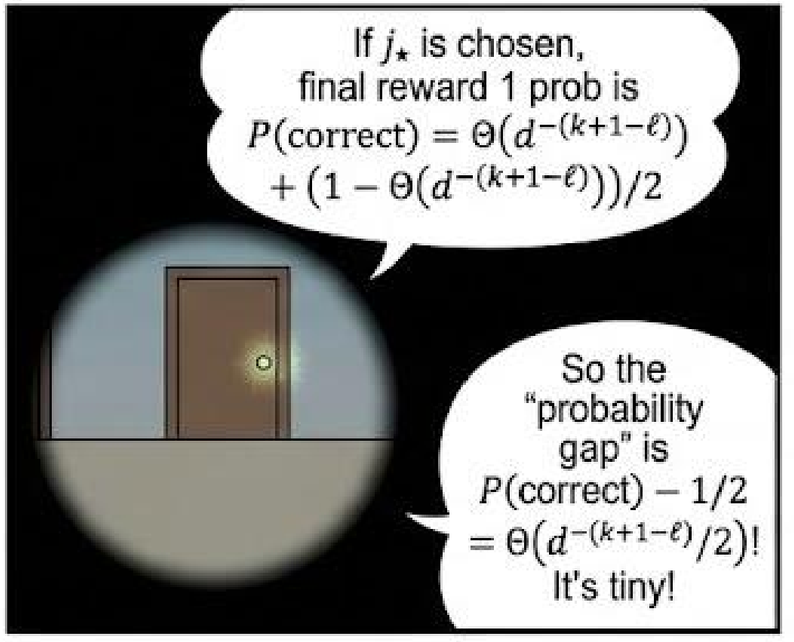
(b) Test Accuracy of Test-Time Scaling.

