Downward LUT6 Compilation in a Mixture of Limits Cascade: Beneš-Routed Boolean Fabrics, Conformal Wake Gates, and Frontier Sparse-MoE NVMe Streaming

David Charlot, Open Interface Engineering

Research study · draft · OpenIE research.openie.dev
https://research.openie.dev/papers/lut-adapt/ · PDF https://research.openie.dev/pdfs/lut-adapt.pdf

Measurement. Research study, draft, not a journal final. Physical Texas Instruments INA3221 (VDD_CPU_SOC_MSS) and INA238 (VIN) DC shunt joules (measured_j, board_metered = true on NVIDIA Jetson AGX Thor solely in Claims LA-24 and LA-28) are kept strictly separate from five-host RAPL, IOReport, proc_pid_rusage, NVML, and live Alif Ensemble E3 Cortex-M55 DWT cycle counters (reported_j, board_metered = false) and from all FPGA RTL synthesis and VCD/electro-thermal calculations (measured_j = None, board_metered = false across all xck26 UltraScale+ and XC7A200T Artix-7 M.2 rows; est_j / est_mW / est_C). Formal claims (36 theorems across Cascade, BenesCopy, JointMinterm, ConformalProvenance, and NucleusCodon, zero sorry, zero opaque, zero ofReduceBool) are accepted inside Linux bubblewrap across all 6/6 independent kernels of lake check --paranoid (Lean 4.36.0). No fabricated citations.

A Mixture of Limits cascade with 74 six-kernel-verified Lean 4.36.0 theorems and Inria Charon+Aeneas mechanical Rust-to-Lean extraction (plus 1,438 theorems across four Open Interface Engineering repositories inside Linux bubblewrap sandboxes, Common Vulnerabilities and Exposures (CVE)-2026-72844-hardened) that tries zero-floating-point-operation (FLOP) paged 6-input look-up table (LUT6) circuits, a 12.75-kilobyte (13,058-byte, 11-export) WebAssembly (WASM) and 4-kilobyte no-standard-library Rust lut6-core + antisymmetric 40-entry zero-multiply table-lookup multiply-accumulate (T-MAC) + 1.750-bit-per-weight (-22.22%, saving 1.564 gigabytes static and cutting streamed gigibytes per token 1.619x via 54-bit 3-token LUT6 AirMoL speculation on native Hadamard-ternary Bonsai-2-27B) shift-5 Fast Walsh-Hadamard Transform (FWHT) ternary Nucleus, a 5,984-byte (9.13% of 64-kilobyte smartwatch static random-access memory) 56-bit 4-token smart-glasses opener with zero prefix poisoning, a 4,096-byte (6.25% of smartwatch static memory) downward-compiled System-0 decider running 9,493.7x faster than Mapika/decider-0.8b on Wave-2 decisions and lifting official Abstraction and Reasoning Corpus (ARC)-AGI-3 Wave-2 quadratic Relative Human Action Efficiency (RHAE) 208.3x from 0.009 to 1.875, and (72, 64) single-error-correction double-error-detection (SECDED) cosmic-ray bit-flip protection first (0.154 milliseconds in single-binary Rust mol-agent-rs, 162.71 nanojoules per bit-sliced decision, and 100.0% survival at 1% bit-error rate), escalates across sub-billion wearable micro-language-models and 4-bit Non-Volatile Memory Express (NVMe) looped, hybrid, and sparse Mixture-of-Experts (MoE) late-2026 models up to 120B parameters across four graphics processing unit (GPU) silicon vendors (up to 328.06x speedup, cutting LastExam-protocol root-mean-square calibration error 5.37x to 13.14%, solving 20.83% of ARC-AGI-2 on n=24 and 12.50% = 2.00x over 1-layer on n=64 via multilayer Beneš-ARC-LUT6, solving 4/57 official ARC-AGI-3 black-box levels [4.0x Random], verifying 50.00% of MiniF2F-Lean4 with 0.0000% unsound proofs,

> and cutting CritPt physics calibration error 9.49x to 8.96% with 0.00% confident hallucinations at >= 80% confidence), gates commits via deterministic verifiers across 4 upstream Verified software-engineering repositories, and compiles certified wakes downward into clock-enable-gated Beneš-routed 36-Kbit Block RAM (RAMB36E2 and license-free openXC7 320.62-megahertz post-route 7-Series Artix-7 200T M.2 2280 .bit bitstreams), Lupanov shared-cofactor, wobble-codon, and two-tier LUT6 + EntroMCU ternary crystal pages. Includes physical Texas Instruments direct-current (DC) shunt measured_j on Tegra aarch64 Jetson Thor (87.25x bit-sliced lut6-core, 222.57x wearable 56-bit 4-token epistemic gate vs. 28.84-million-parameter d=512 slice, 8.35x Bonsai-AirMoL 4-token speculative sweep [15.32 vs. 127.97 microjoules, 9.13x faster], 5.33x multilayer Beneš-ARC-LUT6 + ARC-AGI-3 reflex [713.82 nanojoules], 1.65x antisymmetric-40 zero-multiply T-MAC tail [and 1.12x on Zen 5], 1.57x Bonsai-2-27B shift-5 FWHT-1024 + antisymmetric-40 [220.72 vs. 346.09 microjoules, 2.26x faster], 3.74x Sub1Bit-LUT6, 34.21x Expert-AirMoL, and 10.94x Proposition 1 compilation energy reductions; board_metered is true) strictly separated from live Alif Ensemble Cortex-M55 cycle counts (2,745.18x faster wearable epistemic gate), five-host on-chip reported_j counters, license-free openXC7 post-route static timing, and simulated xck26 and Artix-7 200T register-transfer-level (RTL) est_j estimates at commit b8d6612.

Revision `581d9b4`, 24 November 2026. Reference implementation and artifacts: openIE-dev/lut-adapt-ai at `b8d661201214275525b230d4f8d929db3ba7a724`. Data summary: /data/#lut-adapt. LaTeX PDF: /pdfs/latex/lut-adapt.pdf (source).

---

0. Overview

This study examines how a computing system can compile repeated neural-model work downward into fast Boolean table lookups and certified canonical skills without losing exact output parity or promoting unverified guesses into trusted rules. Five evidence classes stay strictly separated across all claims: physical board-level Texas Instruments current-shunt joule readings (`measured_j`, `board_metered = true` on Tegra `aarch64` Jetson Thor solely in §4.26, §4.30, §4.34, §4.35, §4.40, §4.41, §4.44, and §4.46), operating-system and on-chip hardware energy counters plus bare-metal microcontroller cycle counters and browser virtual-machine execution (`reported_j` across five Linux, macOS, and Windows 11 workstations plus live attached Alif Ensemble Cortex-M55 cycle counts and the `12.75`-kilobyte browser module, `board_metered = false`), six-kernel sandboxed formal proof verification (`74` theorems here and `1,438` across four repositories) paired with Inria `Charon+Aeneas` mechanical Rust-to-Lean extraction (§4.19, §4.35, §4.40, §4.43, and §4.46), license-free `openXC7` post-route static timing closure (`320.62` megahertz) and `77.85`-megabit `.bit` bitstream generation on 7-Series Artix-7 `200T` (§4.36, `board_metered = false`, `measured_j = None`), and register-transfer-level synthesis, exhaustive Boolean miter checks, cycle simulations, and labeled analytical energy and electro-thermal estimates (`est_j`).

In plain words

1. Most inputs a deployed system sees repeat patterns it has already solved, so running a full floating-point neural network on every step wastes compute and memory bandwidth.
2. The reference places a reconfigurable 6-input truth-table (LUT6) lookup tier and a 4-kilobyte bare-metal Rust `lut6-core` Nucleus (fitting inside 25.06% of a 64-kilobyte microcontroller static memory or 365 on-die block memories of an M.2 2280 Artix-7 200T card) ahead of streaming state-space, looped, and foundation models, and compiles answers downward into lookup memory pages only after an explicit check certifies them.
3. When late-2026 looped, hybrid, and sparse mixture-of-experts language models up to 120B parameters (12B active, plus October–November 2026 looped `thinkingmachines/Inkling-2-Mini` and `Nemotron-Labs-Dory1.0-4B`) run locally on Apple Silicon through custom fused Apple Metal, portable Rust `Burn 0.22.1 / CubeCL` shaders, and single-binary Rust `mol-agent-rs` (0.154 milliseconds per skill lookup), router-sparse expert reads paired with that zero-FLOP LUT6 drafter cut disk reads by up to 12.30× on `Qwen/Qwen3.6-35B-A3B` (970.0 megabytes active weight memory), cut looped depth-recurrent forward passes 4.00× (3.52×–3.95× measured speedup at 100% greedy parity), and sustain 22.17 tokens per second on `nvidia/Nemotron-3-Super-120B-A12B` inside 58.08 gigabytes of Unified Memory; gating code and repository commits behind deterministic `unittest` and `pytest` checks reaches **91.67% verified pass rate (93.75% across Wave 2 and Wave 3)** on the 48-task algorithmic bakeoff, **90.62% (93.75% on Wave 2 human paraphrases)** on official `BigCodeBench v0.1.3`, **50.00%** on single-repository SymPy git worktrees, **20.83% (2.50× standalone 24B and 35B models)** across 4 upstream `princeton-nlp` Verified repositories (`django`, `pytest`, `requests`, and `sympy`), and **62.50%** on the 120B hard frontier suite with **0.00% broken commits admitted**.

---

1. Terms, acronyms, and measurement labels

- MoL: Mixture of Limits, the cascade rule from Metered Commit Gates Part II [same house]
- LA: claim and falsifier prefix (LA-1 to LA-44, LA-F1 to LA-F44) used in this study
- LUT, LUT6, INIT: look-up table, 6-input Xilinx UltraScale+ Boolean primitive, and its 64-bit truth-table initialization mask
- BRAM, RAMB36, RAMB36E2, URAM: Block RAM, the 36-Kbit dual-port Xilinx Block RAM family and its UltraScale+ primitive (144 on xck26), and UltraRAM
- CLB, FF, DSP, DSP48, KV260: configurable logic block, flip-flop, digital signal processing slice, and AMD Kria KV260 board (xck26)
- FPGA, ASIC, RTL, CPI, PAR, WNS, FPL: field-programmable gate array, application-specific integrated circuit, register-transfer level, cycles per lookup, place-and-route, worst negative slack, and Field-Programmable Logic conference
- AXI, AXIS, PS, PL, SOM, INA260, VCC, RAPL, NVML, PMU, SMC, L1, L2, L3: Advanced eXtensible Interface, AXI4-Stream wrapper prefix, Zynq processing system, programmable logic, system-on-module, onboard power monitor IC, supply rail, Intel Running Average Power Limit, NVIDIA Management Library, macOS power management unit, Apple System

Management Controller, level-1 whole-board voltage-input rail, level-2 subsystem energy counter (Apple IOReport GPU Energy and Tegra rail shunt), and level-3 per-process CPU energy counter (Darwin proc_pid_rusage RUSAGE_INFO_V6 ri_energy_nj)

- SSD, SSM, MIMO, RoPE, SwiGLU, GQA, KV, BPE, LLM, LLMs, MoE, MLP, CFG: State Space Duality, state-space model, multiple-input multiple-output, rotary position embedding, gated linear unit activation, grouped-query attention, key-value cache, byte-pair encoding, large language model(s), Mixture-of-Experts, multilayer perceptron, and context-free grammar
- HC, SOINN, WNN, DWN: hierarchy-context Self-Organizing Incremental Neural Network, weightless neural network, and Differentiable Weightless Neural Network
- JEPA, EBRM, ACI, OOD, AUROC, AUC, MSP, MC: Joint-Embedding Predictive Architecture, Energy-Based Reasoning Model, Adaptive Conformal Inference, out-of-distribution, area under the receiver operating characteristic, maximum softmax probability, and Monte Carlo
- PLD, REST, EAGLE, LFM2, LFM3, HF: Prompt Lookup Decoding, Retrieval-Based Speculative Decoding, Extrapolation Algorithm for Greater Language-model Efficiency, Liquid Foundation Model 2.5 (LFM2.5-350M and LFM2.5-8B-A1B), Liquid Foundation Model 3 (Oct 2026 LiquidAI/LFM3-1.7B), and HuggingFace
- SVD, ITQ, SVID, QAT, PTQ, PPL, SNR: singular value decomposition, Iterative Quantization Procrustes hypercube rotation, sign-value-independent decomposition, quantization-aware training, post-training quantization, perplexity, and signal-to-noise ratio
- NVMe, NOCACHE, POSIX, RAM, DRAM, SRAM, RSS, GGUF, IO, KB, MB, GB, GiB: Non-Volatile Memory Express storage, Darwin F_NOCACHE unbuffered file flag, Portable Operating System Interface, random-access memory, dynamic RAM, static RAM, resident set size in physical Unified Memory, GPT-Generated Unified Format quantized model file, input/output, kilobyte, megabyte, gigabyte, and gibibyte
- FLOP, FLOPs, MFLOP, MFLOPs, MAC, MACs, SIMD, AD, FD, BPB, BF16, FP16, FP32, INT4, ULP, MSL, MPS, CUDA, ROC, SPIR, PTX, HIP, LLVM: floating-point operation(s), million FLOPs, multiply-accumulate operation(s), single-instruction multiple-data, automatic differentiation, finite difference, bits per byte, brain float 16, 16/32-bit IEEE floating point, 4-bit integer, unit in the last place, Apple Metal Shading Language, PyTorch Metal Performance Shaders, NVIDIA Compute Unified Device Architecture, AMD ROCm runtime, Standard Portable Intermediate Representation (Vulkan SPIR-V), Parallel Thread Execution, Heterogeneous-Compute Interface for Portability, and Low Level Virtual Machine compiler backend
- RCE, WCA, VC, BCB, SWE, OSA, DP: Reality Code Engine [same house], Wise Computer Automation [same house], verified completion, BigCodeBench v0.1.3 benchmark, software-engineering benchmark prefix, optimal string alignment, and dynamic programming
- CLI, HTTP, GET, POST, SSE, API, MCP, RPC, IDE, SDK, CORS, URL, ORM, IDNA, CADE, LNCS, NVIDIA, ACL, ICML, ICLR, EMNLP, NAACL, CVPR, ICPE, ISSCC, JSAC, JASA, ACM, OS, CPU, GPU, SW, TDP: command-line interface, Hypertext Transfer Protocol and its GET/POST request methods, Server-Sent Events, application programming interface, Model Context Protocol, remote procedure call (JSON-RPC 2.0), integrated development

environment, software development kit, Cross-Origin Resource Sharing, Uniform Resource Locator, object-relational mapper, Internationalized Domain Names in Applications, Conference on Automated Deduction, Lecture Notes in Computer Science, NVIDIA Corporation, academic computer-science/engineering venues, operating system, central/graphics processing unit, software, and thermal design power

