Learning and Controlling the Structure of Reasoning in Language Models: Toward Dynamic Ontologies

Soheil Feizi

University of Maryland, College Park

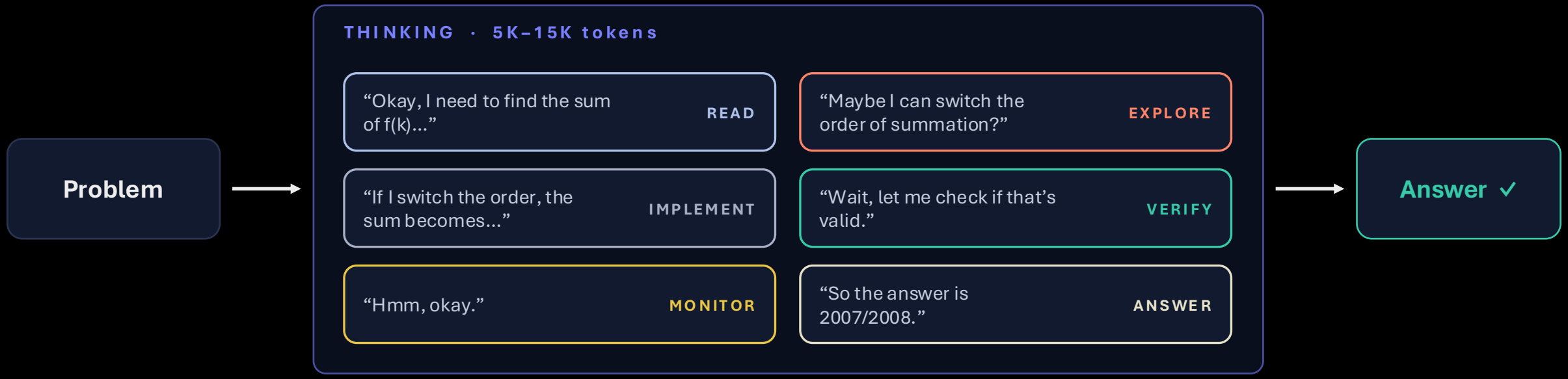
SOLAIR

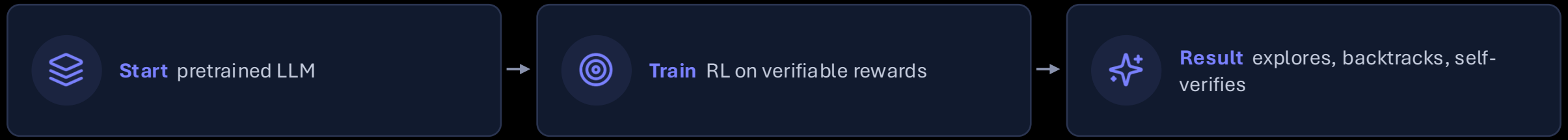
Two years of models that think before they answer



Thinking tokens became a new axis of scaling

How a reasoning model works





Models think ~10× longer. Do they think smart?



- Which path to take?
- When to stop?
- Where to spend compute?
- What is out of reach?

Capability ≠ structured and controllable reasoning

Can we expose the semantic structure underlying the reasoning process?

LLM reasoning through the SOLAIR lens

STRUCTURE



Meaningful states, and relations among them

Do reasoning traces have an interpretable state space?

UNCERTAINTY & DYNAMICS



Probabilistic relations that evolve with new evidence

Are reasoning dynamics stochastic — and do they change with learning?

INFORMATION & RESOURCES



Represent and act efficiently under constraints

Can we compress and control reasoning without losing what matters?

PART 1

Anatomy of a reasoning trace

Schoenfeld's Anatomy of Mathematical Reasoning by Language Models

Ming Li*, Chenrui Fan*, Yize Cheng*, Soheil Feizi, Tianyi Zhou

University of Maryland, College Park

ACL 2026 · Oral github.com/MingLiiii/ThinkARM

Accuracy says if a model got it right — not how it reasoned

Schoenfeld (1985)

think-aloud protocols → episodes of problem solving

failures: poor metacognitive control, not missing knowledge

What we measure

DeepSeek-R1 trace · 9,250 tokens

✓ correct

What ThinkARM sees



thin segments: Read 3% · Plan 6% · Monitor 2% · Answer 3%

Eight episodes, one label per sentence

Read
restate the problem

“The question asks for x …”

Analyze
theory, inference

“By the Pythagorean theorem …”

Plan
commit to a step

“Next, differentiate both sides.”

Implement
calculate

“Substituting x = 3: 2(3) + 5 = 11.”

Explore
tentative ideas

“Maybe try other values of x …”

Verify
check

“Double-check: 2·3 + 5 = 11 ✓”

Monitor
pauses

“Wait.” “Hmm.”

Answer
state the result

“Therefore, the answer is 24.”

Schoenfeld’s six + Monitor + Answer

Ontology lens: episodes ≈ semantic reasoning states

Every sentence gets a label

Question:

Let $f(r) = \sum_{j=2}^{2008} \frac{1}{j^r} = \frac{1}{2^r} + \frac{1}{3^r} + \cdots + \frac{1}{2008^r}$.

Find $\sum_{k=2}^{\infty} f(k)$

Categories:

1. Read	5. Explore
2. Analyze	6. Verify
3. Plan	7. Answer
4. Implement	8. Monitor

[1] Okay, so I have this problem here: I need to find the sum of $f(k)$ from $k = 2$ to infinity, where $f(k)$ is defined as the sum from $j = 2$ to 2008 of $1/(j^k)$. [1] So $f(k)$ is like $1/2^k + 1/3^k + \cdots + 1/2008^k$, right? And then I have to sum that from $k = 2$ all the way to infinity. [8] Hmm, okay, let me think about how to approach this.

[3] First, let me write down what exactly the problem is asking. [1] The sum they want is: $\sum_{k=2}^{\infty} \left[\sum_{j=2}^{2008} \frac{1}{j^k} \right]$. [2] So, that's a double sum. [2] The inner sum is over j from 2 to 2008, and the outer sum is over k from 2 to infinity. [5] Maybe I can switch the order of summation? [2] Because sometimes switching the order makes things easier. [3] Let me check if that's possible here.

[4] If I switch the order, the sum becomes: $\sum_{j=2}^{2008} \left[\sum_{k=2}^{\infty} \frac{1}{j^k} \right]$. Right? [2] Because for each j, we're summing over all k from 2 to infinity. So instead of summing over k first and then j, I sum over j first and then k. [6] That seems valid because all terms are positive, so by Tonelli's theorem, we can interchange the order of summation. [8] Okay, good.

[2] So now the problem reduces to computing, for each j from 2 to 2008, the sum from $k = 2$ to infinity of $1/(j^k)$.

[3] Then sum all those up.

[1547 More Tokens Here]

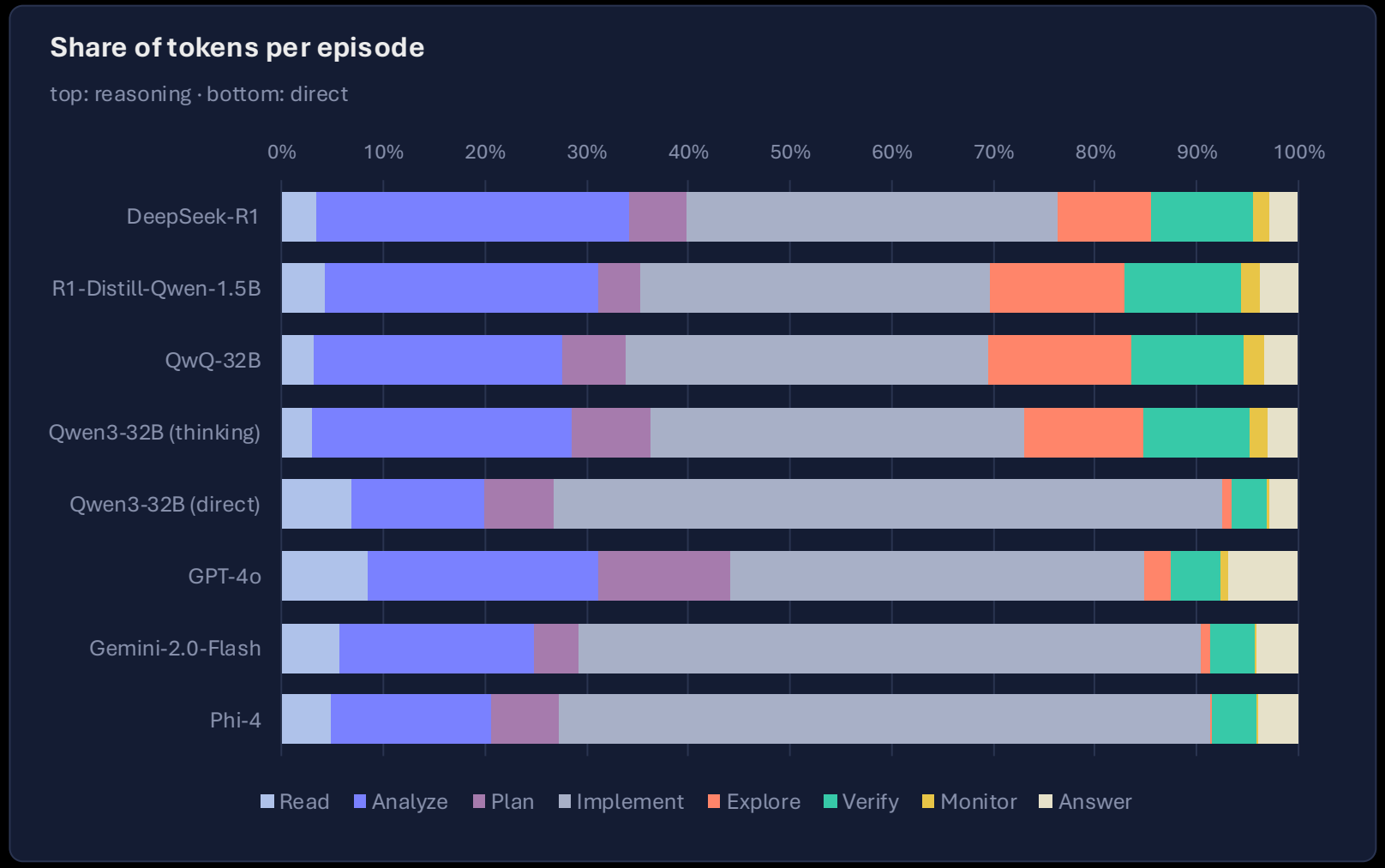
[6] So the telescoping works here. [6] Hence, the method is correct.