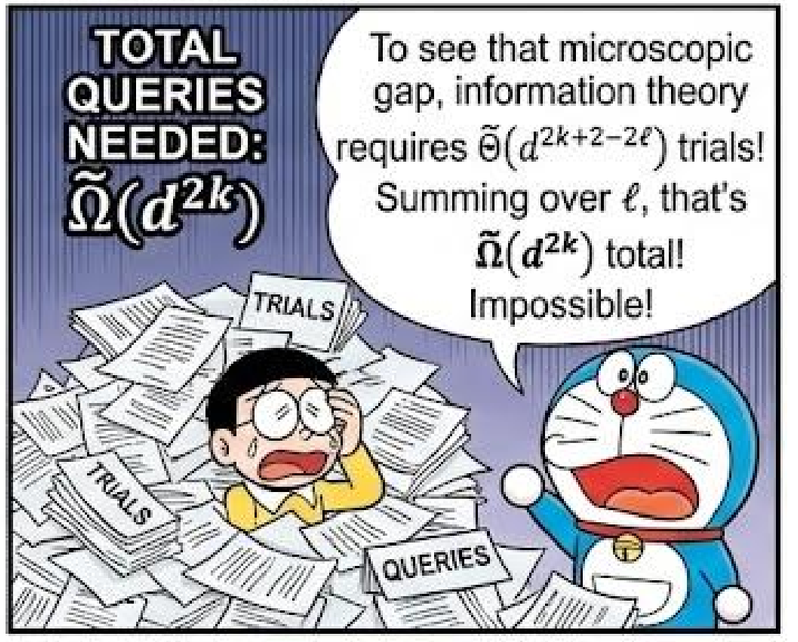


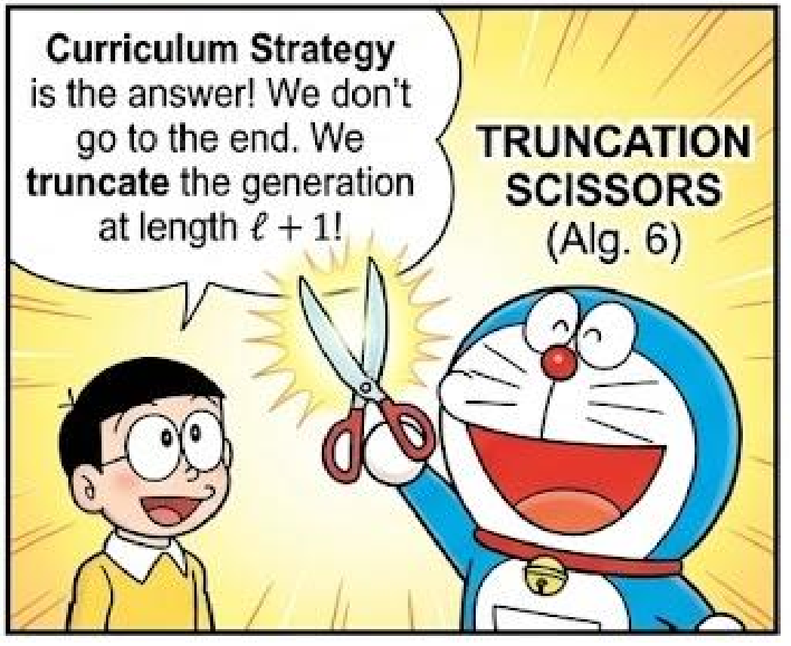
(c) Attention Heat Map of Finetuning.

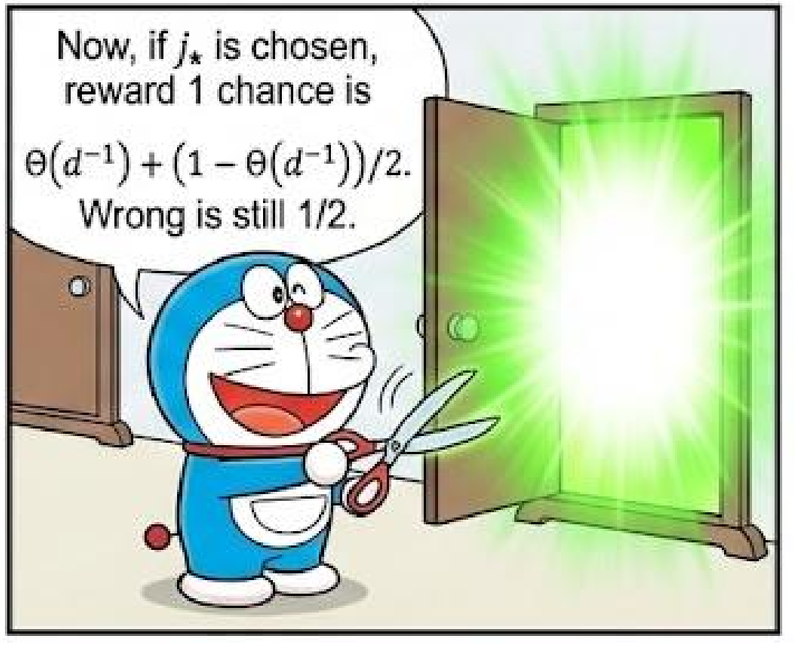
Figure 8. Post-training Dynamics on the parity task with error bar ($d = 8$, $k^\star = 8$, $S^\star = [0, 1, 2]$, $\eta_A = 10^8$, $\eta_B = \eta_C = 10^5$). (a) Test accuracy evolution of finetuning along sample complexity; (b) Test accuracy evolution of Test Time-Scaling along sample complexity; (c) averaged attention heatmap after finetuning.