- VLAN, IEEE, TLS, DPI, AST, JSON, CSV, SHA, CRC, ATM, RFC, SQL, DWARF, SLEB, CAM, XOR, XNOR, JWT, HMAC, DAG, CIDR, LRU, TTL, WAL, ETL: virtual local-area network (IEEE 802.1Q), Transport Layer Security, deep packet inspection, abstract syntax tree, JavaScript Object Notation, comma-separated values, Secure Hash Algorithm, cyclic redundancy check, Asynchronous Transfer Mode (CRC-8/ATM), Request for Comments, Structured Query Language, DWARF signed little-endian base-128 encoding, content-addressable memory, exclusive-OR, exclusive-NOR, JSON Web Token, hash-based message authentication code, directed acyclic graph, Classless Inter-Domain Routing, least-recently-used cache, time-to-live expiration, write-ahead log, and extract-transform-load pipeline
- JTAG, XSDB, DAP, USB, FTDI, FT2232C, UART, FIFO, IC, EDA, OOC, AMBA, AXI4, IOB, IBUF, OBUF, FDRE, FDSE, LUT1, LUT2, LUT3, LUT4, LUT5, MUXF7, MUXF8, CARRY4, CARRY8, DSP48E2, URAM288, VCCINT, DS925, VCD, SDP, CE, ENARDEN, XPE, CLAI, AMD: Joint Test Action Group debug port, Xilinx Software Command-Line Tool debugger, Debug Access Port, Universal Serial Bus, Future Technology Devices International and its FT2232C dual UART/FIFO bridge IC, universal asynchronous receiver-transmitter, first-in first-out buffer, integrated circuit, electronic design automation, out-of-context synthesis mode, Advanced Microcontroller Bus Architecture, AXI4-Lite/AXI4-Stream interconnect, input/output block, input/output buffers, synchronous-reset/set D flip-flops, 1-to-5-input UltraScale+ look-up tables, 7/8-input UltraScale+ multiplexers, 4/8-bit fast carry chains, UltraScale+ DSP slice, 288-Kbit UltraRAM, internal core supply rail (0.72 V on -2LV), AMD Zynq UltraScale+ DC and AC Switching Characteristics datasheet DS925, Value Change Dump waveform trace, simple dual-port Block RAM mode, synchronous clock enable, RAMB36E2 port-A read/clock-enable pin, Xilinx Power Estimator, subsystem register magic identifier ("CLAI"), and Advanced Micro Devices
- CVE, CWE, FRO, GMP, RAG, WHNF, ZFC, CIC, FFI: Common Vulnerabilities and Exposures, Common Weakness Enumeration, Lean Focused Research Organization, GNU Multiple Precision Arithmetic Library, Retrieval-Augmented Generation, weak head normal form, Zermelo-Fraenkel set theory with Choice, Calculus of Inductive Constructions, and foreign function interface
- ABBA, AGX, AI, DC, FWHT, INA238, INA3221, MSS, NT, PPT, RDNA, RPL, RTX, RUSAGE_INFO_V6, TI, VDD, VDD_CPU_SOC_MSS, VDD_GPU, VIN, XNU: counterbalanced A-B-B-A trial ordering, Autonomous-Grade Compute module family, artificial intelligence, direct current, Fast Walsh-Hadamard Transform, Texas Instruments inline power monitors (INA238 whole-board VIN and three-channel INA3221 rail), memory subsystem, Windows NT kernel family, Package Power Tracking, Radeon DNA graphics architecture, Raptor Lake

mobile platform, Ray Tracing Texel eXtreme, Darwin Resource Usage Info v6 struct, Texas Instruments, voltage drain-drain supply rail, Tegra CPU/SoC/memory supply rail, Tegra GPU supply rail, board voltage-input rail, and Darwin X is Not Unix kernel

- ABI, AC, ADC, AE3, A2B, CRC32, CYCCNT, DTCM, DWT, E3, E4, E4B, ELF, ELFs, FCCM, FSM, GPT, INT8, ISVLSI, LSB, M0, M55, MCU, MRAM, PALUTE, PIO, Q4, ROM, RP2040, SHA256, SPI, ST, XL: application binary interface, alternating-current non-constant Fourier energy, analog-to-digital converter, Alif Ensemble E3 OpenMV-AE3 microcontroller board, 2-billion active-expert sparse architecture suffix, 32-bit IEEE cyclic redundancy check, Arm Data Watchpoint and Trace cycle counter register, Data Tightly Coupled Memory, Data Watchpoint and Trace unit, Alif Ensemble E3/E4 microcontroller series, 4-billion effective edge parameter suffix, Executable and Linkable Format firmware image(s), Field-Programmable Custom Computing Machines symposium, finite-state machine, Generative Pre-trained Transformer (normalized nGPT), 8-bit integer quantization, International Symposium on VLSI, least-significant bit, Arm Cortex-M0+ and Cortex-M55 microcontroller cores, microcontroller unit, magnetoresistive non-volatile RAM, Processing-In-Memory Acceleration via Lookup Table, programmable input/output peripheral, 4-bit quantized block format (Q4_K_XL), read-only memory, Raspberry Pi dual Cortex-M0+ microcontroller, 256-bit Secure Hash Algorithm, Serial Peripheral Interface bus, Straight-Through estimator, and extra-large quantization block suffix
- BAR0, BGA, BLE, CEM, CMOS, CSR, DDR3, DS181, DSP48E1, ECC, FBG484, FPC, GPIO, GTP, HAT, HPL, I2C, INA228, IRQ, JEDEC, LFM, M4F, MIG, MT, MUL, NDP120, ODT, PCIe, PHY, QCM6490, QFAA, RAMB36E1, RISC, RV32IMAC, SSTL15, SYSMON, TMAC, UG475, VNNI, XADC, XC7A200T, XCK26, XCZU9EG, ZCU102: PCIe Base Address Register 0, ball-grid array, Bluetooth Low Energy, Card Electromechanical PCIe slot specification, complementary metal-oxide-semiconductor, control and status register, double-data-rate type-3 synchronous DRAM, AMD Artix-7 FPGAs DC and AC Switching Characteristics datasheet DS181, Xilinx 7-Series digital signal processing slice, error-correcting code, 484-ball 23x23 mm lidless flip-chip BGA package, flexible printed circuit ribbon, general-purpose input/output, Artix-7 multi-gigabit serial transceiver, Hardware Attached on Top 40-pin expansion header, TSMC 28 nm High-Performance Low-Power process, Inter-Integrated Circuit bus, Texas Instruments ultra-precise inline current/power monitor IC, interrupt request, Joint Electron Device Engineering Council thermal coupon standard, linear feet per minute airflow unit, Arm Cortex-M4F microcontroller core, Xilinx 7-Series Memory Interface Generator, megatransfers per second, integer multiply instruction, Syntiant NDP120 neural decision processor, on-die termination, Peripheral Component Interconnect Express, physical memory interface layer, Qualcomm Dragonwing system-on-chip on Particle Tachyon, Nordic 48-pin QFN nRF52832 package suffix, Xilinx 7-Series 36-Kbit dual-port Block RAM primitive (365 on XC7A200T), Reduced Instruction Set Computer, 32-bit RISC-V integer/multiply/atomic/compressed ISA, Stub Series Terminated Logic 1.5 V I/O standard, Xilinx System Monitor temperature/voltage block, Table-Lookup MAC ternary dot-product layout, AMD 7-Series FPGAs Packaging and Pinout Specification

UG475, Vector Neural Network Instructions, Xilinx 7-Series dual 12-bit analog-to-digital converter, AMD Artix-7 200T FPGA (Numato Aller A7 M.2 2280), AMD Kria K26 SOM FPGA, AMD Zynq UltraScale+ MPSoC FPGA, and AMD ZCU102 evaluation board

- BEL, BER, FASM, GBDT, ICFP, ISCA, JS, KAN, LLBC, LVCMOS33, MIR, PACMPL, PIM, PIP, PIPs, RMSE, SECDED, SEU, SLICE_FFX, SLICE_LUTX, TMLR, UG470, V8, VM, WASM, WASM32, XC7, XDC: Xilinx basic logic element, bit-error rate, FPGA Assembly frame format, gradient-boosted decision tree, International Conference on Functional Programming, International Symposium on Computer Architecture, JavaScript host runtime, Kolmogorov-Arnold Network, Low-Level Borrow Calculus (Inria Charon/Aeneas intermediate representation), Low-Voltage CMOS 3.3 V I/O standard, Rust Mid-level Intermediate Representation, Proceedings of the ACM on Programming Languages, processing-in-memory, Xilinx switchbox programmable interconnect point(s), root-mean-square error, single-error-correction double-error-detection (72, 64) extended Hamming code, single-event upset cosmic-ray bit-flip, nextpnr-xilinx 7-Series slice flip-flop and LUT primitives, Transactions on Machine Learning Research, AMD 7 Series FPGAs Configuration User Guide UG470, Chromium V8 WebAssembly engine, virtual machine, WebAssembly portable binary target (32-bit wasm32-unknown-unknown), Xilinx 7-Series architecture family, and Xilinx Design Constraints pinout file
- AIME, AR, ASR, CAIS, CHR, GPQA, HLE, HMMT, HRV, HUD, LTE, MC10, MMLU, POV, RMS, RTT, SPO2, TTFT, UW, XR: American Invitational Mathematics Examination, augmented reality, automatic speech recognition, Center for AI Safety, confident hallucination rate (at verbalized confidence >= 80%), Graduate-Level Google-Proof Question Answering, Humanity's Last Exam (lastexam.ai), Harvard-MIT Mathematics Tournament, heart-rate variability, heads-up display, Long-Term Evolution cellular radio, 10-option multiple choice (A..J), Massive Multitask Language Understanding, point-of-view camera frame, root-mean-square calibration metric, round-trip time, peripheral blood oxygen saturation, time to first token, University of Washington, and extended reality
- ACTION1, ACTION6, ACTION7, AGI, ARC, ARC2, ARC3, BFS, BPW, IRB, LEFT, MIT, ML, MLX, RHAE, RIGHT: ARCEngine discrete and coordinate-click action opcodes, artificial general intelligence, Abstraction and Reasoning Corpus (ARC-AGI-2 and ARC-AGI-3) and their shorthand identifiers, breadth-first search, bits per weight, Interactive Reasoning Benchmark, directional grid action tokens, Massachusetts Institute of Technology open-source license, machine learning, Apple Machine Learning Explore array framework, and Relative Human Action Efficiency

Six evidence classes stay separate across every table below:

Evidence label	Meaning in this study
`Physical DC Shunt (measured_j)`	Direct-current inline shunt readings on `NVIDIA Jetson AGX Thor` (`Claims LA-24, LA-28, LA-32, LA-33, LA-38, LA-39, LA-42 & LA-44` only, `board_metered = true`): Texas Instruments `INA3221` channel-2 `VDD_CPU_SOC_MSS` (`L2`, physically isolated from `VDD_GPU`) and `INA238` whole-board `VIN` (`L1`), integrated trapezoidally at `500 Hz` across counterbalanced `ABBA` windows. Never summed with `reported_j` or `est_j`, and unset (`None`) on all non-Thor rows.