[7] Therefore, confident that the answer is 2007/2008.

[7] **Final Answer** $\boxed{\frac{2007}{2008}}$

15 models	100 Omni-MATH problems	410,991 sentences	κ = 0.83

Reasoning models spend the budget differently



66% → 37%

Implement: Qwen3-32B direct vs. thinking

<9% vs. 21–27%

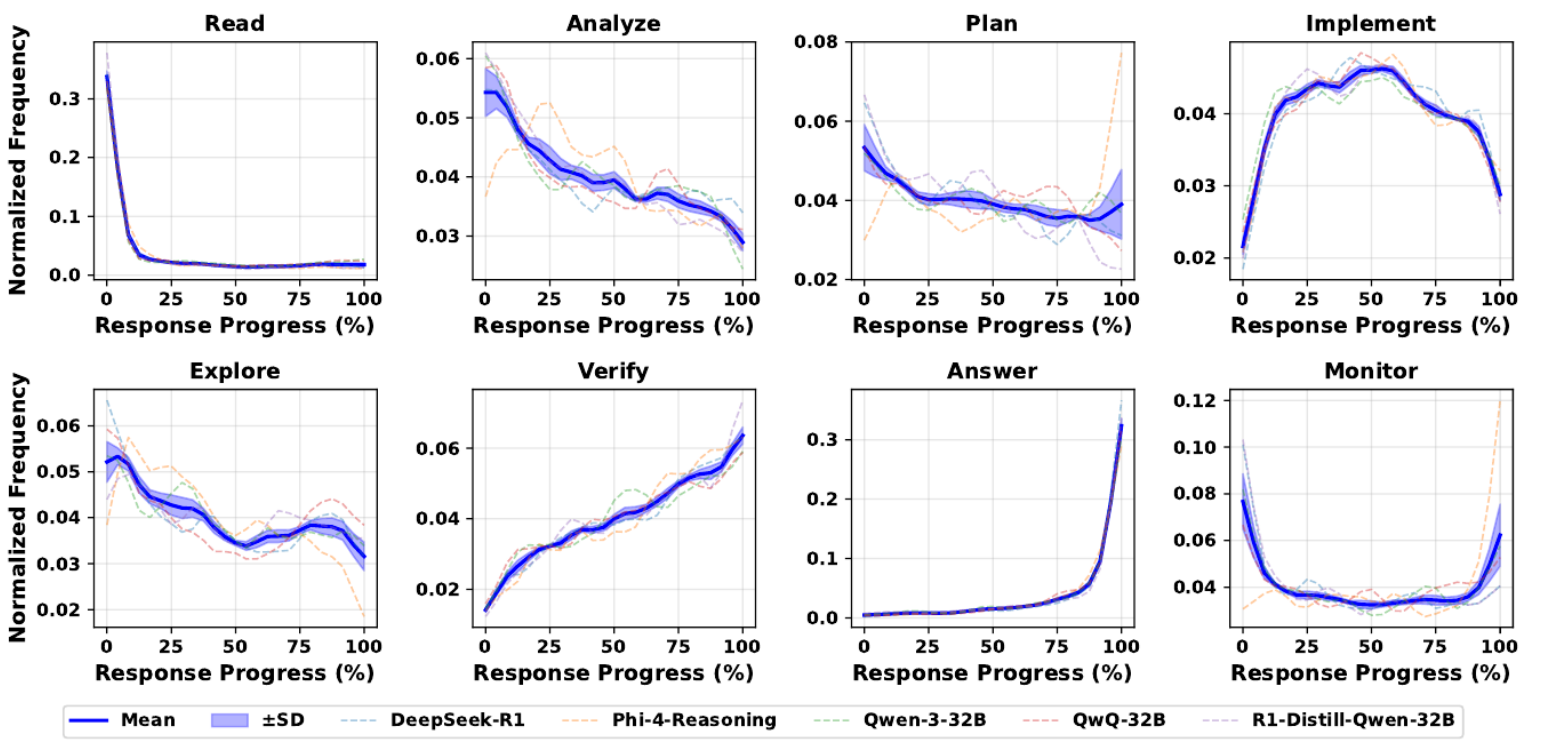
Explore + Verify + Monitor

1.5B ≈ R1

distillation preserves the profile

Research question: can distillation preserve not only outputs, but the ontology of reasoning — states, transitions, and uncertainty?

A three-phase “cognitive heartbeat”



episode frequency along the trace · mean ± SD, 5 reasoning models

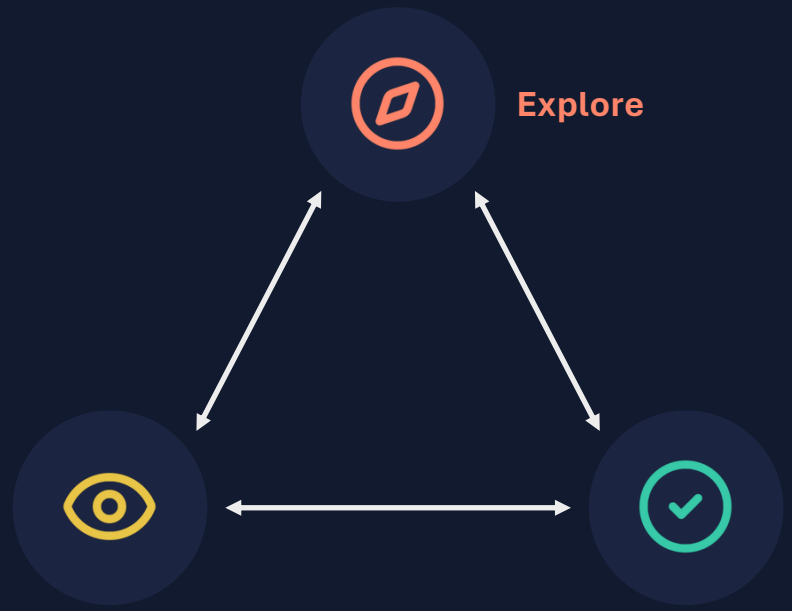
① Initialize
Read, Analyze, Plan decay

② Execute
Implement peaks mid-trace

③ Converge
Verify rises, Answer at the end

Reasoning structure evolves over time

Reasoning runs on Explore ↔ Monitor ↔ Verify loops



Reasoning trace vs. its final answer

top n-grams by mutual information

Exp → Mon	0.286	Exp → Plan	0.275
Ver → Exp	0.282	Exp → Ver	0.261
Mon → Exp	0.277		

Reasoning vs. direct models

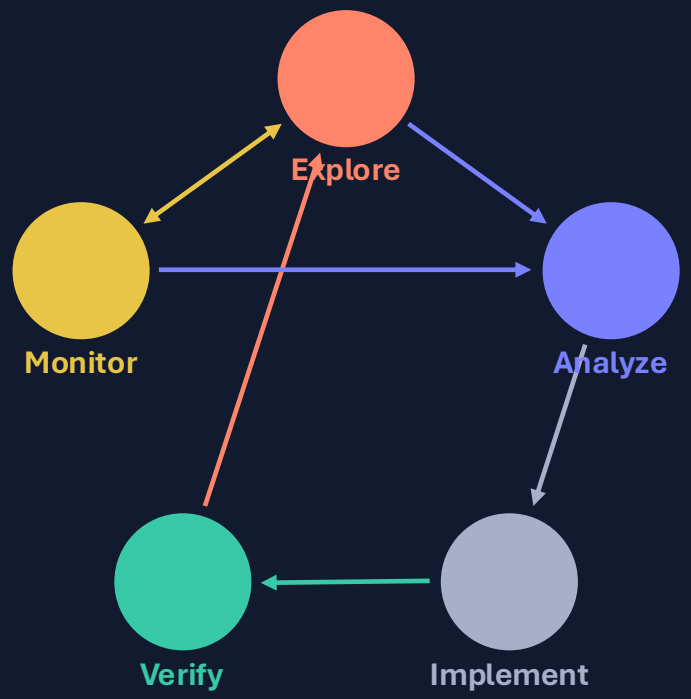
top n-grams by mutual information

Exp → Mon	0.251	Ana → Exp → Ana	0.203
Mon → Exp	0.241	Exp → Ana → Exp	0.203
Exp → Ana → Imp	0.207		

Relations are not deterministic: the next reasoning state depends on the current state, task, model, and context.

Direct responses: feed-forward toward Implement.