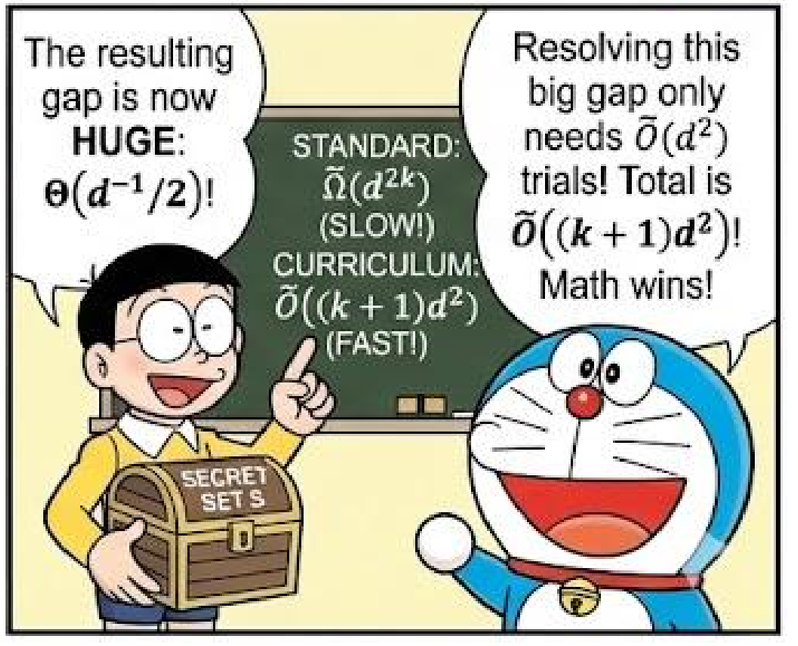

In standard learning, we go to the end. But a wrong set still gets "Reward 1" with probability 1/2! How do we distinguish?
REWARD 1 w/ Prob P(correct)
REWARD 1 w/ Prob P(wrong) = 1/2
STANDARD LEARNING (Go until EOS)
At step ℓ, both correct index $j_\star$ and wrong index j are picked with probability $\Theta(d^{-1})$.


If $j_\star$ is chosen, final reward 1 prob is $P(\text{correct}) = \Theta(d^{-(k+1-\ell)}) + (1 - \Theta(d^{-(k+1-\ell)}))/2$
So the "probability gap" is $P(\text{correct}) - 1/2 = \Theta(d^{-(k+1-\ell)}/2)$! It's tiny!


TOTAL QUERIES NEEDED: $\tilde{\Omega}(d^{2k})$
To see that microscopic gap, information theory requires $\tilde{\Theta}(d^{2k+2-2\ell})$ trials! Summing over ℓ, that's $\tilde{\Omega}(d^{2k})$ total! Impossible!
TRIALS
TRIALS
QUERIES

