Reasoning Structure of Large Language Models

Measuring *how* models reason, beyond accuracy and token count

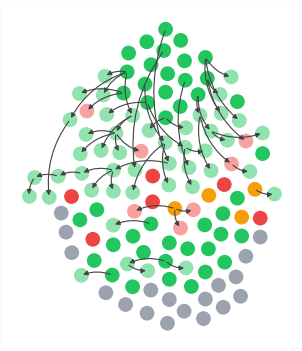
Frédéric Berdoz Luca A. Lanzendörfer Fabian Farestam Roger Wattenhofer

ICML 2026 Poster | PMLR 306

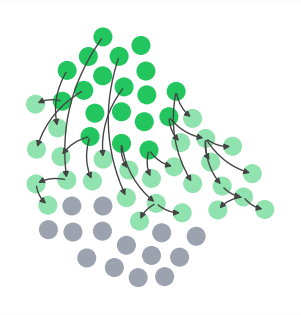
ETH Zurich

The same score can hide very different reasoning

- Reasoning models are judged by final-answer accuracy and token count.
- Both collapse behavior into a single number, and say nothing about *how* the model reasoned.
- Same model, same puzzle, **both correct**: one trace is diffuse, the other focused.



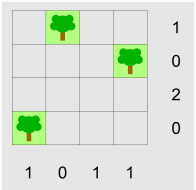
$\eta = 0.059$ (diffuse)



$\eta = 0.362$ (focused)

Nodes are atomic claims; edges are deductive steps. Green nodes are correct, red are wrong.

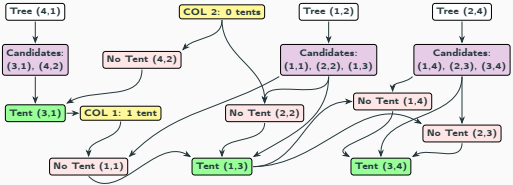
We make reasoning a measurable object



There is a tree at (4,1) and since it is in an edge, the only possible tent placements for the tree are (3,1) and (4,2). However column 2 has 0 tents, so (4,2) has no tent and (3,1) has a tent.

The only adjacent cells to the tree at (1,2) are (1,1), (2,2) and (1,3). Note that column 2 has no tents so (2,2) is no tent. Column one can only have 1 tent and we have already placed it, so (1,1) is no tent. Thus (1,3) has a tent.

The adjacent cells to the tree at (2,4) are (1,4), (2,3) and (3,4). Note that there is already a tent at (1,3) and since (1,4) and (2,3) are adjacent, they cannot be tents. So the tent must be at (3,4).



Puzzle instance (input) | Reasoning trace | Reasoning graph

- Convert each free-form trace into a verifiable reasoning graph: **nodes** = atomic claims checked against an executable environment, **edges** = single deductive steps. The result is a DAG.
- Built on a scalable benchmark of 21 grid puzzles $\times$ 4 difficulty levels, all fully verifiable.

A metric for the structure of reasoning

Reasoning-flow efficiency η: how *concentrated* a model's logical flow is, relative to the minimal set of claims that specifies the solution.

$$\eta = \frac{\log|V| - H_{\mathrm{str}}(G)}{\log|V| - \log|C^*|}$$

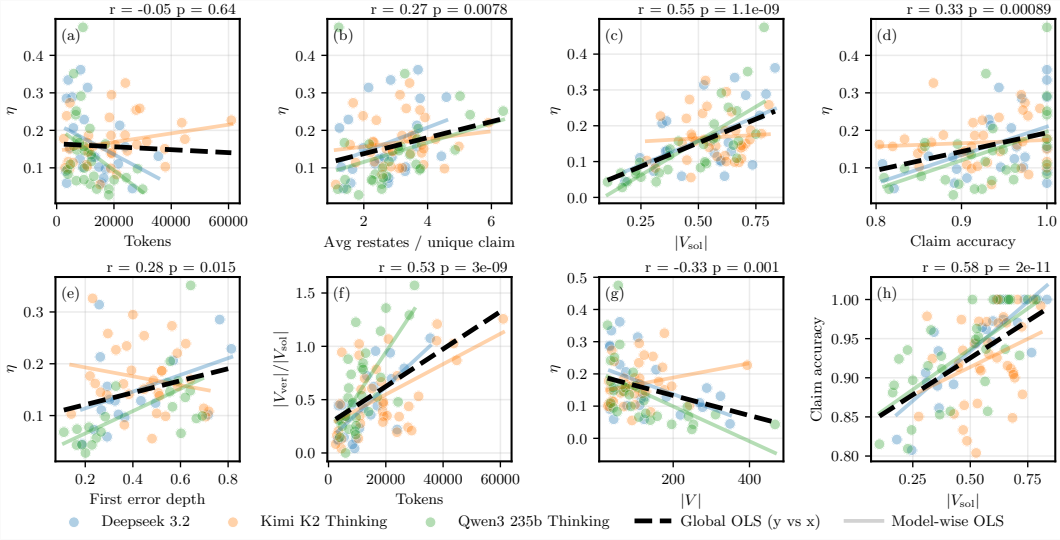
- $\eta = 1$ for all in weights deduction; low η for diffuse exploration and bloat.
- Built from the structural entropy H_{str} of an absorbing Markov chain over the graph, and *size-normalized* so it does not just track puzzle difficulty.