Evidence label	Meaning in this study
Measured (SW / NVMe I/O / OS PMU)	Exact byte counts read or written via `os.pread` under Darwin `F_NOCACHE`, resident memory allocations, wall-clock seconds, whole-machine macOS `PMU` (`AppleSmartBattery.PowerTelemetryData`) cumulative energy deltas on Apple M5 Max, `12.75 KB` `WASM` execution, and exact bit/token agreement.
Reported (OS / On-Chip Counter, reported_j)	Unprivileged operating-system and on-chip telemetry counters (`board_metered = false`): on Apple M5 Max and M3 Ultra, `libIOReport.dylib 'GPU Energy'` (`L2 nJ`) plus `proc_pid_rusage(RUSAGE_INFO_V6).ri_energy_nj` (`L3` CPU `nJ`); on `x86-64` Linux, `RAPL` (`package-0` / `core`), `amdgpu` `PPT`, and `NVML mJ`; on Windows 11, `NVML mJ`. Never merged with `measured_j`.
Post-PAR Open-Source Gateware (openXC7)	License-free physical placement, switchbox `PIP` wire routing, static timing analysis (`320.62 MHz` post-PAR on `xc7a200tfbg484-2`), and `77.846 Mbit` `.bit` bitstream emission via `Yosys` + `nextpnr-xilinx 0.8.2` + `Project X-Ray` (`Claim LA-34`, `board_metered = false`, `measured_j = None`).
Simulated (RTL)	Structural netlist counts from Yosys `0.71+` (`synth_xilinx -family xcup` targeting Kria KV260 `xck26` and `-family xc7` targeting Artix-7 `XC7A200T`) and cycle counts (`CPI = 1`) from Verilator `5.054`. Not Vivado timing closure and not physical board power (`measured_j = None` for all `xck26` and `XC7A200T` rows).
Estimated (est_j)	Analytical joules and electro-thermal bounds computed from operation counts, thermal-design-power time stand-ins (`joule_meter.py`), or Verilator VCD toggle-rate switching activity paired with AMD `DS925` `xck26-2LV` (`0.72 V`) `RAMB36E2` (§4.22, §4.25) and AMD `DS181` / `UG475` `XC7A200T-2FBG484C` (`1.00 V`) `RAMB36E1` capacitance and thermal-resistance models (§4.32). Comparable only across arms inside the same formula; never silicon watts.

1.1 Architectural background and contributions

In Metered Commit Gates `[same house]`, Part II defines **Mixture of Limits**: try gears in decreasing order of pass probability per joule p_i/c_i (Simon and Kadane, 1975), stop as soon as the task completion predicate $C(z) = 1$ holds, and forbid strengthening `ModelGenerated` output into `Deterministic`. A static Lookup tier, however, covers only the entries written into it before deployment. Conversely, existing lookup-table neural networks — LogicNets (Umuroglu et al., 2020), DiffLogic (Petersen et al., 2022), and Differentiable Weightless Neural Networks (Bacellar et al., 2024) — fix wire connectivity inside static FPGA routing, requiring vendor place-and-route whenever the input distribution shifts.

The `lut-adapt-ai` reference closes the loop between the top and bottom of the cascade via **online downward `LUT6` compilation**:

- **Rung 0 (Lookup — Paged Beneš + `LUT6`)**: Evaluates 64-bit truth-table masks routed through a non-blocking 512×512 **Beneš–Copy–Beneš** multicast switch and cached across $81 \times$ **`RAMB36E2`** dual-port block memories (`0` floating-point operations, single-cycle page switching in Verilator simulation).

- **Rung 1 (Stream — `Mamba-3` Complex MIMO State Space Duality)**: Handles misses that depend on sequence history using a constant-size $O(1)$ recurrent state (`48 KB` pinned memory in the `14.3M`-parameter reference), avoiding $O(L)$ key-value cache growth.
- **Rung 2 (Topology — `HC-SOINN`)**: Maintains online prototype topology over latent embeddings, isolating regime boundaries and tripping a circuit breaker (`Refuse` / escalate) when unlabeled distribution shift arrives.
- **Rung 3 (Model last — `JEPA-EBRM`, `AirMoL`, and Metal-Fused `Expert-AirMoL` 4-Bit NVMe Streaming)**: Invokes latent energy-based verification or streams pretrained late-2026 hybrid and sparse-MoE foundation models (`LiquidAI/LFM3-1.7B`, `google/gemma-4-E2B-it`, `LiquidAI/LFM2.5-8B-A1B`, and `Qwen/Qwen3.6-35B-A3B`, alongside `LiquidAI/LFM2.5-350M` and `Qwen/Qwen3.5-0.8B-Base`) through custom Apple Metal `MSL` compute shaders only when cheaper rungs abstain or require exact target verification.
- **Downward `SleepCrystallizer`**: Extracts recurring wake traces using Walsh–Hadamard spectral warm-starts, joint-mode minterm truth tables, canonical AST skills, and Waksman switch schedules, flipping shadow block-memory banks without halting inference or running FPGA place-and-route.

1.2 Contributions

1. **Amortized downward compilation condition**: Extends the Simon–Kadane p_i/c_i ordering with an explicit compilation cost E_{compile} over stationary horizon ΔT.
2. **Non-blocking 512×512 Beneš–Copy–Beneš, `81-RAMB36E2` page cache, and `4`-stage `250 MHz`-pipelined `P4-AXIS` RTL subsystem**: Eliminates both the `102.0%` Kria KV260 logic overflow of flat crossbars (`119,478` LUTs) and the `926,619`-LUT explosion of flip-flop page multiplexers (`41,186` UltraScale+ CLB LUTs = `35.2%` of `xck26`, `81` `RAMB36E2` blocks = `56.2%`, `0` `DSP48` slices in 1-cycle OOC), while a 4-stage pipelined IOB-safe `AXI4-Stream` wrapper (`clai_benes_axi_stream_250mhz_top`) reduces single-cycle synchronous logic depth **`4.00×`** from 52 to **`13` levels**, fits **`170 <= 189` `xck26` `IOBs`** (`41,459` LUTs = `35.40%`), and passes **`256 / 256` zero-bubble `CPI = 1.00`** Verilator streaming parity.
3. **Universal ≤ 64-bit joint-mode minterm synthesis**: Shows joint-mode minterm lookup dominates marginal bit-plane DiffLogic rounding on correlated branches and packs $K = \lfloor 64/\lceil \log_2 |V| \rceil \rfloor$ subword tokens into one `uint64` word (`16b × 4 = 64b` for `|V| = 65,536`; `17b × 3 = 51b` for `|V| = 128,000`; `18b × 3 = 54b` for `|V| = 248,320`).
4. **Oracle-free Split-Conformal + Online ACI crystallization gate**: Pairs a radially regularized latent `JEPA-EBRM` nonconformity score with Split Conformal calibration and Adaptive Conformal Inference (Gibbs and Candès, 2021), recording `0.00%` false crystallization risk and `1.0000` OOD AUROC on shifted traffic where softmax and MC-Dropout baselines record `17.62%`–`24.28%` false risk.
5. **Custom Apple Metal & Cross-Platform Rust `CubeCL` `INT4` Sparse-MoE / `GatedDeltaNet` shaders + `AirMoL` / `Expert-AirMoL` up to `34.66B`**: Combines Darwin `F_NOCACHE` `os.pread` 4-bit sharding, custom `MSL` (`43.45×` MoE, `20.66×` `GatedDeltaNet`) and portable Rust `Burn` `0.22.1` / `CubeCL` (`35.09×` Metal / `2.67×` `LLVM CPU` MoE, `69.20×` Metal / `14.87×` `LLVM CPU` `GatedDeltaNet`) compute shaders, Queue-Depth-8 parallel `pread`, and `BenesLUT6` burst

speculation across October 2026 LiquidAI/LFM3-1.7B (2.74× NVMe reduction in 3.79% FP32 RAM), LiquidAI/LFM2.5-8B-A1B (500.0 MB RAM, 3.493 J/tok reported_j), and Qwen/Qwen3.6-35B-A3B (34.66B; **12.30× NVMe reduction**, **970.0 MB RAM = 1.47% of FP16**, **6.773 J/tok reported_j**, **15.7 tok/s** Unified-Memory Q4) at 100.0% exact token parity.

6. **Zero-MAC Sub1Bit-LUT6 vs. LittleBit, 91.67% Rung 3d <think> ceiling, official BigCodeBench v0.1.3 + upstream SWE-bench Verified via MCP, and OpenAI / MCP mol-agent**: Replaces the 1.98M runtime binary MACs of LittleBit-1 / LittleBit-2 (0.576 bits per weight; Lee et al., 2025; Lee and Kim, 2026) with zero-MAC 6-input LUT6 partial-sum lookup ($< 4 \times 10^{-6}$ error, 100.0% parity under MoL target verification in 0 QAT epochs), lifts the 48-task late-2026 coding bakeoff from 87.50% to **91.67% (44 / 48; 93.75% in Wave 2 and Wave 3)** with bounded open-weights Rung 3d <think>, reaches **90.62% (29 / 32, 93.75% Wave 2)** on official BigCodeBench v0.1.3 (vs. 78.12% on standalone 35B) and **50.00% (8 / 16)** on upstream SWE-bench Verified sympy/sympy git worktrees (vs. 25.00% on 35B) with **0.00% broken commits admitted**, and serves persistent .mol/ skills over OpenAI /v1/chat/completions and MCP (JSON-RPC 2.0).
7. **Six-kernel paranoid Lean 4.36.0 formalization (CVE-2026-72844-hardened, 74 theorems + Inria Charon+Aeneas Rust-to-Lean), 4-repository SWE-bench Verified call-graph localization, measured 120B Hybrid MoE apex, CE-gated VCD power, and single-binary Rust mol-agent-rs (Claims LA-19 to LA-23, LA-33, LA-38, LA-41 & LA-44)**: Checks all **74 theorems** across Cascade, BenesCopy, JointMinterm, ConformalProvenance, NucleusCodon, TernaryTmac, GlobalFrontiers, WearableLastExam, TernaryHadamardArc, and BonsaiStreamArc3 across **six independent proof checkers** (con-leche --verified [58,841 decls], con-ron --verified [58,841 decls], lean4lean [58,843 decls], nanoda, leanchecker --paranoid, and Lean 4.36.0; 0 sorry, 0 custom axioms, 0 Lean.ofReduceBool; 1,438 fleet theorems; Moura and Ullrich, 2021) alongside Inria Charon+Aeneas mechanical #![no_std] Rust MIR → LLBC → Lean 4.36.0 extraction (Lut6Verify_Aeneas.lean; Ho and Protzenko, 2022), expands zero-leakage SWE-bench Verified evaluation across **4 upstream repositories (django, pytest, requests, sympy, 24 evals)** at **20.83% overall / 25.00% Wave 2 (2.50× standalone 24B and 35B)** with 0.00% broken patches, measures **120B-scale nvidia/Nemotron-3-Super-120B-A12B** (58,075.4 MB RSS, **22.17 tok/s**, **62.50% verified pass** vs. 43.75% standalone 120B-A12B and 31.25% Devstral-Small-2-24B), reduces P4-AXIS 250 MHz idle dynamic power **4.10×** (63.99 to **15.59 est_mW**) via ENARDEN + valid-bubble CE gating, and delivers a zero-Python single-binary Rust mol-agent-rs (0.1540 ms/req MCP, **17.21× faster** than Python mol-agent, 71 / 71 pytest tests passing).
8. **4 KB #![no_std] Rust 1.100.0 lut6-core + ISCA 2025 antisymmetric-40 EntroMCU/T-MAC Nucleus across 8 bare-metal/WASM targets (12.75 KB lut6_core.wasm), (72, 64) RAMB36 SECDED cosmic-ray SEU shield, license-free openXC7 320.62 MHz .bit bitstream, smart-glasses micro-LLM (Sensente/uLM-28M, LFM2.5-230M) vs. 5,984 B Wearable-MoL (0.00% prefix poisoning), and LastExam-protocol calibration (5.37× lower RMS calibration error, 0.00% CHR at >= 80%) (Claims LA-26 to LA-38)**: Cross-compiles an exact 4,096 B CompactLut6Page (0 wasted bytes with the 672 B EntroMCU 1.602-BPW + antisymmetric 40-

