

A 150M-Parameter Model That Reasons for Less Than a Cent

Kabui, Charles

2026-08-19

[Read at ToKnow.ai](#)

A 150M-Parameter Model That Reasons for Less Than a Cent

Thinking in a hidden state, at a fraction of a cent per task

29.5%

pass@2 on the public
ARC-AGI-1 evaluation

\$0.0007

computed inference cost
per task solved

11x

cheaper per task than
GPT-5.6 Luna (Low)

August 19, 2026

ToKnow.ai

Pathway, the Palo Alto lab behind the Dragon Hatchling architecture, reports that its 150-million-parameter model BDH-CQ scores 29.5% on ARC-AGI-1, the public reasoning benchmark from the ARC Prize Foundation, at a computed cost of \$0.0007 per task, under one-tenth of a cent. That is roughly 11 times cheaper per task than OpenAI's GPT-5.6 Luna, which scores 34.2%, even after Luna's July price cut. The trick is architectural: instead of generating

a text chain of thought, the model reads example demonstrations into a recurrent memory, solves the query by iterating over a hidden state, and decodes only the final answer. An independent black-box audit by Bielik and NYU reproduced the 29.5% score without access to the weights.

The practical point is what thinking costs. Every reasoning step a frontier model emits is a token it must generate and pay for, so one ARC task can cost a large model real money. BDH-CQ's whole pipeline fits in about 0.85 GPU-seconds on an H200. The model is not a general rival yet, and its own paper shows it still collapses on ordering and nesting tasks. But on cost per unit of reasoning, a small model that never verbalizes its thinking undercuts frontier APIs by an order of magnitude.

This is the strongest evidence yet that reasoning does not have to be spoken to be real. If latent-state thinking scales the way Pathway's early tests suggest, with Transformer-like scaling from 1B to 600B parameters, the arms race to write ever-longer chains of thought has a quieter, cheaper alternative.

Read More: [A 37-Author Survey Says Latent Space, Not Language, Is Where AI Should Think](#)

Sources:

- [BDH-CQ: In-Context Learning with Recurrent Latent Reasoning \(arXiv\)](#)
- [Pathway's 150M-Parameter Model Breaks the ARC-AGI-1 Cost-Efficiency Frontier \(Pathway\)](#)
- [BDH-CQ paper page \(Hugging Face\)](#)
- [arc-task-gen: ARC-AGI-1 task generator \(GitHub\)](#)
- [The Dragon Hatchling: The Missing Link between the Transformer and Models of the Brain \(arXiv\)](#)

***Disclaimer:** For information only. Accuracy or completeness not guaranteed. Illegal use prohibited. Not professional advice or solicitation. **Read more:** [/terms-of-service](#)*