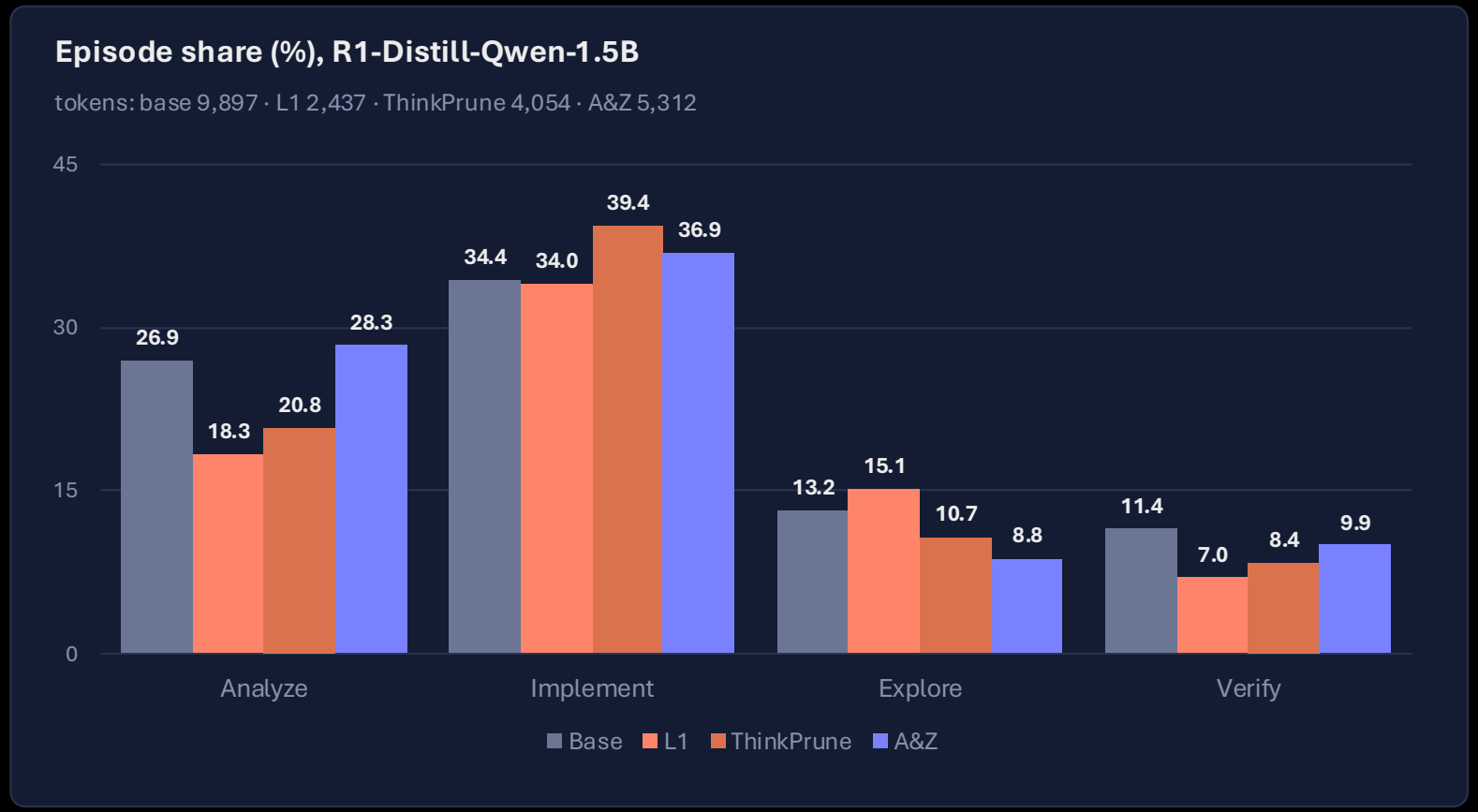
Toward a stochastic ontology of reasoning



State	$z_t \in \{$ Analyze, Explore, Implement, Verify, … $\}$
Dynamics	$P(z_{t+1} \mid z_t,$ task, model, context $)$
Uncertainty	confidence / entropy over reasoning state and outcome
Adaptation	the state space or transition dynamics may change with training, new evidence, or domain shift

Move from a fixed behavioral taxonomy to learned, probabilistic, and evolving reasoning ontologies.

Compression of reasoning can preserve — or distort — its structure



L1, ThinkPrune

cut Verify and Analyze; lose verification loops

A&Z

same profile, 46% fewer tokens

SOLAIR lens: equal token reduction can produce very different semantic distortion.

Same savings, different reasoning — **we return to this in Part 3.**

Can we define distortion over reasoning states and transitions, not only final accuracy?

Which transitions predict correctness?



Resolved exploration → correct

Explore → Monitor / Analyze

Unresolved exploration → wrong

more Explore; Explore → Verify / Answer

Not every transition carries equal semantic value.

PART 2

Steering reasoning as it unfolds

Off-the-Shelf LLMs as Process Scorers: a Training-Free Alternative to PRMs

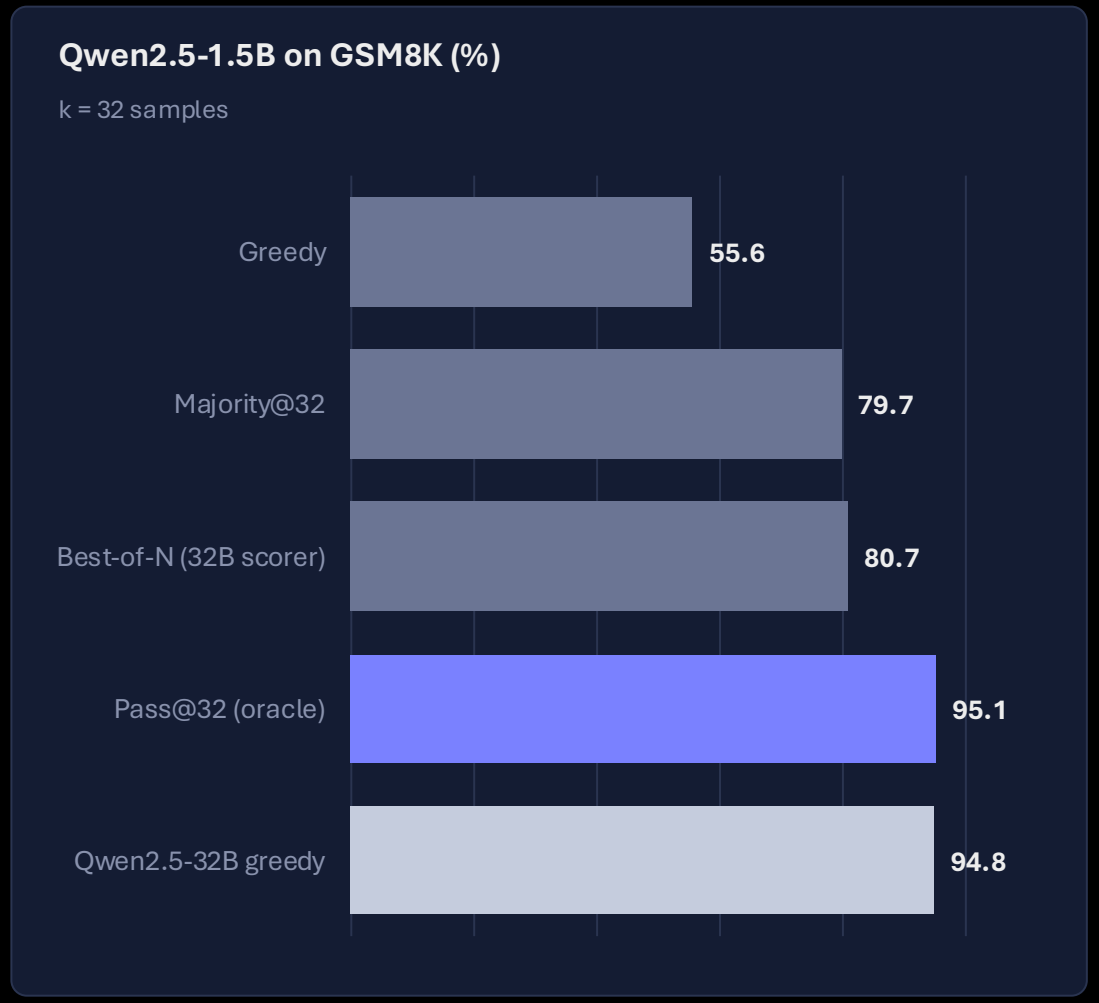
Atoosa Chegini, Soheil Feizi

University of Maryland, College Park

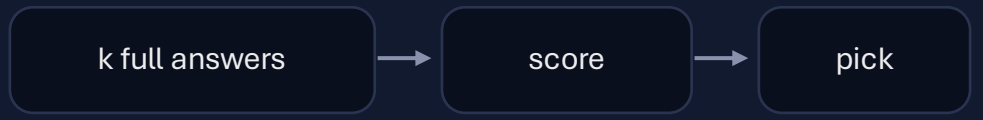
arXiv:2606.01682 · EMNLP 2026

SOLAIR lens: *controlling transitions in a structured reasoning process*

Small models can get there — selection is the bottleneck

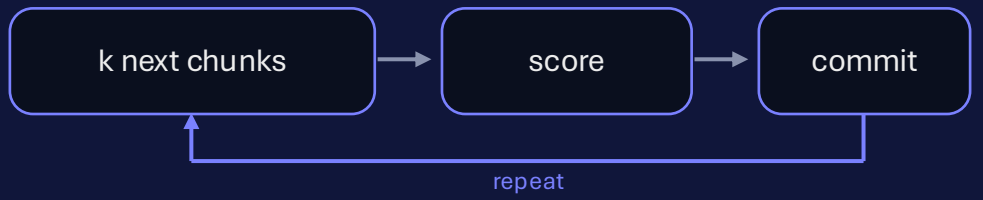


POST-HOC SELECTION



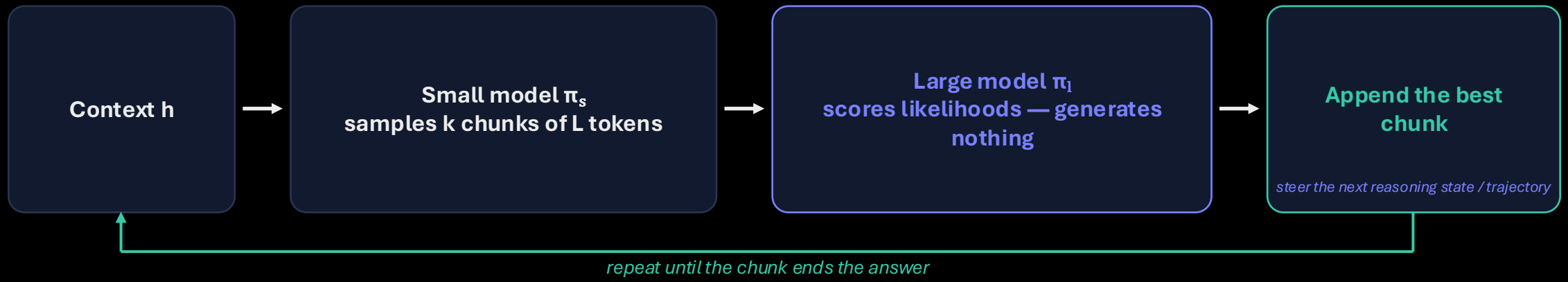
Too late: every candidate has already committed to its early mistakes. A strong scorer can only choose among finished paths.

GUIDED GENERATION



In time: steer before an error propagates. PRM (Process-reward-model) guided search does this, but needs a reward model trained on step labels.

