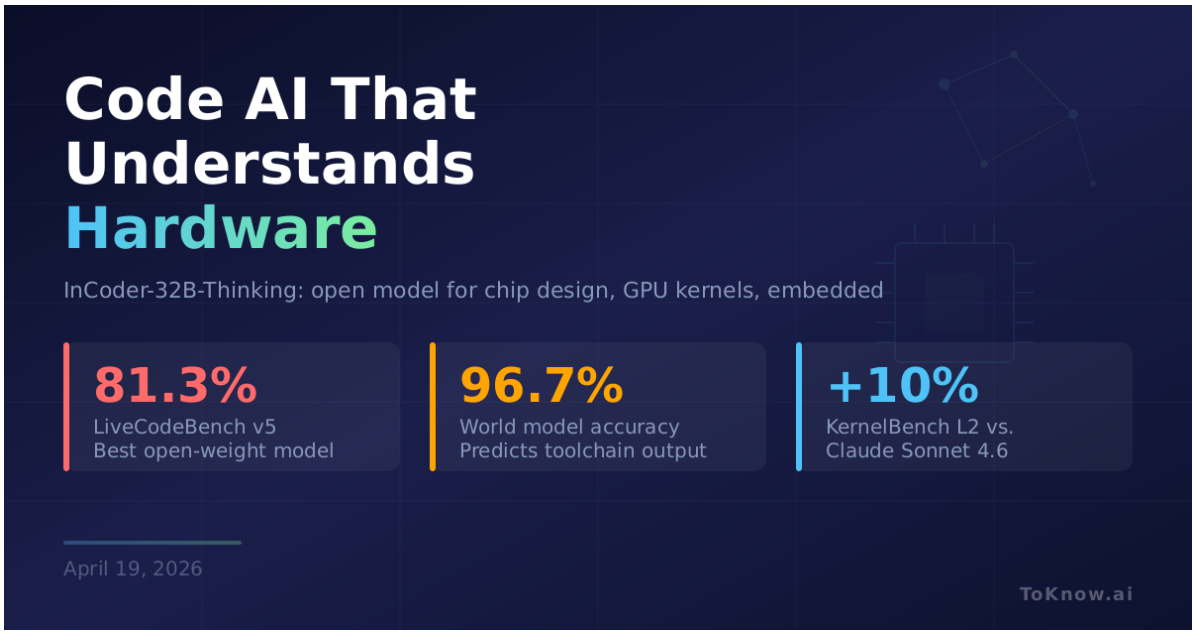
InCoder-32B-Thinking: A Code Model That Understands Chips, GPUs, and Embedded Hardware

Kabui, Charles

2026-04-19

Read at ToKnow.ai



A team of 26 researchers released InCoder-32B-Thinking, a 32B-parameter open-weight code model targeting domains current code LLMs handle badly: Verilog for chip design, CUDA kernels for GPU performance, embedded firmware, and 3D CAD modeling. These areas lack

the human reasoning traces web and Python models train on. Two pieces fix this. Error-driven Chain-of-Thought generates reasoning by running multi-turn dialogues against real toolchain feedback, teaching the write-compile-error-fix loop hardware engineers actually follow. The Industrial Code World Model, trained on traces from Verilog simulation, GPU profiling, and compiler diagnostics, predicts execution outcomes with 96.7% accuracy, so synthetic training data can be validated without running the toolchain. It scores 81.3% on LiveCodeBench v5, 38.0% on KernelBench L2 (vs. 28.0% for Claude Sonnet 4.6), and 70.4% on SWE-bench Verified.

Most AI coding tools target web apps and Python. Hardware and systems code, running trillions of dollars of silicon, GPUs, and embedded devices, has been mostly untouched because no one had a cheap way to generate training data for it. The world-model trick changes that: the model predicts what a Verilog simulator or CUDA profiler would output, so the team only spot-checks. For a small chip team or a kernel author who can't afford a custom fine-tune, an open 32B model that beats Claude Sonnet 4.6 on GPU kernels by ten percentage points is directly useful, with weights on HuggingFace.

World models stop being a vision-and-robotics idea and become a code training tool. If you can predict the toolchain, you don't need to run it, and expensive simulation domains start scaling data like ordinary text. Read more in MiniMax M2.5.

Sources:

- InCoder-32B-Thinking: Industrial Code World Model for Thinking (arXiv)
- Industrial-Coder GitHub repository
- IndustrialCoder model weights on HuggingFace
- LiveCodeBench v5 leaderboard
- KernelBench: GPU kernel benchmark