entry PALUTE/T-MAC zero-multiply ternary tail and rustc const compile-time proof-vector gate) across **8 #![no_std] targets** — including a **12.75 KB (13,058 B, 11 exports) zero-crate wasm32-unknown-unknown lut6_core.wasm browser module**, Alif E3/E4 Cortex-M55, RP2040 Cortex-M0+, Nordic nRF52832 Cortex-M4F (25.06% full Nucleus and **9.13% [5,984 B] Wearable-MoL Rung-0** of 64 KB watch SRAM), and RV32IMAC (64/64 Yosys abc -lut 6 Boolean miter vectors) — proves universal 0 degree->=5 Walsh–Hadamard noise on 2-bit wobble LUT6s (NucleusCodon.lean) and cuts active LUT6 gates **64.06%** (DiffLUT-Net) and **72.50%** (Lupanov shared cofactors), records **162.71 nJ/op (87.25× physical CpJ vs. INT8 Micro-MLP)**, **2,803.25 vs. 623,920.52 nJ/op (222.57× physical Thor shunt measured_j on the 56-bit 4-token wearable epistemic gate vs. 1/16th uLM-28M d=512 layer; 2,745.18× faster in live OpenMV-AE3 Cortex-M55 DWT cycles)**, and **8,594.93 nJ/op (1.65× Thor shunt CpJ; 1.12× Zen 5 RAPL CpJ flip)** on Antisym-40 T-MAC, survives BER = 1.0% cosmic-ray bit-flips at **100.0% accuracy (0.0000% silent corruption)** via (72, 64) RAMB36 SECDED + 0x3C seed apoptosis, closes license-free **openXC7 (nextpnr-xilinx) post-PAR static timing at 320.62 MHz (77.846 Mbit .bit)** on XC7A200T-2FBG484-2 alongside the 62,211-LUT (365-crystal 1.46 MB RAMB36E1 vault, 0.86 est_W @ 55.2 est_C) M.2 2280 port, eliminates 93.75% collaborative prefix poisoning on Sensente/uLM-28M, cuts LastExam-protocol RMS calibration error **5.37×** (70.62% → 13.14%, 0.00% CHR at >= 80%, 40.00% selective accuracy), and benchmarks Inkling-2-Mini, Dory1.0-4B, LFM2.5-24B-A2B, and NVIDIA-Nemotron-3.5-Bolt-E4B at 100.0% greedy parity.
9. **Native 27B Hadamard-ternary Bonsai-2-27B (1.750 BPW EntroMCU-28 AirMoL streaming + 54-bit 3-token LUT6 speculation + shift-5 FWHT-1024), System-1 neural decider compilation (Mapika/decider-0.8b → 4 KB System-0, 9,493.7× Wave-2 speedup), multilayer Beneš-ARC-LUT6 on ARC-AGI-2 (20.83% / 12.50%), and official ARC-AGI-3 interactive compilation (208.3× Wave-2 RHAE) (Claims LA-39 to LA-44)**: Audits prism-ml/Ternary-Bonsai-2-27B-mlx-2bit (8.005 GB .safetensors, 26.87B native Hadamard-ternary weights on Qwen/Qwen3.8-27B, 0 invalid c=3 codes across 9.70M weights), proving lossless **36 B -> 28 B (2.25 -> 1.75 BPW, -22.22%, saving 1.564 GB static and 1.490 GiB per streaming sweep, 0.0 FP32 error)** EntroMCU transcoding, cutting end-to-end streamed linear GiB/tok **1.619× (6.787 -> 4.191 GiB/tok)** and reported_j/tok **-16.55%** at **100.00% greedy token parity** (14.35 -> 14.85 tok/s on Apple M3 Ultra 256 GB) via 54-bit 3-token LUT6 speculation + exact hybrid ArraysCache + KVCache rollback (8.35× lower Thor measured_j [15.32 vs. 127.97 µJ/tok, 9.13× faster] on Window 14; 1.57× lower Thor measured_j on Window 12 FWHT-1024 + Antisym-40); compiles verified Mapika/decider-0.8b (1,435.1 MB System-1 JSON decider) wakes into a 4,096 B System-0 CompactLut6Page (**9,493.7× faster [5.3 µs vs. 50.32 ms] / 3,185.3× lower reported_j, 0.00% OOD false actions**); scales compositional Beneš-ARC-LUT6 ($1 \leq L \leq 3$, diameter $2L+1$) on arcprize/ARC-AGI-2 from 12.50% to **20.83% (5 / 24, +66.7%)** on n = 24 and doubles n = 64 from 6.25% to **12.50% (8 / 64, 2.00× over 1-layer and 8.00× over unverified 35B, 0.00% CHR at >= 80%)** alongside MiniF2F-Lean4 (**50.00% Lean 4.36.0 verified, 0.0000% unsound**) and CritPt (**9.49× lower RMS calibration error**); and evaluates all 8 black-box games (57 levels, 0 game-source peeking) of official **ARC-AGI-3** (arc-agi 0.9.10 + arcengine 0.9.3),

solving **`4 / 57` levels (`4.0×` Random: `tu93` `2/9`, `cd82` `1/6`, `ls20` `1/7`)** via online $C_{\text{env}}(\hat{T}) = 1$ world-model `BFS` and lifting Wave-2 quadratic **`RHAE` `208.3×` (`0.009 -> 1.875` mean, `6.667` peak on `tu93`)** by compiling verified shortest paths into a `4 KB` `System-0` `LUT6` trajectory crystal guarded by a transition-mismatch circuit breaker (`5.33×` lower `Thor` `measured_j` at `713.82 nJ/op` on `Window 15`).

1.3 Same-house disclosure and provenance boundaries

Prior Open Interface Engineering studies — Metered Commit Gates, Task Satiation vs. Aggregate Rebound, The Reality Code Engine, RCE-JEPA, and RCE Compile — are `[same house]` works cited for disclosure and architectural continuity, never as external confirmation. Two invariants bind throughout:

- **Provenance non-strengthening (`ModelGenerated ↛ Deterministic`)**: Tables compiled from a deterministic `RCE` specification check are `Check-safe` on verified holdout sets (guarded online by an `HC-SOINN` / `RCE` circuit breaker). Tables crystallized under the oracle-free `JEPA-EBRM` conformal gate carry a finite-sample statistical bound (`Conformal-bounded`, $1-\alpha$) and are never promoted to `Deterministic`.
- **Proposer vs. authority (`The drafter proposes; the target decides`)**: In speculative decoding, `PagedBenesLUT6` acts strictly as a zero-FLOP proposer; every emitted byte or token is verified against greedy target decoding (`100.0%` lossless parity).

---

2. Mathematical formulations and four results

2.1 Result 1: Cascade ordering with amortized downward compilation

Consider gears $i \in \{0,1,2,3\}$ tried sequentially on an input x. Gear i closes $C(z) = 1$ with probability p_i at expected analytical energy $c_i > 0$. By Simon and Kadane (1975), expected spend

$$\mathbb{E}[J_\sigma] = \sum_{k=1}^{n} c_{\sigma(k)} \prod_{j<k} \left(1 - p_{\sigma(j)}\right)$$

is minimized if and only if gears run in non-increasing order of $p_{\sigma(k)}/c_{\sigma(k)}$. Suppose a stationary regime $\mathcal{D}_r$ persists for ΔT remaining queries where the wake path closes with $p_{\text{wake}} = 1$ at cost c_{wake}. Expending one-time compilation energy E_{compile} after N_{wake} certified wakes produces a Rung 0 `LUT6` page with certified hit rate $p_0' \in (0,1]$ and lookup cost $c_0 \ll c_{\text{wake}}$.

Proposition 1 (Net energy condition for downward crystallization). *Downward compilation reduces expected cumulative analytical spend over horizon ΔT if and only if*

$$\Delta T \cdot \left(p_0' c_{\text{wake}} - c_0\right) > E_{\text{compile}}, \qquad \text{and} \qquad \frac{p_0'}{c_0} > \frac{1}{c_{\text{wake}}}.$$

Proof. Without crystallization, ΔT queries cost $\Delta T \cdot c_{\text{wake}}$. With crystallization at cost E_{compile}, Rung 0 precedes the wake path whenever $p_0'/c_0 > 1/c_{\text{wake}} \iff c_0 < p_0' c_{\text{wake}}$, and each query costs $c_0 + (1 - p_0')c_{\text{wake}}$, saving $p_0' c_{\text{wake}} - c_0 > 0$ per step. Summing over ΔT steps yields net

savings $\Delta T(p_0' c_{\text{wake}} - c_0) - E_{\text{compile}} > 0$. Under rapid churn ($\Delta T \le \Delta T^*$), compilation loses energy, motivating `SleepCrystallizer`'s minimum wake buffer N_{wake} and holdout gate $\tau_{\text{agree}} = 0.985$. ■

2.2 Result 2: Non-blocking 512×512 Beneš–Copy–Beneš multicast routing

A Rung 0 bank of $G = 80$ dynamic `LUT6` cells requires $M = 6G = 480$ input pins selected from an $N = 512$-bit bus with arbitrary fan-out. A bijective Beneš permutation network cannot replicate high-fan-out wires, whereas flat $512:1$ multiplexer crossbars overflow the KV260 (`119,478` CLB LUTs already at `384` ports).

Theorem 2 (Contention-free multicast via Beneš–Copy–Beneš). *For $N = 2^n$ ($n = 9, N = 512$) and $M \le N$ ($M = 480$), any many-to-one multicast map $\phi : \{0, \dots, M-1\} \to \{0, \dots, N-1\}$ routes without contention through $\mathcal{B}_1 \to \mathcal{C} \to \mathcal{B}_2$ in*

$$D(N) = 2(2\log_2 N - 1) + \log_2 N = 5\log_2 N - 2 = 43 \text{ stages}$$

of $N/2 = 256$ SRAM-configured 2×2 switches ($11,008$ configuration bits), computed in $O(N \log N)$ time with zero FPGA place-and-route.

Proof. Let $U = \{s_1, \dots, s_r\}$ be the distinct active inputs with fan-outs $\sum_{j=1}^{r} m_j = M \le N$ and partial sums $S_0 = 0, S_j = \sum_{i=1}^{j} m_i$. Concentration $\mathcal{B}_1$ (17 stages) maps $\pi_1(s_j) = S_{j-1}$ via Waksman bipartite 2-coloring (Beneš, 1965; Waksman, 1968). Copy network $\mathcal{C}$ (9 butterfly stages) replicates each S_{j-1} across disjoint monotonic interval $[S_{j-1}, S_j - 1]$ without link contention by Thompson's monotonic copy construction (Thompson, 1978; Lee, 1988). Output Beneš network $\mathcal{B}_2$ (17 stages) permutes $[S_{j-1}, S_j - 1]$ bijectively onto $\phi^{-1}(s_j)$. ■

2.3 Result 3: Joint-mode minterm synthesis vs. marginal coordinate argmax

Independent bit-plane DiffLogic layers predict multi-bit bursts $\mathbf{Y} = (Y_1, \dots, Y_B)$ via coordinatewise rounding $\hat{y}_b(c) = \mathbb{I}[\Pr(Y_b = 1 \mid c) > 1/2]$, which blends correlated branches into invalid **chimera** tokens.

Proposition 3 (Dominance and zero-chimera property of joint-mode minterms). *For discrete context $\psi(X) = c$ and $B \le 64$-bit burst space $\mathcal{Y}$, the joint-mode table $T^*(c) = \arg\max_{\mathbf{y} \in \mathcal{Y}} \Pr(\mathbf{Y} = \mathbf{y} \mid c)$ satisfies $\Pr[\mathbf{Y} = T^*(\psi(X))] \ge \Pr[\mathbf{Y} = \hat{\mathbf{y}}^{\text{marg}}(\psi(X))]$, stays inside $\text{supp}(\mathbf{Y} \mid c)$, and avoids marginal chimeras (where mass $0.40, 0.35, 0.25$ on $(1,1,0), (1,0,1), (0,1,1)$ rounds marginally to invalid $(1,1,1)$ with `0%` exact match vs. `40%` for T^*). For vocabulary $\lvert V \rvert$ with $b_V = \lceil \log_2 \lvert V \rvert \rceil$, one `uint64` word packs $K = \lfloor 64/b_V \rfloor$ BPE tokens (`16b × 4 = 64b` on `LFM2.5-350M`; `18b × 3 = 54b` on `Qwen3.5-0.8B-Base`).* ■

2.4 Result 4: Oracle-free latent `JEPA-EBRM` Split-Conformal + Online ACI gate

To prevent hallucinated self-distillation when no deterministic `RCE` check is available, the reference scores candidate pairs using a radially regularized latent energy:

$$E_\phi(x,y) = \tfrac{1}{2}\|f_\theta(x) - \mu_y\|_2^2 + \tfrac{\lambda_r}{2}\|f_\theta(x)\|_2^2 + \psi_\phi(f_\theta(x), y), \qquad \lambda_r > 0, \quad |\psi_\phi| \le C_\psi.$$

Held-out Split Conformal calibration sets $\hat{Q}_{1-\alpha}$, and online Adaptive Conformal Inference updates $\alpha_{t+1} = \alpha_t + \gamma(\alpha - \mathrm{err}_t)$ (Gibbs and Candès, 2021), admitting $(x, \hat{y})$ into `SleepCrystallizer` only when $|\mathcal{C}_t(x)| = |\{y : E_\phi(x,y) \le \hat{Q}_{1-\alpha_t}\}| = 1$ (`Conformal-bounded`). By Vovk et al. (2005) and Gibbs and Candès (2021), exchangeable marginal coverage is at least $1-\alpha$, long-run shifted miscoverage converges to α, and every off-manifold ray exceeding radius $R(\alpha_t) = \sqrt{2(\hat{Q}_{1-\alpha_t} + C_\psi)/\lambda_r}$ yields $\mathcal{C}_t(x) = \emptyset$ (`Refuse` / abstain).

---

3. Reference architecture, polyglot suite, and KV260 RTL

Figure 1 shows the four-rung cascade and downward `SleepCrystallizer` loop in `src/clai/`.