An off-the-shelf LLM as the process scorer — no training



LGS · Likelihood-Guided Selection

$S(c) = \ell_l(c \mid h)$ the large model's length-normalized log-probability

CGS · Contrastive-Guided Selection

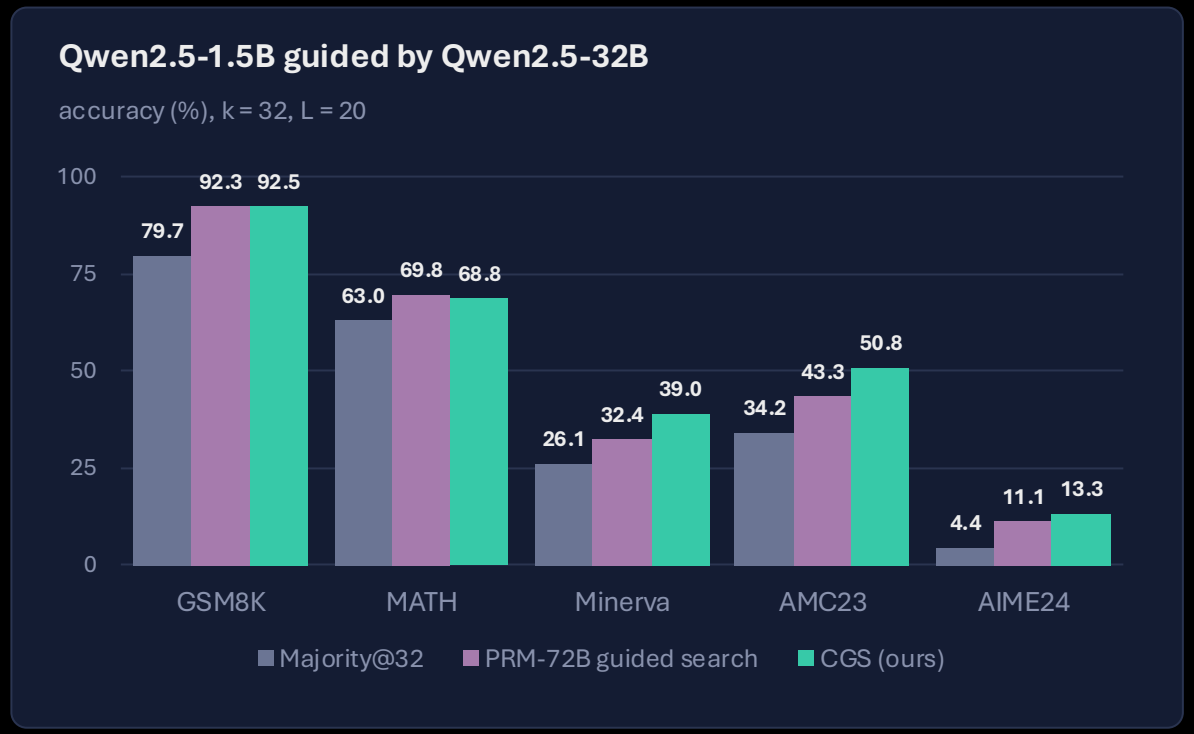
$S(c) = \ell_l(c \mid h) - \ell_s(c \mid h)$ where the large model's preference diverges from the small model's

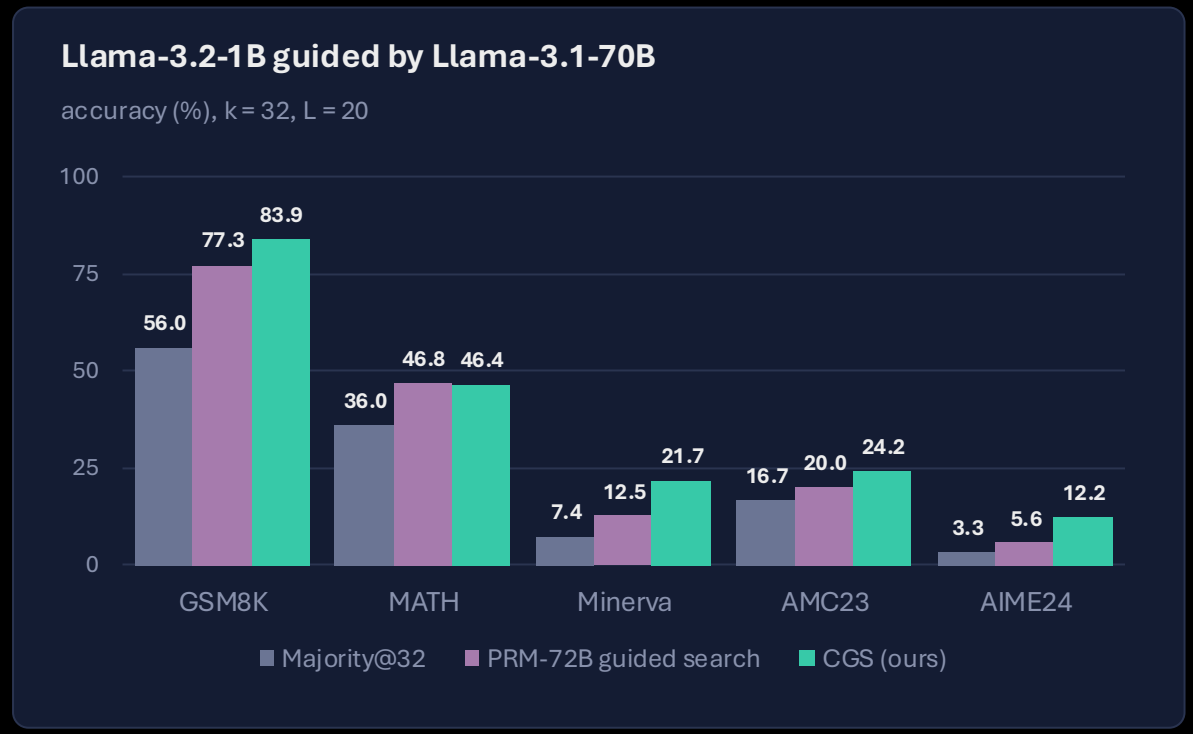
Post-hoc selection
no training · too late

PRM guided search
in time · needs a trained PRM

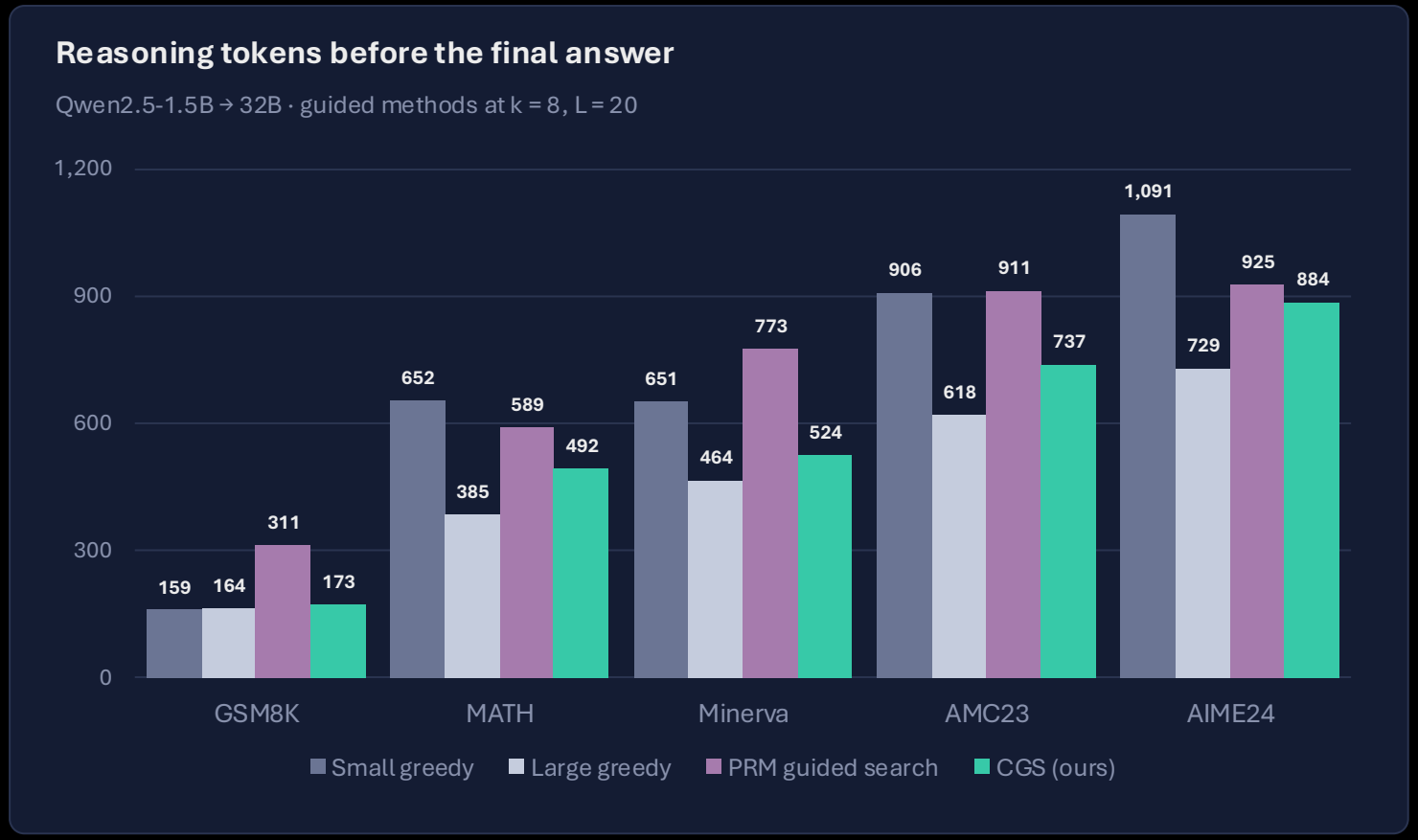
Chunk-level guidance (ours)
in time · training-free