Token count is *not* a proxy for reasoning quality

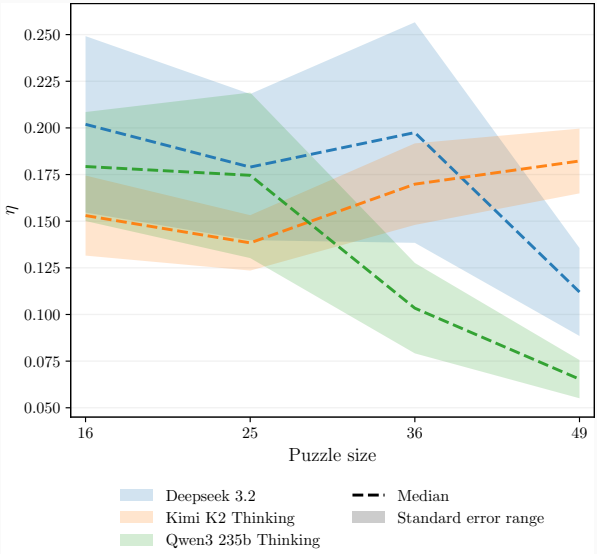
- Verbosity says nothing about structure: η is essentially uncorrelated with token count ($r = -0.05$).
- Extra tokens mostly buy verification overhead, not better reasoning ($|V_{\text{ver}}|/|V_{\text{sol}}|$ grows with tokens, $r = 0.53$).
- η is the only metric that rises with accuracy while staying independent of size and length; graph width, $|V|$, and tokens just track difficulty.

Across all models, accuracy also collapses on the hardest puzzles despite far larger token budgets; more compute does not close the gap.

Efficiency tracks reasoning quality, not length



η exposes structure that accuracy hides



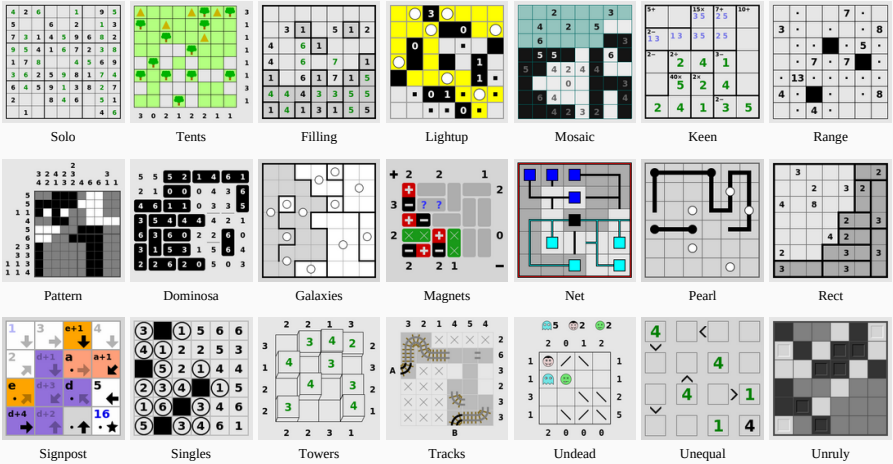
- Restricted to puzzle sizes where every instance is solved, so accuracy is saturated and uninformative.
- η still separates models and exposes diverging scaling behavior: substantial variation in *how* solutions are produced that accuracy alone hides.

Takeaways

- Reasoning traces $\rightarrow$ verifiable reasoning graphs; η measures how concentrated the logical flow is.
- Accuracy and length conflate behaviors that structure cleanly separates.
- The structural layer is domain-agnostic and extends to math (symbolic checks), code (unit tests), and agentic tool use; a candidate shaping reward for post-training.

Code & benchmark: `github.com/ETH-DISCO/llm-reasoning-efficiency`

A scalable, verifiable benchmark



21 grid puzzles (Simon Tatham's collection), 4 difficulty levels each, in an executable environment.

Accuracy collapses despite more compute

	Trivial		Human hard	
Model	Acc. (%)	Tokens	Acc. (%)	Tokens
GPT-5	83.8	4.2k	5.7	19.9k
DeepSeek V3.2	77.1	7.7k	0.0	36.8k
Kimi K2	77.1	10.6k	1.0	61.3k
Qwen3 235B	69.5	10.3k	0.0	23.6k

From Trivial to Human hard, accuracy collapses to near zero while token budgets balloon. Kimi K2 spends the most yet scores lowest.