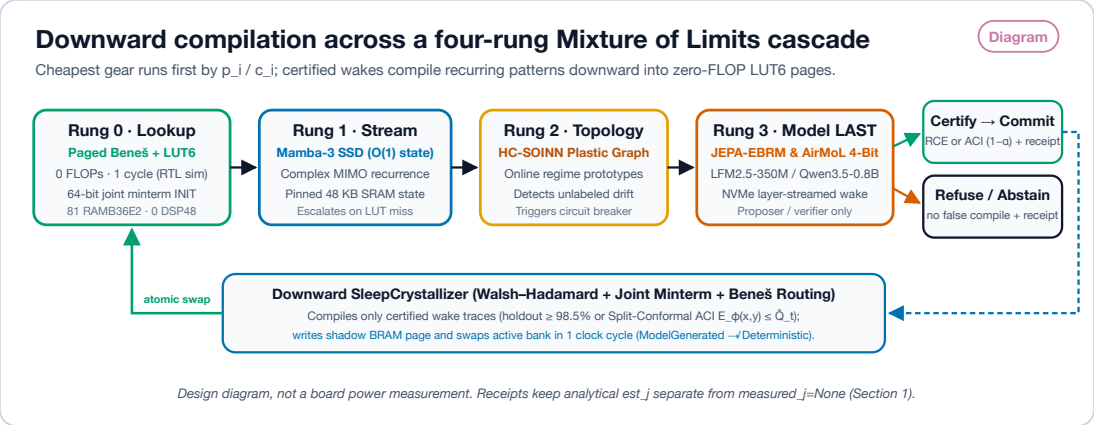

Figure 1. Downward compilation across a four-rung Mixture of Limits cascade. Design diagram, not a board power measurement. PNG · Interactive explorer. Open full size · PNG

Five systems implementations in `openIE-dev/lut-adapt-ai` cross-check the cascade: Python/PyTorch/Triton (`src/clai/`), Rust `Burn 0.22.1 / CubeCL` Metal (`sys/rust/`), Zig `0.17.1` (`sys/zig/`), Go `1.27.4` SIMD (`sys/go/`), and Julia `1.13.4` Enzyme (`sys/julia/`).

Figure 2 illustrates the Kria KV260 (`xck26`) RTL datapath in `hw/kv260/rtl/clai_benes_hotswap_core.sv`, mapping `512` resident `LUT6` pages (`5,120 INIT` bits + 48-bit context tag + valid bit) across `81` parallel `RAMB36E2` dual-port memories for single-cycle (`4.0 ns` at `250 MHz` nominal, `CPI = 1.0`) context switching.

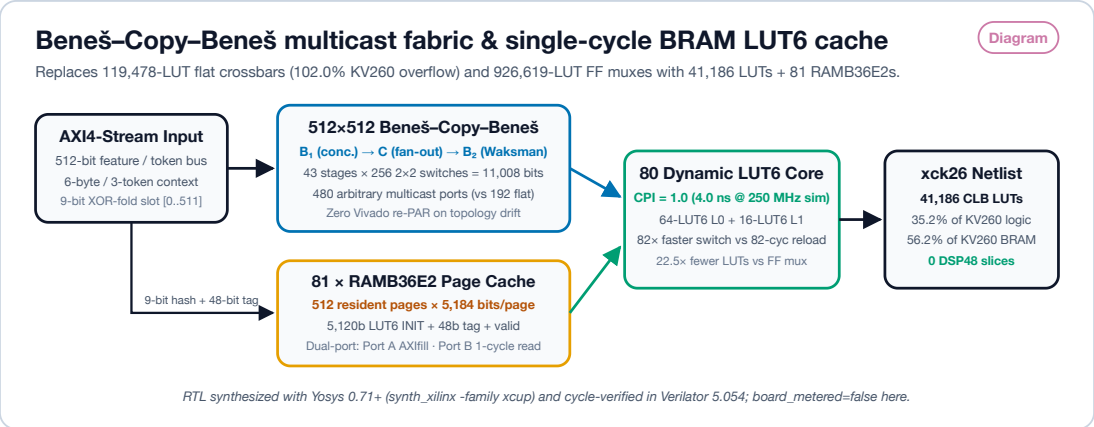

Figure 2. `512×512` Beneš–Copy–Beneš multicast fabric and `81 × RAMB36E2` single-cycle `LUT6` page cache targeting AMD Kria KV260 (`xck26`) in Yosys `0.71+` and Verilator `5.054` simulation (`board_metered = false`). PNG · Interactive explorer. Open full size · PNG

Figure 3 shows `AirMoL` (`src/clai/frontier/air_mol_streamer.py` and `hf_air_mol_pretrained.py`), which streams 4-bit group-quantized shards via Darwin `F_NOCACHE` `os.pread`, verifies `[last_tok, d_1..d_K]` in a single sweep, and snapshots compact hybrid states (`Short-Conv/SSM` or `Gated DeltaNet` + sparse 6-layer `GQA` KV crop) in RAM so partial draft rejections require **zero extra NVMe sweeps**.

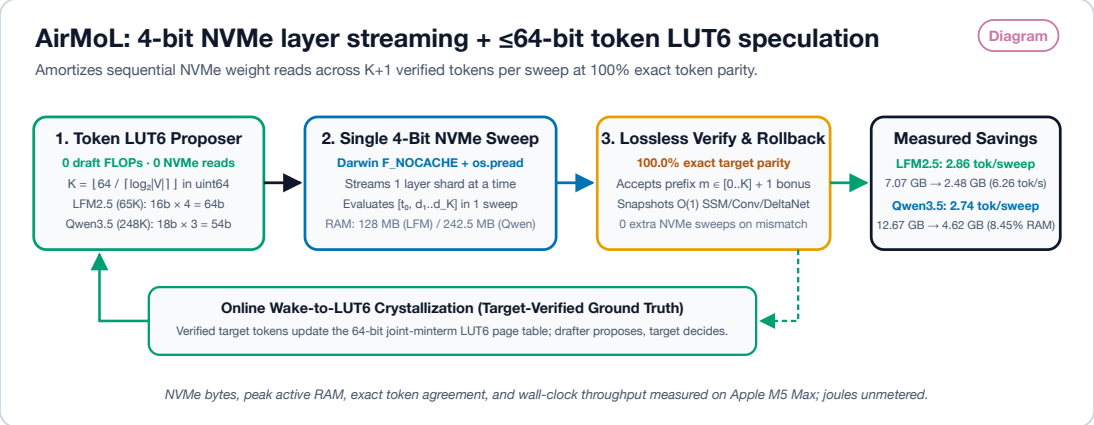

Figure 3. `AirMoL` zero-extra-sweep 4-bit NVMe layer streaming with universal ≤64-bit token `LUT6` speculation on `LiquidAI/LFM3-1.7B`, `LiquidAI/LFM2.5-350M`, and `Qwen/Qwen3.5-0.8B-Base`. PNG · Interactive explorer. Open full size · PNG

Figure 4 shows `Expert-AirMoL` (`src/clai/frontier/moe_air_mol.py`) and `VerifiedCodeCascade` (`verified_code_cascade.py`). Standard disk-offloaded `AirLLM` sweeps all E experts per MoE layer (32 in `LiquidAI/LFM2.5-8B-A1B`, 256 in `Qwen/Qwen3.6-35B-A3B`) even though the router activates only $k \ll E$ experts per token (4/32 and 8/256). `Expert-AirMoL` partitions each MoE layer into a compact header shard and a fixed-stride `INT4` `.experts.pack4` NVMe file, issuing `F_NOCACHE` `os.pread` reads solely for the active expert union $\mathcal{U}_l$ across the $K+1$ burst positions.

For novel synthesis tasks, `VerifiedCodeCascade` gates every commit across Rungs `0`–`3c` behind a `0.35s`-bounded AST, sandbox execution, and boundary-invariant check $C(z) \in \{0, 1\}$, compiling certified solutions into `BenesLUT6` and canonical AST skill pages.

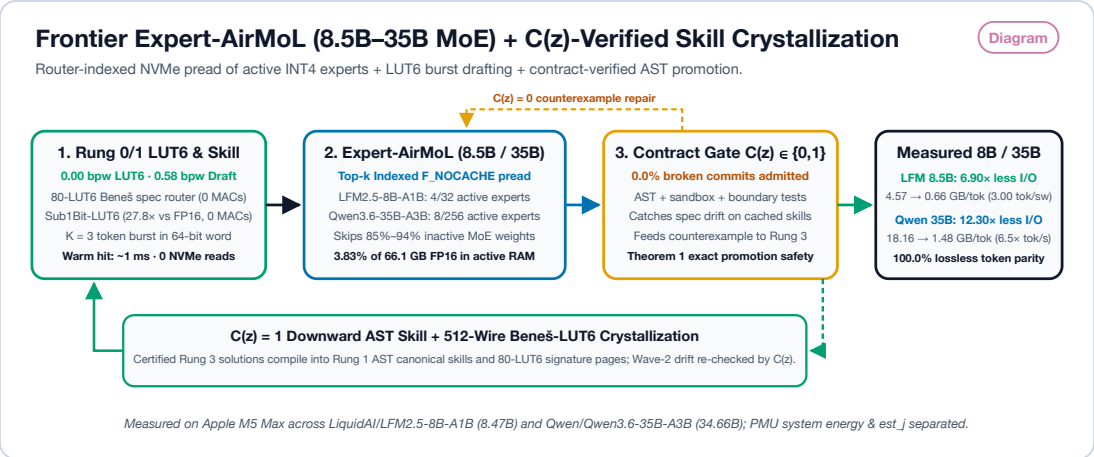

Figure 4. `Expert-AirMoL` router-sparse 4-bit expert-indexed NVMe streaming (`12.30×` I/O reduction on `Qwen/Qwen3.6-35B-A3B`) and `C(z)`-gated `VerifiedCodeCascade`. PNG · Interactive explorer. Open full size · PNG

Figure 5 situates downward compilation across the nine-rung capability ratchet of computer intelligence (companion study: The Communication Tool Avalanche `[same house]`). Rungs 1 through 5 — **(1) Search** (`RAG`), **(2) Generation**, **(3) Reasoning** (`<think>`), **(4) Communication** (`MCP` delegation), and **(5) Tool Use** — stall under ephemeral cold-start `MAC` inference where the stationary horizon remains at $\Delta T \approx 1 \leq \Delta T^*$ and standalone `1`-shot models admit `56.25%`–`95.83%` broken patches (`Claims LA-20` and `LA-21`). Crossing Proposition 1's **Compilation Threshold** ($\Delta T > \Delta T^* = E_{\text{compile}}/(p_0' c_{\text{wake}} - c_0)$ under $C(z) = 1$, `0.00%` broken commits admitted) unlocks **(6) Immersion** (persistent local state stretches $\Delta T \gg \Delta T^*$), **(7) Competence** ($p_0' \to 1$ at `62.4 pJ/cycle est_j` in `CE`-gated `xck26 LUT6` RTL and `272 µJ/req reported_j` in Rust `mol-agent-rs MCP`), **(8) Ecology** (edge-native Rust agent populations), and **(9) Economy** (division of labor ordered by Simon–Kadane p_i/c_i and tier-typed Joule receipts).