Matches a trained PRM, beats voting by up to 28 points





Shorter traces than PRM guidance



PART 3

Learning to stop without learning to stop

Self-Supervised Confidence Training Improves Reasoning Efficiency

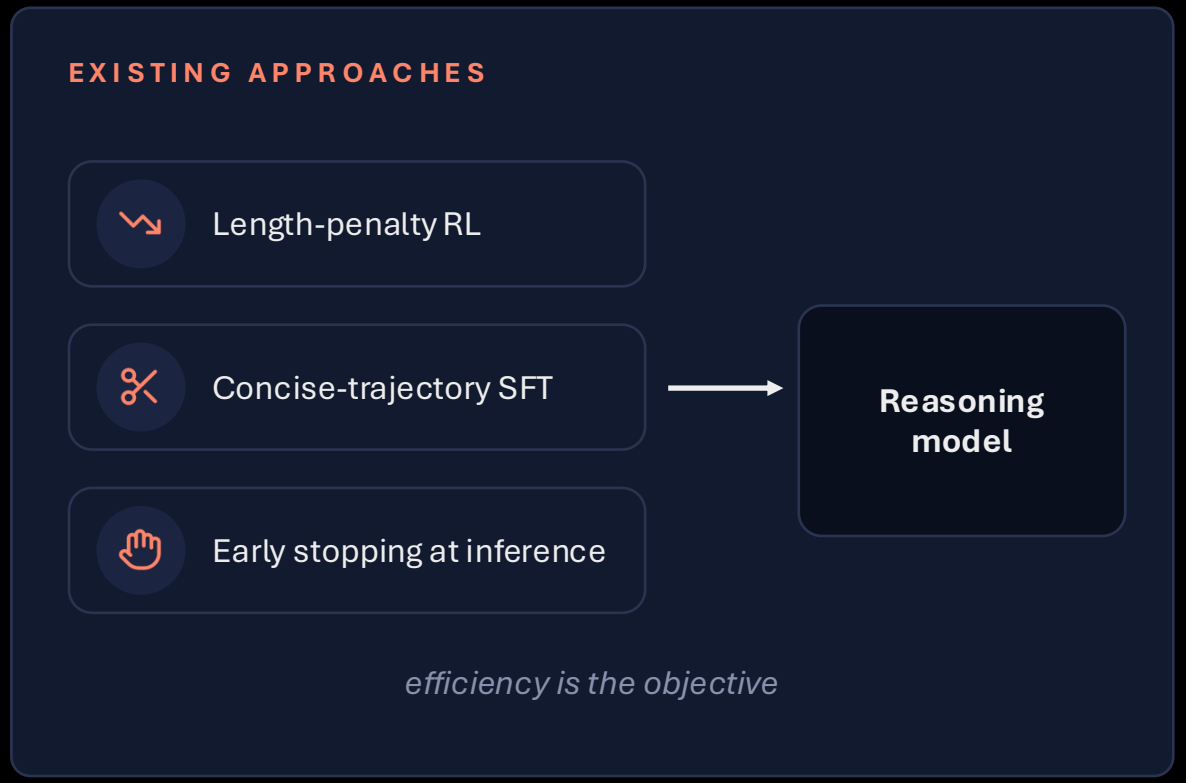
Parsa Hosseini, Akasha Tigalappanavara, Sumit Nawathe, Chenrui Fan, Sourya Basu, Genta Indra Winata, Anirban Das, Soheil Feizi, Nima Chitsazan

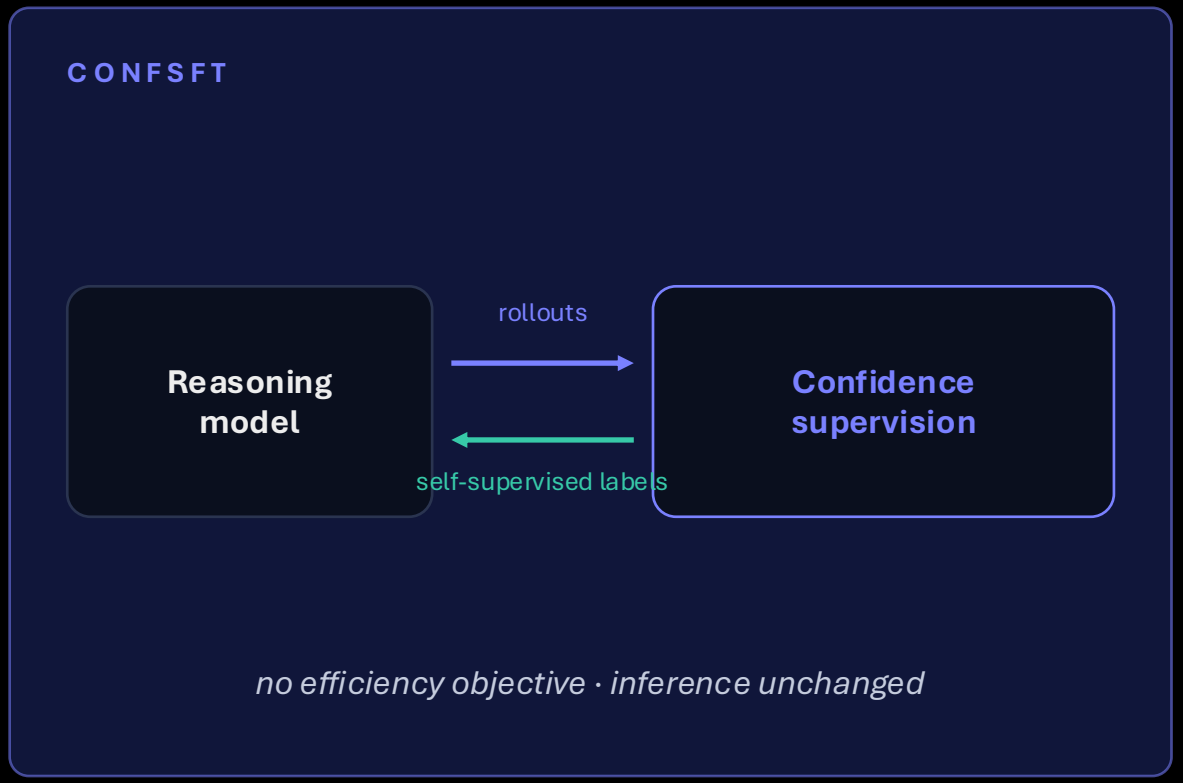
University of Maryland · AI Foundations, Capital One

arXiv:2609.31619 · September 2026

SOLAIR lens: *uncertainty as an internal signal for controlling reasoning dynamics*

Teach confidence, not efficiency





Can efficient reasoning emerge on its own — no bigger model, no efficiency objective?

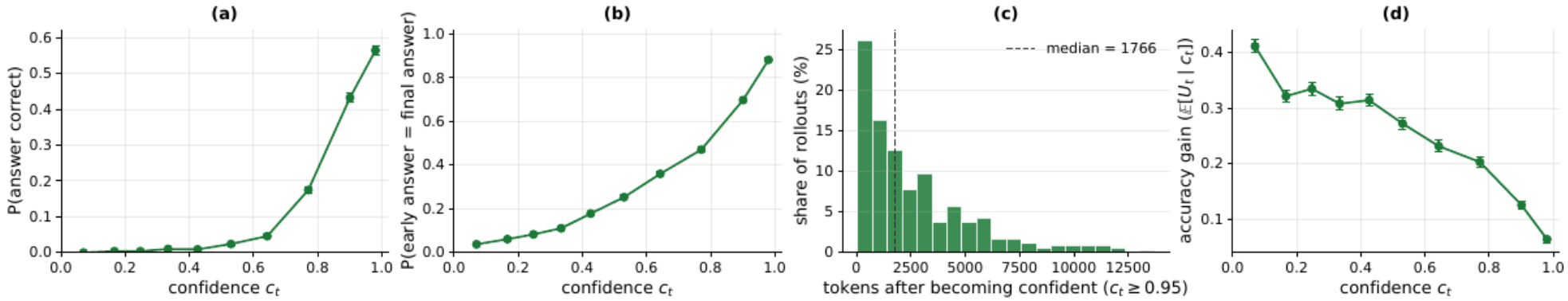
Confidence: a self-supervised signal about the reasoning state

PROBE AT STATE t

reasoning so far ... → "The final answer is \boxed{" → 5 5 → $c_t = (0.80 \cdot 0.65)^{1/2} = 72\%$

no gold answer, no judge

p = .80 p = .65



Nemotron-Nano-8B · 480 AIME-2024 traces

0 → 0.57
P(correct) vs. confidence (a)

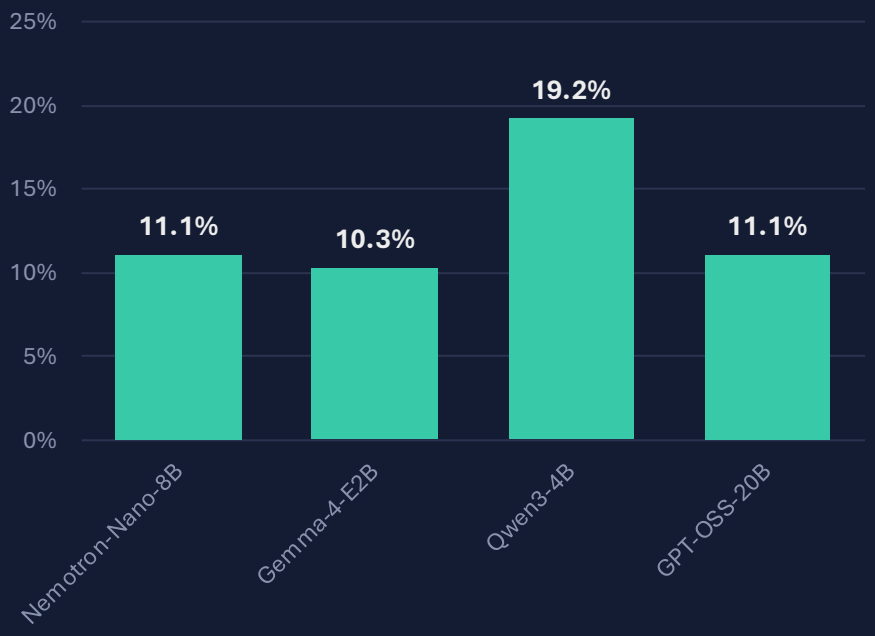
1,766 tokens
generated after $c_t \geq 0.95$ (median) (c)

