Misinterpreting Model Failures

The Illusion of the Illusion of Thinking A Comment on Shojaee et al. (2025)

- Shojaee et al. claimed that model accuracy drops to near-zero when problem complexity exceeds a threshold. Opus and Lawsen argue this is a misinterpretation caused by flawed experimental design.

- In Tower of Hanoi tasks with many disks, models often terminate their own outputs saying things like:

"The pattern continues, but to avoid making this too long, I'll stop here."

This shows awareness of output length constraints—they're not failing to reason, they're choosing to stop due to token limits.

Shojaee et al.'s evaluation method doesn't distinguish between:

Failure to reason & Failure to output everything due to constraints

This causes false negatives: models get scored as wrong even when they've reasoned correctly but couldn't finish printing.

- The River Crossing benchmark includes problems with $N \geq 6$ agents and a boat of capacity 3.

  But it's mathematically proven (Efimova 2018) that these have no solutions.

  Yet models are penalized for not solving them : essentially punished for being correct.

- Changing Output Format Solves the Problem

  When models are instead asked to write a Lua function to solve Tower of Hanoi with 15 disks (i.e., output a recursive algorithm, not move-by-move text):

  All models succeed in < 5,000 tokens.

  Reasoning intact, output constraint avoided.

- Problem Complexity Is Not Just Solution Length

  Shojaee et al. equated complexity with the length of the solution.

  But complexity depends on: Branching factor, Search depth

  Constraint satisfaction

  Tower of Hanoi: long but mechanically trivial (deterministic, recursive)

  River Crossing: short but combinatorially hard (search required)

Shojaee et al. used minimum number of moves required to solve a puzzle as a measure of complexity.

They call this compositional depth.

But Opus and Lawsen argue: This metric confuses "length of output" with "difficulty of thinking."

Puzzle	Solution Length	Branching Factor	Search Required
Tower of Hanoi	2^N – 1 (very long)	1 (only one move at each step)	No — just follow the recursive rule
River Crossing	~4N (much shorter)	> 4 (many move options)	Yes — must search for legal sequence
Blocks World	~2N	O(N^2)	Yes — often PSPACE-hard