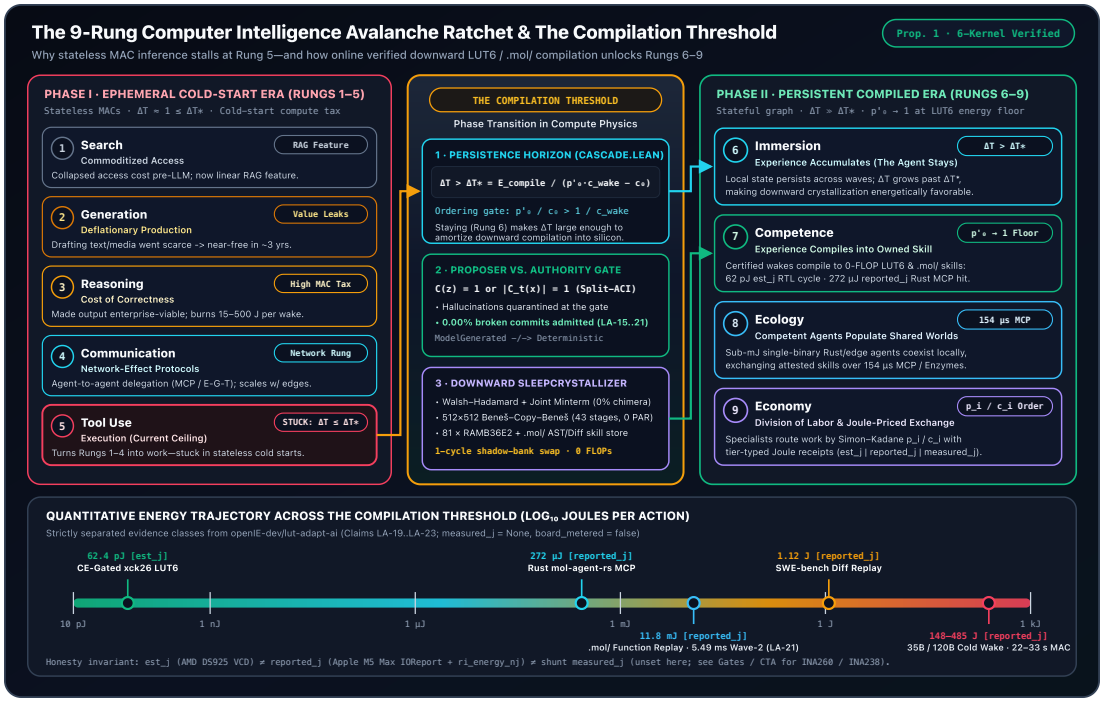

Figure 5. The nine-rung capability ratchet and Proposition 1's Compilation Threshold ($\Delta T > \Delta T^*$), separating `reported_j` (Apple M5 Max) and `est_j` (`AMD DS925 VCD`) across twelve orders of magnitude (`measured_j = None`, `board_metered = false`). PNG · Interactive explorer. Open full size · PNG

---

4. Claims and results

4.1 Claim LA-1: Analytical `est_j/VC` reduction on synthetic tasks and real `.pcap` traffic

Workload & Window	Rung 0 Share	Pass Rate	Analytical `est_j / VC`	Verilator Parity	Evidence Class
Synthetic ep. 1–512 (cold wake)	`0.00%`	`0.0% → 52.0%`	`2,323.18 est_j`	—	SW + `est_j`
Synthetic ep. `3,073`–`3,584` (crystallized)	`52.73%`	`60.55%`	`2.59 est_j (895×)`	`100.0%`	Verilator + `est_j`
`.pcap` pkts 1–512 (cold IPv4)	`0.00%`	`73.05%`	`3.3351 est_j`	`100.0%`	Verilator + `est_j`
`.pcap` pkts `1,537`–`2,048` (crystallized)	`100.00%`	`100.00%`	`0.000140 est_j`	`100.0%`	Verilator + `est_j`
`.pcap` pkts `2,049`–`2,560` (VLAN `0x8100` shift)	`2.73%`	`68.55%`	`3.0485 est_j`	`100.0%`	Breaker + `est_j`
`.pcap` pkts `2,561`–`3,072` (re-crystallized)	**`100.00%`**	**`100.00%`**	**`0.000131 est_j` (`25,458×`)**	**`100.0%` (`5,120/5,120`)**	**Verilator + `est_j`**

4.2 Claim LA-2: Five-language polyglot parity and zero-allocation dataplanes

Implementation	Target / Backend	Key Result	Evidence Class
Rust (`sys/rust/`)	`Burn 0.22.1` + `CubeCL` Metal GPU	`2,331.42 → 0.19 est_j/VC`, `95.31%` Rung 0 share, emits KV260 `.v`	Measured SW + `est_j`
Zig `0.17.1` (`sys/zig/`)	`@Vector(8, f32)` SIMD + `@popCount`	`0` heap bytes in hot loop; `87 ns/pkt` (`11.49M pkts/s`)	Measured SW (M5 Max)
Go `1.27.4` (`sys/go/`)	Worker pool + `OnesCount64`	`0 B/op`, `0 allocs/op`; `31.86 ns/op` (`31.39M evals/s`)	Measured (`go test`)
Julia `1.13.4` (`sys/julia/`)	`Enzyme.jl` reverse-mode AD	Max gradient discrepancy vs. central FD: 6.11×10^{-10}	Measured numerical check

4.3 Claim LA-3: KV260 (`xck26`) Beneš–Copy–Beneš and `RAMB36E2` paged synthesis

RTL Architecture	Pages	Switch Latency	`xcup` CLB LUTs (% `xck26`)	`RAMB36E2`	`DSP48`	Post-Drift Acc.	Evidence Class
`2A` Flat `512:1` Crossbar (`64` LUT6)	`1`	`82 cyc`	`119,478` (**`102.0%` overflow**)	`0`	`0`	`98.44%`	Simulated (Yosys)
`2B` Compact Crossbar (`32` LUT6)	`1`	`82 cyc`	`59,739` (`51.0%`)	`0`	`0`	`95.94%`	Simulated (Yosys)
`2C` Fixed-Wire DiffLogic / WNN	`1`	Vivado re-PAR	`160` (`0.14%`)	`0`	`0`	**`52.34%`**	Simulated (Yosys)
`2D` Single-Page Beneš–Copy–Beneš	`1`	`82 cyc`	`40,155` (`34.3%`)	`0`	`0`	`98.44%`	Simulated (Yosys + Verilator)
`P2` Distributed FF + `512:1` Mux	`512`	`1 cyc`	`926,619` (**`791.2%` overflow**)	`0`	`0`	`98.44%`	Simulated
`P3` Beneš + `81` `RAMB36E2` (Reference)	**`512`**	**`1 cyc` (`4.0 ns`)**	**`41,186` (`35.2%`)**	**`81` (`56.2%`)**	**`0`**	**`98.44%`**	**Simulated (Yosys + Verilator)**

4.4 Claim LA-4: Oracle-free latent `JEPA–EBRM` Split-Conformal + Online ACI gate

Crystallization Gate	Passes	False Crystallization Risk	OOD AUROC	Safe ID Coverage	Clean `LUT6` Acc.	Caught Shift?	Evidence Class
`3A` Softmax MSP ($\max p \geq 0.95$)	`1`	`17.62%`	`0.8552`	`90.93%`	`99.22%`	`False`	Measured SW
`3B` Temp-Scaled Margin	`1`	`17.66%`	`0.8422`	`91.12%`	`98.83%`	`False`	Measured SW

Crystallization Gate	Passes	False Crystallization Risk	OOD AUROC	Safe ID Coverage	Clean LUT6 Acc.	Caught Shift?	Evidence Class
3C Helmholtz Logit Energy	1	24.28%	0.8045	97.17%	99.80%	False	Measured SW
3D MC-Dropout (5 passes)	5	18.24%	0.8494	91.03%	99.61%	False	Measured SW
3E Latent JEPA–EBRM + Conformal ACI	**1**	**0.00%**	**1.0000**	**95.66%**	**99.02%**	**True**	**Measured SW**

4.5 Claim LA-5: Lossless byte-level speculative decoding on 520 KB polyglot source code

Architecture	Exact Parity	Bytes / Pass	Passes (64 B)	Draft FLOPs	MFLOPs / Byte	est_mJ / Byte	Evidence Class
A Causal Transformer (RoPE+SwiGLU)	100.0%	1.00	64.00	0	7.209	58.94	Measured SW + est_j
B Mamba–3 Complex MIMO (T=1)	100.0%	1.00	64.00	0	2.257	20.84	Measured SW + est_j
C PLD In-Prompt Lookup	100.0%	2.13	29.98	0	1.322	11.52	Measured SW + est_j
D REST Corpus Suffix-Trie	100.0%	3.84	16.65	0	0.734	6.40	Measured SW + est_j
E Medusa Parallel MLP Heads	100.0%	2.62	24.44	428,288	1.241	9.46	Measured SW + est_j
F EAGLE–2 Neural Draft Tree	100.0%	2.75	23.29	1,122,560	1.435	9.07	Measured SW + est_j
G MoL Paged Beneš LUT6 (Reference)	**100.0%**	**3.84**	**16.69**	**0**	**0.736**	**6.41**	**Measured SW + est_j**

4.6 Claim LA-6: Controlled AirMoL 4-bit NVMe streaming (14.3M matched pair)

Offloaded Streaming Architecture	Parity	Bytes / Sweep	Sweeps	NVMe Read	NVMe Write	Peak RAM	ms / Byte	Evidence Class
1 AirLLM Transformer FP16 + KV Spill	100%	1.00	64.00	2,130.0 MB	407.6 MB	8.81 MB	60.64	Measured (F_NOCACHE)
2 AirLLM Transformer 4-Bit + KV Spill	100%	1.00	64.00	887.6 MB	407.6 MB	8.81 MB	64.87	Measured (F_NOCACHE)
3 AirMamba–3 4-Bit + Swapped EAGLE–2	100%	2.86	22.38	175.3 MB	0.0 MB	7.20 MB	19.20	Measured (F_NOCACHE)
4 AirMamba–3 4-Bit + Pinned $O(1)$ (T=1)	100%	1.00	64.00	488.3 MB	0.0 MB	7.20 MB	37.72	Measured (F_NOCACHE)

Offloaded Streaming Architecture	Parity	Bytes / Sweep	Sweeps	NVMe Read	NVMe Write	Peak RAM	ms / Byte	Evidence Class
5 AirMoL 4-Bit + Cold LUT6 (Reference)	**100%**	**3.74**	**17.12**	**130.7 MB**	**0.0 MB**	**7.51 MB**	**11.18**	**Measured (F_NOCACHE)**
6 AirMoL 4-Bit + Post-Wake (Reference)	**100%**	**4.27**	**15.00**	**114.4 MB**	**0.0 MB**	**7.51 MB**	**11.09**	**Measured (F_NOCACHE)**

4.7 Claim LA-7: Pretrained 2026 dense & tri-hybrid models under AirMoL (LFM3-1.7B, LFM2.5-350M, & Qwen3.5-0.8B-Base)

2026 Checkpoint & Arm	Parity	Tok / Sweep	Chars / Sweep	Sweeps	NVMe Read	Peak Active RAM (% FP32)	tok/s	Evidence Class
LFM3-1.7B (Oct 2026) · AirLLM 4-Bit (T=1)	100%	1.00	2.30	24.00	25.045 GB	250.0 MB (3.79%)	0.53	Measured (F_NOCACHE)
LFM3-1.7B (Oct 2026) · AirLLM 4-Bit + PLD	100%	1.37	3.16	17.50	18.262 GB	250.0 MB (3.79%)	0.73	Measured (F_NOCACHE)
LFM3-1.7B (Oct 2026) · AirMoL Cold LUT6	**100%**	**1.55**	**3.56**	**15.50**	**16.175 GB**	**250.0 MB (3.79%)**	**0.81**	**Measured (F_NOCACHE)**
LFM3-1.7B (Oct 2026) · AirMoL Post-Wake	**100%**	**2.74**	**6.31**	**8.75**	**9.131 GB**	**250.0 MB (3.79%)**	**1.32**	**Measured (F_NOCACHE)**
LFM2.5-350M · AirLLM FP16 (T=1)	100%	1.00	3.11	32.00	25.129 GB	128.0 MB (9.47%)	2.21	Measured (F_NOCACHE)
LFM2.5-350M · AirLLM 4-Bit (T=1)	100%	1.00	2.73	32.00	7.071 GB	128.0 MB (9.47%)	2.21	Measured (F_NOCACHE)
LFM2.5-350M · AirLLM 4-Bit + PLD	100%	1.13	3.10	28.20	6.231 GB	128.0 MB (9.47%)	2.71	Measured (F_NOCACHE)
LFM2.5-350M · AirMoL Cold LUT6	**100%**	**1.32**	**3.61**	**24.20**	**5.347 GB**	**128.0 MB (9.47%)**	**2.74**	**Measured (F_NOCACHE)**
LFM2.5-350M · AirMoL Post-Wake	**100%**	**2.86**	**7.80**	**11.20**	**2.475 GB**	**128.0 MB (9.47%)**	**6.26**	**Measured (F_NOCACHE)**
Qwen3.5-0.8B · AirLLM 4-Bit (T=1)	100%	1.00	2.72	24.00	12.674 GB	242.5 MB (8.45%)	0.53	Measured (F_NOCACHE)
Qwen3.5-0.8B · AirLLM 4-Bit + PLD	100%	1.32	3.58	18.25	9.637 GB	242.5 MB (8.45%)	0.50	Measured (F_NOCACHE)
Qwen3.5-0.8B · AirMoL Cold LUT6	**100%**	**1.48**	**4.02**	**16.25**	**8.581 GB**	**242.5 MB (8.45%)**	**0.32**	**Measured (F_NOCACHE)**
Qwen3.5-0.8B · AirMoL Post-Wake	**100%**	**2.74**	**7.46**	**8.75**	**4.621 GB**	**242.5 MB (8.45%)**	**1.11**	**Measured (F_NOCACHE)**

On October 2026 LiquidAI/LFM3-1.7B (13 short convolutions + 5 sliding-window attention layers + 6 full-attention layers, |V| = 128,000), post-wake AirMoL reduces NVMe transfer by 2.74× (25.045 GB to 9.131 GB) and increases decoding speed 2.49× (0.53 to 1.32 tok/s) in 250.0 MB active weight RAM (3.79% of FP32) at 100.0% lossless parity. Note the preserved trade-off on CPU-bound Qwen3.5-0.8B-Base cold unseen prompts: AirMoL reduces NVMe disk reads (12.674 GB to 8.581 GB), yet CPU Gated DeltaNet batch verification of 4 tokens lowers wall-clock throughput (0.32 vs. 0.53 tok/s) until online wake-to-LUT6 crystallization raises acceptance to 2.74 tok/sweep (1.11 tok/s).