0.41 → 0.05
gain from continuing (d)

Up to 25% fewer tokens at matched accuracy

Average token reduction (%)

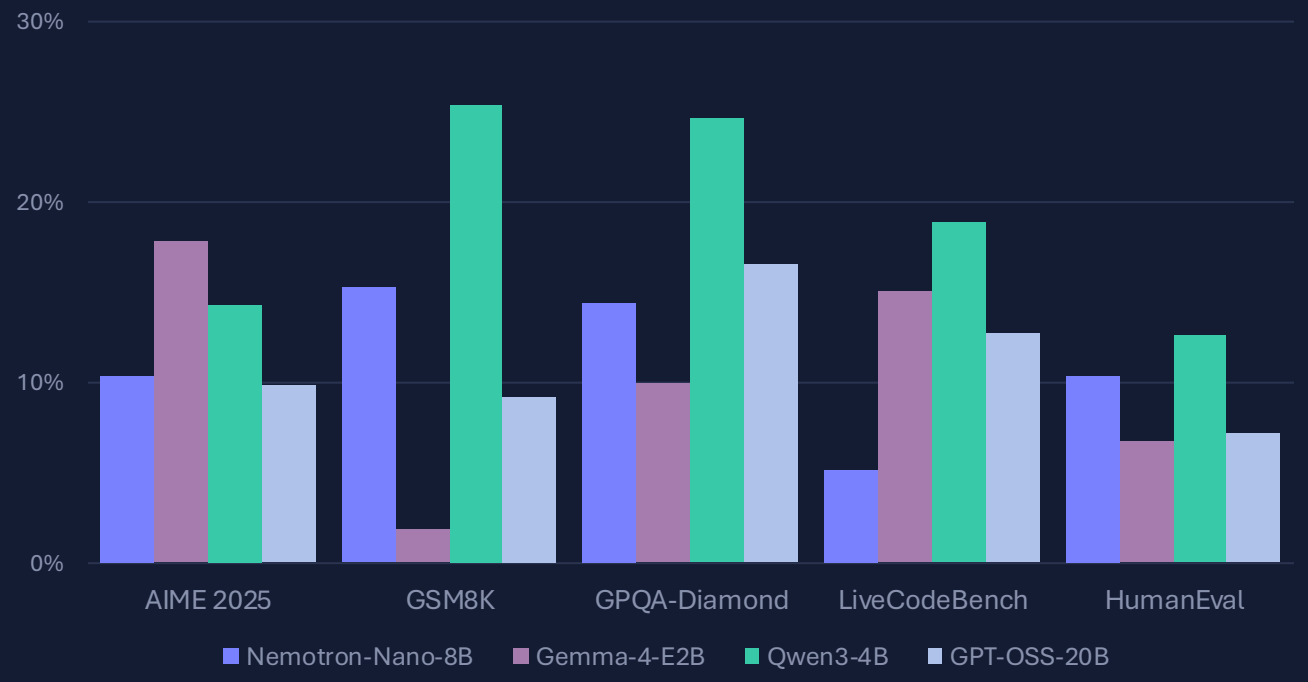
5 benchmarks



Δ accuracy: +1.1 / −0.7 / +0.1 / −0.6 (n.s.)

Token reduction by benchmark (%)

trained on AIME only

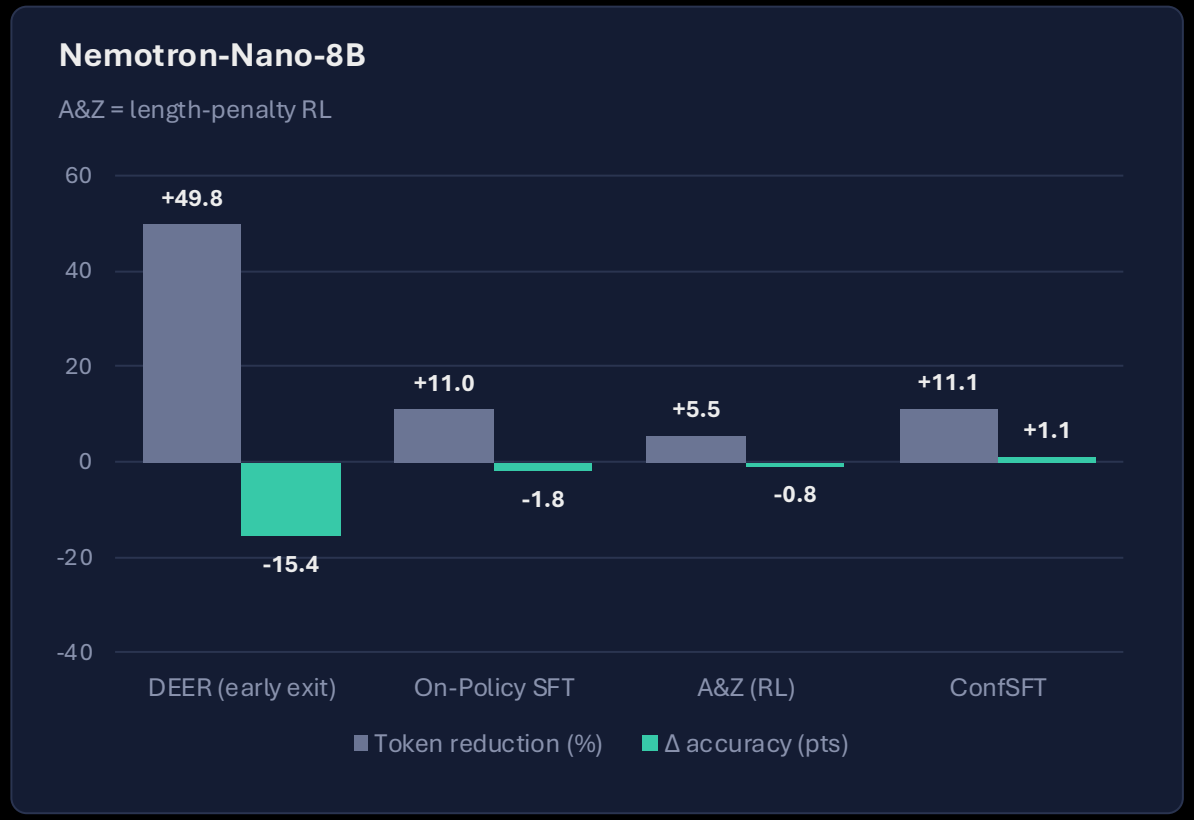


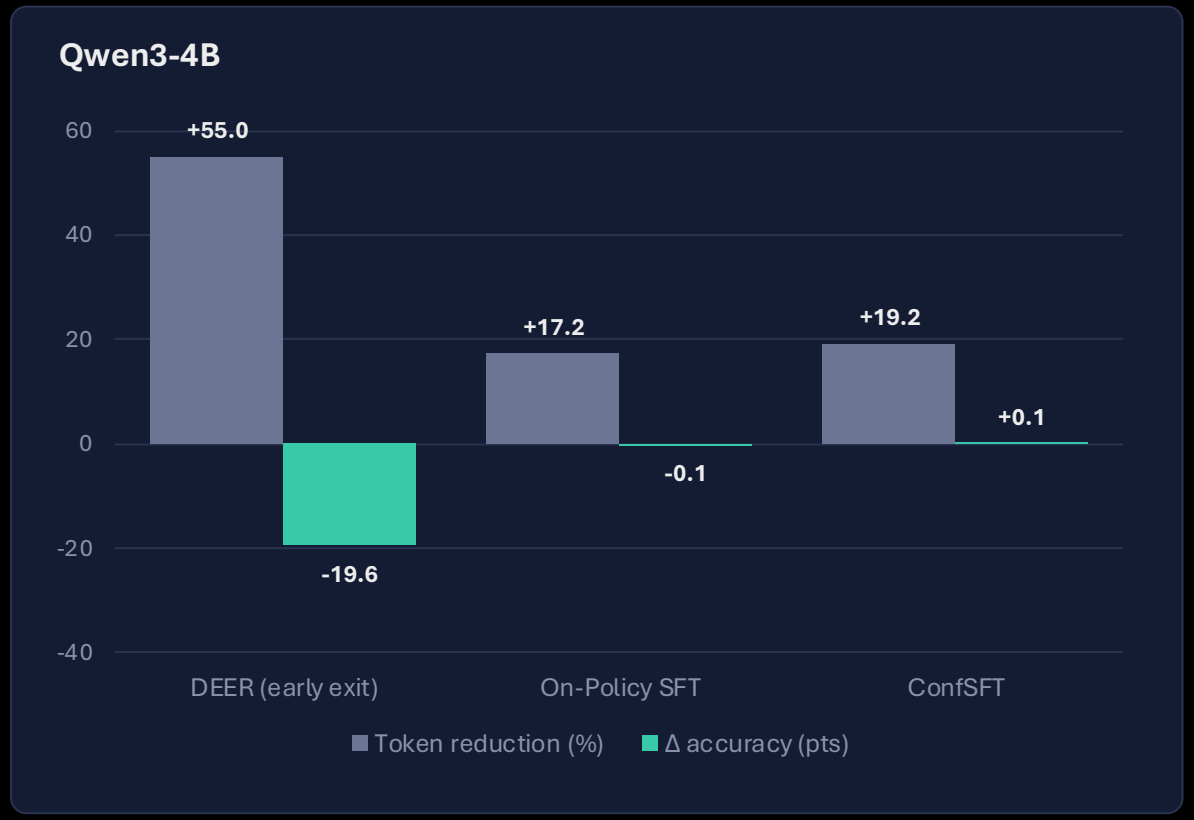
Transfers: math → science, code

Accuracy unchanged (95% CIs)

pass@8 preserved

On par with explicit efficiency training, safer than early exit





DEER: −15 points
code collapses (LCB 44 → 17)