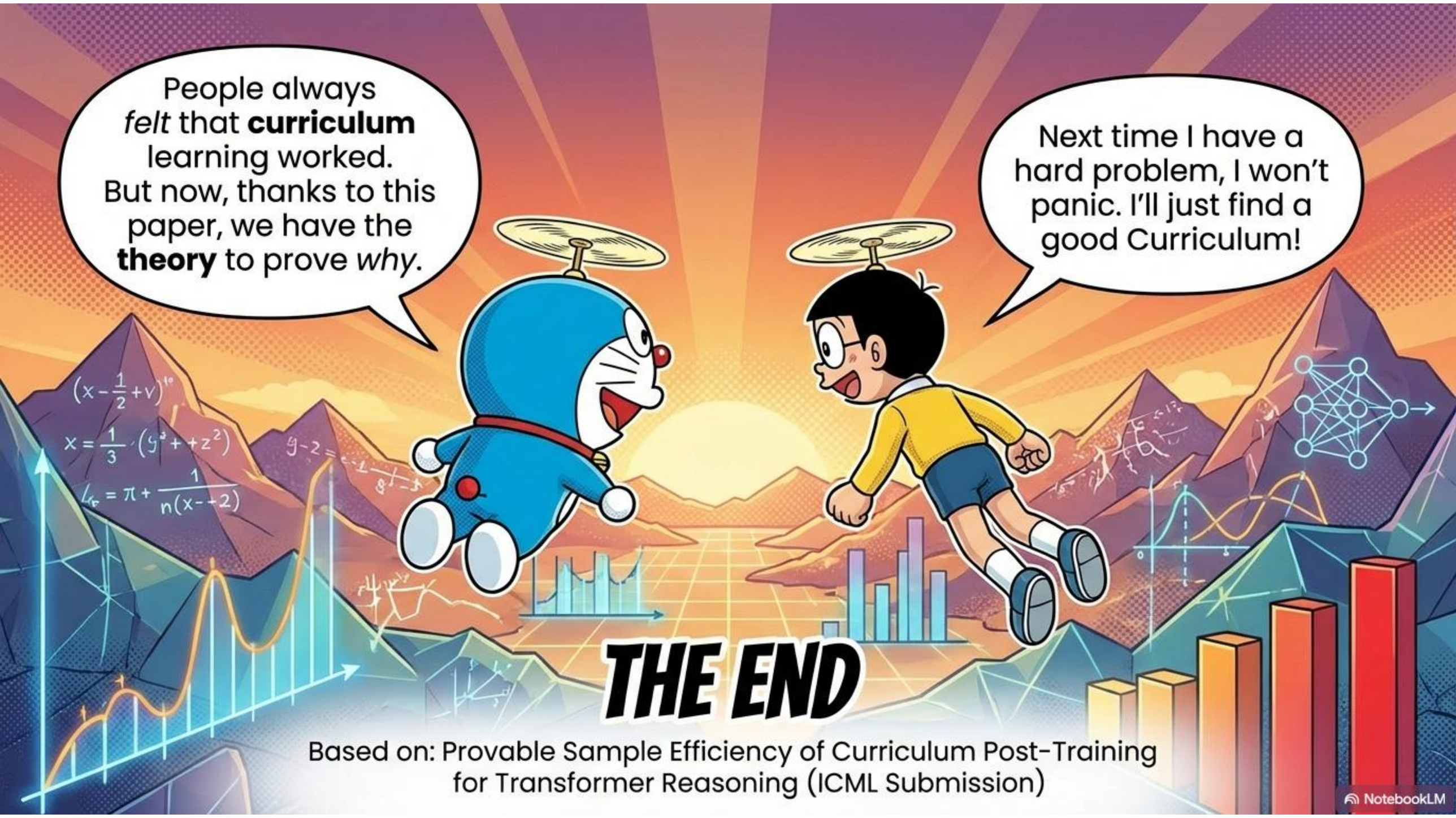

Curriculum Strategy is the answer! We don't go to the end. We **truncate** the generation at length $\ell + 1$!
TRUNCATION SCISSORS (Alg. 6)


Now, if $j_\star$ is chosen, reward 1 chance is $\Theta(d^{-1}) + (1 - \Theta(d^{-1}))/2$. Wrong is still 1/2.


The resulting gap is now **HUGE**: $\Theta(d^{-1}/2)$!
STANDARD: $\tilde{\Omega}(d^{2k})$ (SLOW!)
CURRICULUM: $\tilde{O}((k+1)d^2)$ (FAST!)
SECRET SET S
Resolving this big gap only needs $\tilde{O}(d^2)$ trials! Total is $\tilde{O}((k+1)d^2)$! Math wins!

GADGET SUMMARY CHECKLIST

1. **The Problem:** Hard reasoning has sparse rewards → Exponential Bottleneck. **[CHECK]**
2. **The Solution:** Decompose the task! Use Depth-Increasing or Hint-Decreasing gadgets. **[CHECK]**
3. **The Guarantee:** The Coverage Coefficient proves Polynomial Sample Complexity. **[CHECK]**
4. **The Result:** Works for both Post-Training (RL) and Test-Time Scaling. **[CHECK]**


People always *felt* that **curriculum** learning worked. But now, thanks to this paper, we have the **theory** to prove *why*.
Next time I have a hard problem, I won't panic. I'll just find a good Curriculum!
THE END
Based on: Provable Sample Efficiency of Curriculum Post-Training for Transformer Reasoning (ICML Submission)
NotebookLM