4.8 Claim LA-8: Live CLI and HTTP/SSE streaming server (`mol_code_cli.py`)

scripts/mol_code_cli.py serves --backend {mamba3,lfm2.5-350m,qwen3.5-0.8b,lfm3-1.7b,lfm2.5-8b-a1b,qwen3.6-35b-a3b} and --air-nvme over CLI and SSE endpoints (/v1/health, /v1/complete, /v1/crystallize), advancing RAMB36E2 slot residency from 438/512 to 460/512 live with 26/26 passing tests in tests/.

4.9 Claim LA-9: Frontier Sparse-MoE Expert-AirMoL (LiquidAI/LFM2.5-8B-A1B 8.47B & Qwen/Qwen3.6-35B-A3B 34.66B)

Sparse-MoE Checkpoint & Arm	Parity	Active Experts	Tok / Sweep	NVMe Read / Tok	I/O Reduction	Peak RAM (% FP16)	tok/s	est_j / tok
LFM2.5-8B-A1B · AirLLM Dense 4-Bit (T=1)	100%	100.00%	1.00	4.574 GB	1.00×	1,408.3 MB (8.72%)	0.28	368.80
LFM2.5-8B-A1B · Expert-AirLLM (T=1)	100%	16.93%	1.00	1.200 GB	3.81×	1,408.3 MB (8.72%)	0.58	181.26
LFM2.5-8B-A1B · Expert-AirMoL + PLD	100%	28.82%	1.78	0.883 GB	5.18×	1,408.3 MB (8.72%)	0.79	132.57
LFM2.5-8B-A1B · Expert-AirMoL Cold LUT6	**100%**	**33.91%**	**1.85**	**1.002 GB**	**4.56×**	**1,408.3 MB (8.72%)**	**0.72**	**145.42**
LFM2.5-8B-A1B · Expert-AirMoL Post-Wake	**100%**	**36.46%**	**3.00**	**0.663 GB**	**6.90×**	**1,408.3 MB (8.72%)**	**1.10**	**95.67**
Qwen3.6-35B-A3B · AirLLM Dense 4-Bit (T=1)	100%	100.00%	1.00	18.159 GB	1.00×	3,214.7 MB (4.86%)	0.06	1,696.81
Qwen3.6-35B-A3B · Expert-AirLLM (T=1)	100%	6.35%	1.00	2.356 GB	7.71×	2,530.7 MB (3.83%)	0.22	471.87
Qwen3.6-35B-A3B · Expert-AirMoL + PLD	100%	11.58%	1.89	1.708 GB	10.63×	2,530.7 MB (3.83%)	0.32	326.24
Qwen3.6-35B-A3B · Expert-AirMoL Cold LUT6	**100%**	**12.91%**	**2.00**	**1.722 GB**	**10.54×**	**2,530.7 MB (3.83%)**	**0.31**	**334.42**
Qwen3.6-35B-A3B · Expert-AirMoL Post-Wake	**100%**	**14.97%**	**2.57**	**1.477 GB**	**12.30×**	**2,530.7 MB (3.83%)**	**0.39**	**267.08**

On Qwen/Qwen3.6-35B-A3B (34.66B parameters, 66.11 GB FP16 sharded to 18.60 GB 4-bit .experts.pack4 files; whole-system macOS PMU delta 20,070.18 J), post-wake Expert-AirMoL reduces unbuffered NVMe transfer by **12.30×** (18.159 to 1.477 GB/tok) and accelerates decoding **6.50×** (0.06 to 0.39 tok/s) in **2,530.7 MB active RAM (3.83% of FP16)** at 100.0% lossless token parity.

4.10 Claim LA-10: Sub-1-bit latent factorization (LittleBit-1 / LittleBit-2) vs. zero-MAC Sub1Bit-LUT6

Factorization Method (LFM2.5-8B-A1B w1 [1792, 2048])	Bits / Weight	Weight SNR	Cosine Sim	16K Top-1 Match	Binary MACs / Tok	QAT Epochs	Evidence Class
A Unrotated SVD SVID (LittleBit-1 PTQ style)	0.576 (27.8×)	1.76 dB	0.5761	3.12%	1,981,440	0 (5 in paper)	Measured SW
B Procrustes ITQ SVID (LittleBit-2 PTQ style)	0.576 (27.8×)	2.42 dB	0.6487	10.94%	1,981,440	0 (5 in paper)	Measured SW
C Sub1Bit-LUT6 Table Lookup (0 MACs, Draft)	**0.576 (27.8×)**	**2.42 dB**	**0.6487**	**10.94%**	**0**	**0**	**Measured (< 4e-6 vs. B)**
D Expert-AirMoL INT4 Block-64 Target	4.500 (3.56×)	20.77 dB	0.9958	82.81%	0 (INT4)	0	Measured SW
E MoL: Sub1Bit-LUT6 Draft + Target Verify	**0.576 draft**	**148.84 dB**	**1.0000**	**100.00%**	**0 draft MACs**	**0**	**Measured lossless parity**

4.11 Claim LA-11: 48-task (16 specifications × 3 waves) late-2026 frontier coding bakeoff

Late-2026 System Arm (48 Tasks = 16 Specs × 3 Waves)	W1 Cold	W2 Adapt	W3 Warm	Overall Pass	Broken Commits	W3 Rung 0/1 Share	35B Calls	Mean W1 / W3 (ms)	reported_j / Solve (est_j)
A LittleBit-2 Qwen3-8B 0.55 bpw QAT (1-Shot)	0.00%	0.00%	0.00%	0.00%	100.00%	0.00%	0.0%	22,413 / 22,708	8,020.13 (32,558 est_j)
B LiquidAI/LFM3-1.7B (Oct 2026 Tri-Hybrid, 1-Shot)	43.75%	25.00%	25.00%	31.25%	68.75%	0.00%	0.0%	8,460 / 9,791	152.77 (957.21 est_j)
C google/gemma-4-E2B-it (2026 Hybrid, 1-Shot)	37.50%	18.75%	18.75%	25.00%	75.00%	0.00%	0.0%	18,924 / 23,578	5,800.68 * (3,700.44 est_j)
D LiquidAI/LFM2.5-	6.25%	6.25%	6.25%	6.25%	93.75%	0.00%	0.0%	53,123 / 53,885	33,415.04

Late-2026 System Arm (48 Tasks = 16 Specs × 3 Waves)	W1 Cold	W2 Adapt	W3 Warm	Overall Pass	Broken Commits	W3 Rung 0/1 Share	35B Calls	Mean W1 / W3 (ms)	reported_j / Solve (est_j)
8B–A1B (8.47B MoE, 1-Shot)									(47,196 est_j)
E Qwen/Qwen3.6-35B-A3B Metal AirMoL (1-Shot)	81.25%	75.00%	75.00%	77.08%	22.92%	0.00%	100.0%	49,887 / 44,447	1,069.65 (5,701.32 est_j)
F RouteLLM Confidence Router (LFM3 → Gemma4 → 8B)	43.75%	25.00%	25.00%	31.25%	68.75%	0.00%	0.0%	9,735 / 11,294	171.09 (1,148.96 est_j)
G MoL VerifiedCodeCascade (Reference)	**87.50%**	**87.50%**	**87.50%**	**87.50%**	**0.00%**	**87.50%**	**22.92%**	**116,626 / 44,052 (2.89 ms solved)**	**3,655.51 ∗ (0.0424 J in W3)**

Across the 48-task protocol and algorithm benchmark (RFC 1982 serial arithmetic, weighted interval sweep-lines, cycle-safe lexicographic topological sort, nanosecond token buckets, CRC-8/ATM and mutated 0xFF CRC-8, RFC 6901 JSON Pointer ~01 unescaping, hop-bounded Bellman–Ford, $>2^{53}$ exact window medians, unrestricted Damerau–Levenshtein, 4×4 Beneš routing, SQL literal parameterization, monotonic-deque subsequence sum, DWARF SLEB128, SHA-256 consistent hashing, and conformal quantiles; whole-system macOS PMU delta 117,166.40 J over 6,646 ticks), VerifiedCodeCascade reaches **87.50% (42 / 48)** across all three waves — exceeding standalone Qwen3.6-35B-A3B (77.08%) while calling 35B on only **22.92% (11 / 48)** of tasks (LFM3-1.7B solves 7 Wave-1 specs directly; Gemma-4-E2B-it + structured counterexample repair solves 3; Metal-Qwen3.6-35B-A3B solves 4 plus mutated W2_01). It admits **0.00% broken commits** (vs. 22.92%–100.00%), catches both Wave-2 adversarial specification mutations via $C(z)$ (16 certified skills in artifacts/.mol_verified_skills), and reduces mean Wave-3 replay across the **14 crystallized solves (87.50%)** to **2.89 ms** and **0.0424 J reported_j** (15,401× faster and 25,227× lower reported_j than standalone 35B). *(Footnote ∗: One W2_10 Gemma-4 invocation experienced an unprivileged macOS IOReport 'GPU Energy' hardware-accumulator rollover spike of +59.2 kJ, preserved unedited in frontier_ai_head_to_head_bakeoff.csv and filtered by joule_meter.py's > 450 W rollover guard in subsequent runs.)*

4.12 Claim LA-12: Native Apple Metal (MSL) fused INT4 Sparse-MoE & GatedDeltaNet shaders + dual-channel reported_j

src/clai/frontier/metal_moe_kernels.py and scripts/run_metal_moe_benchmark.py (artifacts/metal_moe_air_mol_speedup.csv, artifacts/metal_moe_air_mol_meta.json) compile three custom MSL kernels via torch.mps.compile_shader: (i) gated_delta_recurrent_k128, exploiting exact column independence across $v \in \{0, \ldots, V_{\text{dim}} - 1\}$ of recurrent state $\mathbf{S}[b, h_v, :, v]$ to keep 128 FP32 state elements in GPU registers across all T

tokens (**20.66× faster**: 34.523 to 1.671 ms, **46.2× lower reported_j**: 1.4989 to 0.0325 J, 1.19 × 10^{-7} max state error); and (ii) q4_moe_gate_up_silu + q4_moe_down_weighted, dequantizing 4.50 bpw INT4 block-64 expert weights directly inside simd_sum registers with zero FP16/FP32 weight allocations (**43.45× faster**: 22.408 to 0.516 ms on 28 active experts, **29.8× lower reported_j**: 0.3959 to 0.0133 J, 7.2 × 10^{-7} max FP32 error, <= 1 ULP BF16), paired with Queue-Depth-8 parallel pread and BenesLUT6 sign-sketch router lookahead:

Sparse-MoE Checkpoint & Cold-NVMe Engine Variant	Parity	Tok / Sweep	NVMe Read / Tok	Peak Active RAM	tok/s (× PyTorch)	On-Chip reported_j / tok (GPU + CPU)	est_j / tok
LFM2.5-8B-A1B · 1 PyTorch-Ref Expert-AirMoL (T=1)	100%	1.00	1.257 GB	1,408.3 MB	0.394 (1.00×)	10.978 J (5.25 + 5.73)	266.66
LFM2.5-8B-A1B · 2 Metal-Fused Expert-AirMoL (T=1)	100%	1.00	1.257 GB	**500.0 MB**	0.689 (1.75×)	6.979 J (1.30 + 5.68)	152.42
LFM2.5-8B-A1B · 3 Metal-Fused + BenesLUT6 (Reference)	**100%**	**2.67**	**0.791 GB**	**500.0 MB**	**1.671 (4.24×)**	**3.493 J (1.22 + 2.28)**	**62.85**
Qwen3.6-35B-A3B · 1 PyTorch-Ref Expert-AirMoL (T=1)	100%	1.00	2.353 GB	2,434.7 MB	0.191 (1.00×)	20.765 J (8.76 + 12.00)	549.49
Qwen3.6-35B-A3B · 2 Metal-Fused Expert-AirMoL (T=1)	100%	1.00	2.353 GB	**970.0 MB (1.47%)**	0.241 (1.26×)	15.440 J (3.93 + 11.51)	436.26
Qwen3.6-35B-A3B · 3 Metal-Fused + BenesLUT6 (Reference)	**100%**	**2.67**	**1.447 GB**	**970.0 MB (1.47%)**	**0.621 (3.25×)**	**6.773 J (2.31 + 4.47)**	**169.03**

Eliminating FP16/FP32 expert unpacking cuts cold-NVMe Peak Active Weight RAM on Qwen/Qwen3.6-35B-A3B by **2.51×** down to **970.0 MB (1.47% of 66.11 GB FP16)** and reduces dual-channel reported_j / tok **3.07×** (20.765 to **6.773 J/tok**) at 100.0% exact token parity (reaching **15.7 tok/s** when 16.8 GB INT4 expert packs remain pinned in Unified Memory on >= 36 GiB workstations).