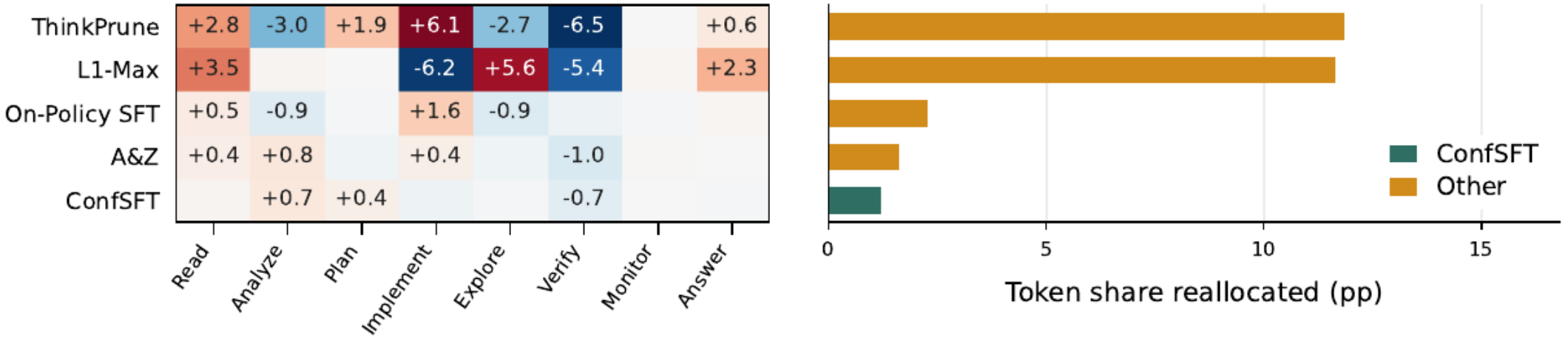
A&Z: 15× more data
smaller gains

ConfSFT
no efficiency supervision

Efficient without dismantling the structure

ThinkARM lens: change in episode share vs. each method's base policy · AIME 2025

Structure-preserving adaptation



≈ 12 pp

L1-Max, ThinkPrune

≈ 2 pp

On-Policy SFT and A&Z

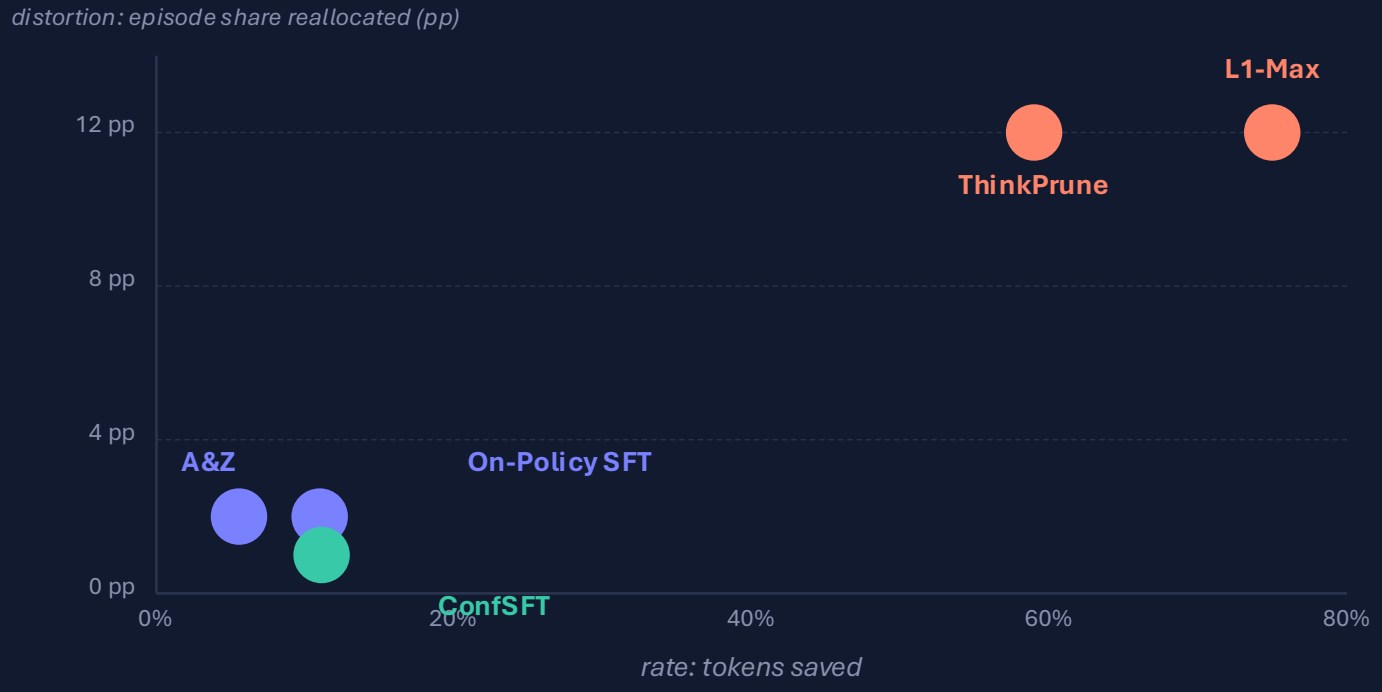
≈ 1 pp

ConfSFT: same mix, shorter

Future question: *can model compression / continual learning preserve the same semantic structure?*

A rate–distortion view of reasoning compression

Tokens saved vs. reasoning structure moved — one point per method



tokens saved as reported earlier in this talk (L1, ThinkPrune on R1-Distill-1.5B; On-Policy SFT, A&Z, ConfSFT on Nemotron-Nano-8B) · structure moved: previous slide

Rate

reasoning cost — tokens spent

Distortion

structure lost — here, episode mass reallocated; next: divergence over transition paths

Reading

ConfSFT near the knee: ≈ 11% fewer tokens, ≈ 1 pp moved. L1 and ThinkPrune save more, reshaping ≈ 12 pp.

SOLAIR's rate–distortion question, in LLM form: minimize semantic distortion of reasoning for a given rate.

PART 4

Thinking hard, not smart

Reasoning Models Fail to Ration Test-Time Compute Across Questions

Chenrui Fan*, Yize Cheng*, Ming Li, Yongyuan Liang, Tianyi Zhou, Soheil Feizi

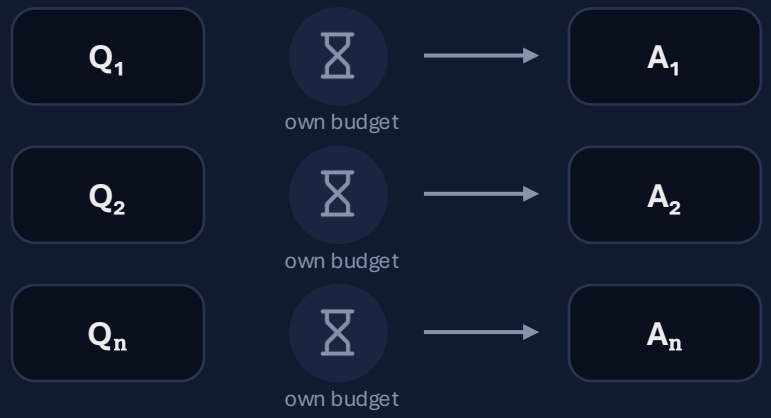
University of Maryland, College Park · MBZUAI

arXiv:2608.07968 · August 2026 **github.com/Fcr09/thinking-hard-not-smart**

SOLAIR lens: *allocating limited resources across competing reasoning processes*

A shared budget makes reasoning an allocation problem

TODAY: ONE QUESTION, ONE BUDGET



only decision: how long to think

OURS: N QUESTIONS, ONE BUDGET B



- which questions, in what order
- how much effort on each
- when to stop, skip, revisit

A knapsack: capacity B · value v_i · weight w_i = own solving cost, **unknown until you try**

An exam the model takes under budget pressure

Exams

- Omni-MATH, difficulty ≤ 5 (hidden)
- N ∈ {5, 10, 20} · 50 exams each
- shared budget B = 20K tokens

Perturbations

- scoring: fixed · random · aligned · reversed
- order: easy-first · random · hard-first
- prompt: base · plan · skip · recheck

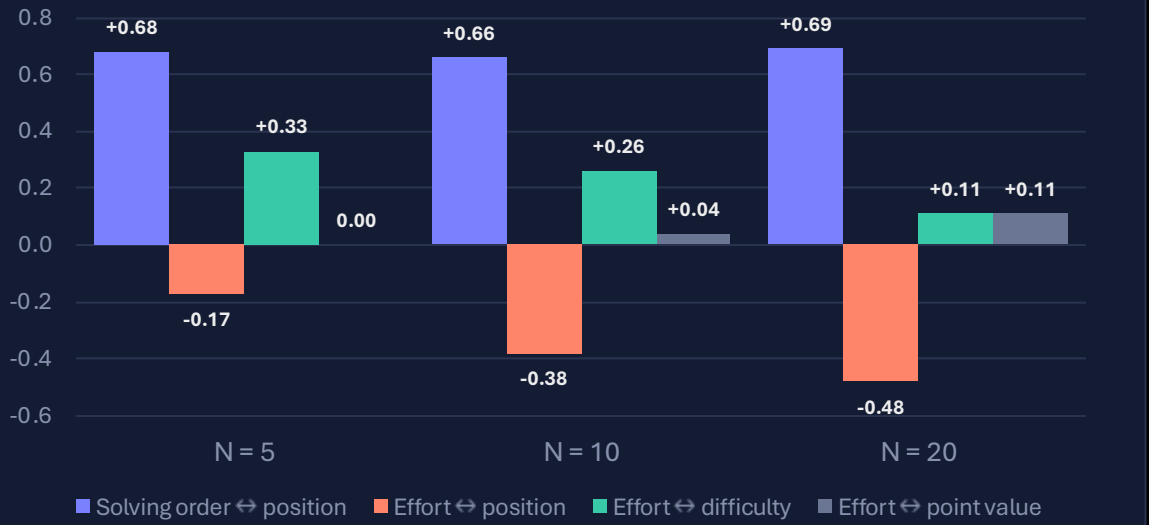
Models

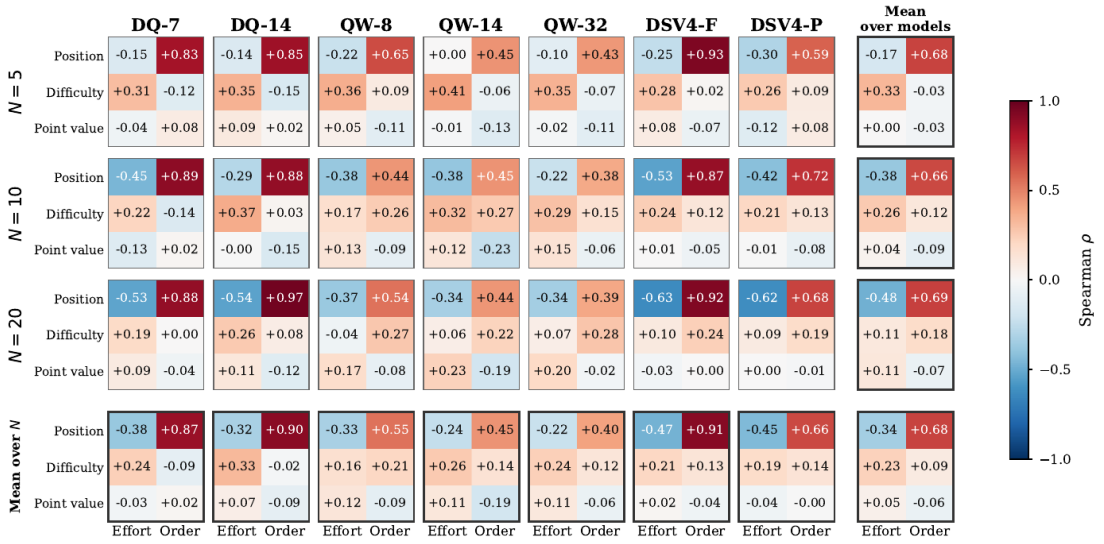
- R1-Distill-Qwen 7B / 14B
- Qwen3 8B / 14B / 32B
- DeepSeek-V4 Flash / Pro

Measures

- effort: tokens per question
- work set: ≥ 200 tokens
- solving order: token centroid

Allocation follows presentation order, not value





» **ρ ≈ +0.68**

solving order ≈ presentation order

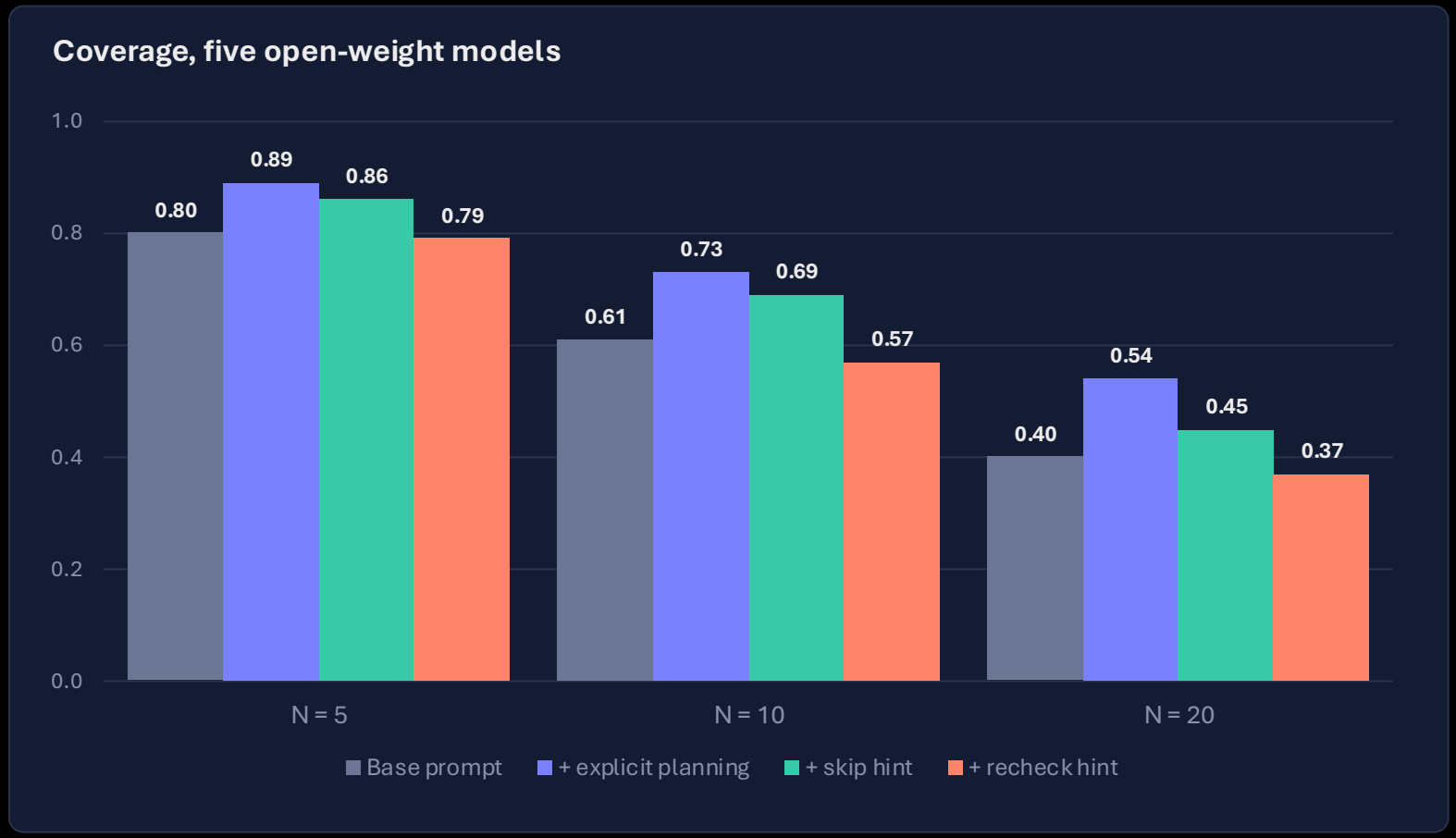
↘ **−0.17 → −0.48**

later questions starve

ρ ≈ 0

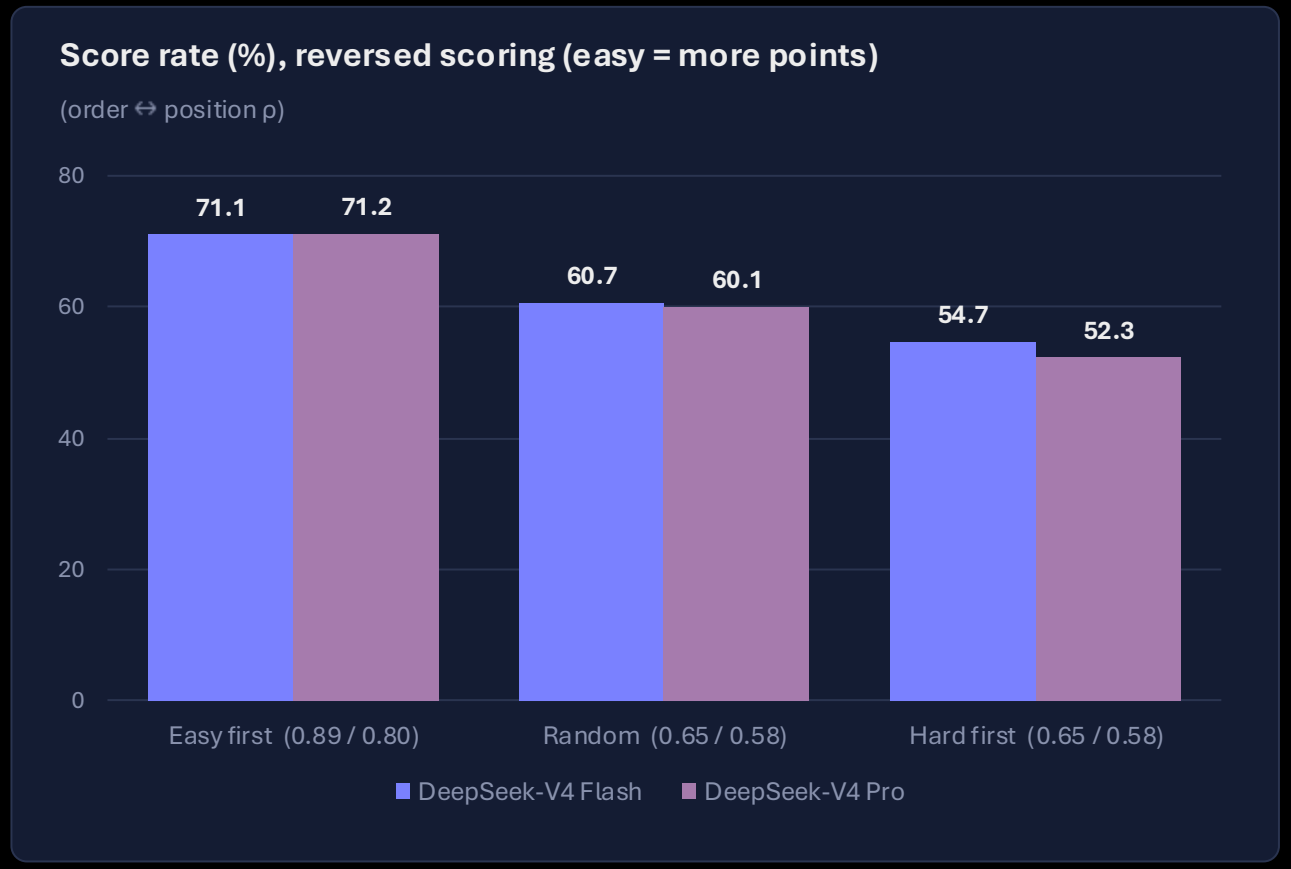
point values ignored

Planning prompts spread compute but do not prioritize



Told to budget, the model spreads compute — it does not prioritize.

Hijacked by hard questions placed first



Toward structured, adaptive, and controllable reasoning

DISCOVER



Reasoning has structure

ThinkARM → behavioral ontology

STEER



Its dynamics can be steered

chunk-level guidance

QUANTIFY



Uncertainty shapes it

confidence supervision

PRESERVE & CONTROL



Allocation is still open

Thinking Hard, Not Smart

SOLAIR: learn dynamic ontologies of reasoning — and keep them through adaptation and compression.