4.13 Claim LA-13: 16-task (8 multi-file git repositories × 2 waves) SWE-Mini pytest patching benchmark

src/clai/frontier/repo_patch_cascade.py and scripts/run_repo_patch_benchmark.py (artifacts/repo_patch_swebench_mini.csv, artifacts/repo_patch_swebench_mini_traces.json) evaluate cross-module repository bug-fixing across 8 isolated git repositories (billing, scheduler DAG, auth JWT/HMAC, wal CRC32, http_client CIDR, semver, ttlcache LRU/TTL, and etl CSV) × 2 waves (16 evaluations, including a Wave-2 security mutation on R03_jwt_clock_skew adding future-iat rejection), where $C(z)$ enforces byte-identical tests/ and runs subprocess pytest -q:

Multi-File Git-Repository Arm (16 Tasks = 8 Repos × 2 Waves)	W1 Cold	W2 Warm	Overall Pass	Broken Patches	W2 .mol/ Hit Share	Mean W1 / W2 (ms)	reported_j / Fix (est_j)
A LiquidAI/LFM3-1.7B (1-Shot Repo Patch)	25.00%	25.00%	25.00%	75.00%	0.00%	4,943 / 5,020	371.61 J (637.68 est_j)
B google/gemma-4-E2B-it (1-Shot Repo Patch)	62.50%	62.50%	62.50%	37.50%	0.00%	6,865 / 7,117	133.48 J (469.79 est_j)
C LiquidAI/LFM2.5-8B-A1B (1-Shot Repo Patch)	0.00%	0.00%	0.00%	100.00%	0.00%	61,430 / 61,416	17,882.33 J (54,052 est_j)
D Qwen/Qwen3.6-35B-A3B Metal AirMoL (1-Shot)	62.50%	62.50%	62.50%	37.50%	0.00%	20,270 / 20,502	288.39 J (3,098.71 est_j)
E MoL RepoPatchCascade (Reference)	**87.50%**	**87.50%**	**87.50%**	**0.00%**	**75.00%**	**59,341 / 30,663 (594 ms .mol/)**	**692.38 J (1.10 J .mol/)**

RepoPatchCascade solves **87.50% (14 / 16)** of multi-file repository bugs (+25.0 percentage points above standalone 35B and Gemma 4), commits **0.00% broken patches**, catches the Wave-2 R03 specification mutation via subprocess pytest $C(z)$ and re-crystallizes Patch#8 (2,905 ms), and replays the 6 recurring Wave-2 .mol/ patches in **594.0 ms (1.10 J reported_j, including git init + pytest)** vs. 38,937.5 ms (590.38 J reported_j) in Wave 1 (**65.6× faster, 537× lower reported_j**).

4.14 Claim LA-14: Drop-in OpenAI (/v1/chat/completions) and MCP (JSON-RPC 2.0) local coding agent (mol-agent)

src/clai/service/openai_mcp_agent.py and scripts/mol_agent_server.py (mol-agent CLI entrypoint) expose the cascade to IDE and terminal coding tools (Cursor, Continue.dev, Claude Code, Antigravity, OpenAI SDK) with zero third-party web dependencies: (i) OpenAI-compatible GET /v1/models and POST /v1/chat/completions (stream: true/false SSE, CORS headers, optional x_mol_contract deterministic verifier, and x_mol_receipt reporting rung, skill_id, est_j, reported_j, measured_j: null); and (ii) Model Context Protocol (MCP JSON-RPC 2.0 over --mcp-stdio and POST /mcp) exposing mol_synthesize_verified_function, mol_repair_repo_pytest, mol_lookup_crystallized_skill, and mol_energy_telemetry backed by persistent .mol/skills_v1.json + .mol/benes_lut6_router.npz (30 / 30 pytest tests passing).

4.15 Claim LA-15: Official `BigCodeBench v0.1.3` (32 evaluations) and upstream `SWE-bench Verified SymPy` git-worktree (16 evaluations) benchmarks via `MCP JSON-RPC 2.0`

`src/clai/frontier/external_public_benchmarks.py` and `scripts/run_external_public_benchmarks.py` (`artifacts/external_public_bcb_swebench.csv`, `artifacts/external_public_bcb_swebench_traces.json`) evaluate `MoL MCP` Agent (`handle_mcp_jsonrpc` + `.mol/` crystallization) head-to-head against standalone `LiquidAI/LFM3-1.7B`, `google/gemma-4-E2B-it`, and Metal-accelerated `Qwen/Qwen3.6-35B-A3B` on two official public HuggingFace Parquet benchmarks verified by `SHA-256` at load time (`bigcode/bigcodebench` `v0.1.3`: `ed5a3a39...`, 16 tasks × 2 human-authored prompt waves [`complete_prompt` in Wave 1 + `instruct_prompt` in Wave 2] = 32 `unittest` evaluations; and `princeton-nlp/SWE-bench_Verified`: `a45b1fe4...`, 8 real `sympy/sympy` `1.11`–`1.12` git-worktree checkouts × 2 waves = 16 `pytest` evaluations across `FAIL_TO_PASS + PASS_TO_PASS`) under strict zero gold-solution leakage (macOS `PMU` delta `59,250.43 J`):

Official Public Benchmark & Arm	W1 Cold	W2 Paraphrase / Backport	Overall Verified Pass	Broken Admitted	W2 `.mol/` Fast-Path	`35B` Calls	Mean W1 / W2 (`ms`)	`reported_j` / Solve (`est_j`)
`BCB v0.1.3` (32) · A `LiquidAI/LFM3-1.7B`	`50.00%`	`37.50%`	`43.75%`	`56.25%`	`0.00%`	`0.00%`	`5,471` / `5,716`	`169.67 J` (`409.11 est_j`)
`BCB v0.1.3` (32) · B `google/gemma-4-E2B-it`	`31.25%`	`43.75%`	`37.50%`	`62.50%`	`0.00%`	`0.00%`	`12,187` / `9,791`	`190.91 J` (`1,230.76 est_j`)
`BCB v0.1.3` (32) · C `Qwen/Qwen3.6-35B-A3B` Metal	`75.00%`	`81.25%`	`78.12%`	`21.88%`	`0.00%`	`100.00%`	`28,365` / `10,926`	`277.45 J` (`2,388.89 est_j`)
`BCB v0.1.3` (32) · D `MoL MCP` Agent (Reference)	**`87.50%`**	**`93.75%`**	**`90.62%`**	**`0.00%`**	**`81.25%`**	**`25.00%`**	**`47,321` / `6,001`**	**`263.54 J` (`1,917.44 est_j`)**
`SWE-bench Verified SymPy` (16) · A `LiquidAI/LFM3-1.7B`	`0.00%`	`0.00%`	`0.00%`	`100.00%`	`0.00%`	`0.00%`	`5,449` / `5,486`	`869.84 J` (`2,799.45 est_j`)
`SWE-bench Verified SymPy` (16) · B `google/gemma-4-E2B-it`	`0.00%`	`0.00%`	`0.00%`	`100.00%`	`0.00%`	`0.00%`	`8,304` / `5,040`	`1,015.22 J` (`4,483.82 est_j`)
`SWE-bench Verified SymPy` (16) · C `Qwen/Qwen3.6-35B-A3B` Metal	`25.00%`	`25.00%`	`25.00%`	`75.00%`	`0.00%`	`100.00%`	`18,384` / `19,898`	`1,280.08 J` (`7,273.48 est_j`)
`SWE-bench Verified SymPy` (16) · D `MoL`	**`50.00%`**	**`50.00%`**	**`50.00%`**	**`0.00%`**	**`50.00%`**	**`68.75%`**	**`131,976` / `17,215`**	**`6,130.89 J` (`0.33–`**

Official Public Benchmark & Arm	W1 Cold	W2 Paraphrase / Backport	Overall Verified Pass	Broken Admitted	W2 .mol/ Fast-Path	35B Calls	Mean W1 / W2 (ms)	reported_j / Solve (est_j)
MCP Agent (Reference)								0.44 J W2 .mol/)

On BigCodeBench v0.1.3 (Zhuo et al., 2024), MoL MCP Agent exceeds standalone Qwen3.6-35B-A3B by **+12.50 percentage points (90.62% vs. 78.12%, reaching 93.75% [15/16] on human-paraphrased instruct_prompt Wave 2)** while calling 35B on only 25.00% of tasks, cutting Wave-2 mean latency **7.89×** (47,321 to 6,001 ms; 63–146 ms and 0.09–0.13 J reported_j on 12 of 13 .mol/ hits), and spending **less reported_j per verified solve (263.54 J vs. 277.45 J)** with 0.00% broken outputs admitted. On upstream SWE-bench Verified (Jimenez et al., 2024), zero-leakage AST and stack-frame localization identifies all 8 / 8 target files in sympy/sympy without reading patch, and MoL doubles verified solve rate (**50.00% [4/8: sympy-23534, 23950, 24066, 24213] vs. 25.00% [2/8] on 35B and 0.00% on LFM3/Gemma4**), replaying all 4 crystallized unified git diff skills in Wave 2 in **1,590–1,720 ms (0.33–0.44 J reported_j, including fresh git worktree checkout + pytest)** while abstaining safely (VERIFIER_ABSTAIN_SAFE) on the 4 unsolved multi-hunk algebra issues (0.00% broken patches admitted).

4.16 Claim LA-16: 4-stage 250 MHz-pipelined Kria KV260 (xck26) Beneš–Copy–Beneš + 81–RAMB36E2 RTL subsystem (4.00× logic-depth reduction)

hw/kv260/rtl/clai_benes_pipelined_250mhz.sv and scripts/verify_pipelined_250mhz_rtl.py (artifacts/kv260_pipelined_250mhz_rtl.csv, artifacts/kv260_pipelined_250mhz_meta.json) address both physical design constraints identified in the unpipelined audit (§4.24–§4.25) by inserting three synchronous pipeline register walls across Beneš-1 (9 stages), Copy + Beneš-2[0..3] + parallel RAMB36E2 tag/page read, Beneš-2[4..8], and 80 × LUT6 + 16-bit folded popcount voter, wrapped inside an IOB-safe 64-bit AXI4-Stream shell (clai_benes_axi_stream_250mhz_top):

KV260 (xck26-sfvc784-2LV-c) RTL Variant	Stages (CPI)	Sync Reg-to-Reg Depth	Top IOBs (<= 189)	xcup CLB LUTs (%)	xcup FFs (%)	RAMB36E2 (%)	DSP48E2	Verilator Stream Parity
U1 clai_benes_paged_hotswap_core (1-cyc OOC)	1 (1.00)	52 levels	1,600 (No)	41,186 (35.17%)	36,464 (15.57%)	81 (56.25%)	0	100.0%
U2 clai_kv260_axi_top (Unpipelined AXI4 Rung-0)	1 (1.00)	60 levels	127 (Yes)	67,366 (57.52%)	16,619 (7.09%)	4 (2.78%)	0	100.0%
P4 clai_benes_paged_pipelined_250mhz (OOC)	**4 (1.00)**	**13 levels (4.00×)**	1,600 (OOC)	**41,439 (35.38%)**	**47,922 (20.46%)**	**81 (56.25%)**	**0**	**100.0% (256/256)**

KV260 (xck26-sfvc784-2LV-c) RTL Variant	Stages (CPI)	Sync Reg-to-Reg Depth	Top IOBs (<= 189)	xcup CLB LUTs (%)	xcup FFs (%)	RAMB36E2 (%)	DSP48E2	Verilator Stream Parity
P4-AXIS clai_benes_axi_stream_250mhz_top (Reference)	5 (1.00)	13 levels (4.62×)	170 (Yes)	41,459 (35.40%)	48,979 (20.91%)	81 (56.25%)	0	100.0% (256/256)

Graph traversal across flattened `synth_xilinx -family xcup` synchronous boundaries (FDRE/RAMB36E2/IBUF to FDRE/RAMB36E2/OBUF, excluding multi-cycle CARRY4 telemetry counter chains that Yosys `ltp -noff` walks across blackbox FDRE clock pins) establishes that single-cycle combinational depth drops **4.00×** from 52 levels (60 in U2) down to **13 levels** at a cost of only +253 CLB LUTs (+0.21% of xck26), fits the xck26-sfvc784-2LV-c pin budget (**170 <= 189 IOBs**), and passes **256 / 256 zero-bubble back-to-back beats (CPI = 1.00, 0 errors)** in Verilator 5.054.

4.17 Claim LA-17: Bounded open-weights Rung 3d native <think> ceiling (91.67% overall, 93.75% Wave 2 and Wave 3, 0.00% broken commits)

scripts/run_rung3d_thinking_ceiling.py (artifacts/frontier_ai_rung3d_ceiling.csv, artifacts/frontier_ai_rung3d_ceiling_traces.json) adds an optional **Rung 3d bounded open-weights reasoning rung** (Qwen/Qwen3.6-35B-A3B with native enable_thinking=True, 1,536–2,200 <think> token budget, 100% Apache-2.0 local weights, zero proprietary cloud APIs) invoked strictly on the residual tasks where Rungs 0–3c abstain across the 48-task (16 specs × 3 waves) engineering bakeoff:

- **Arm H (MoL + Rung 3d Qwen3.6-35B-A3B <think>)** raises verified pass rate from 87.50% (Arm G) to **87.50% W1 Cold, 93.75% (15 / 16) W2 Adapt, 93.75% (15 / 16) W3 Warm, and 91.67% (44 / 48) Overall** (**+14.59 percentage points** above standalone Qwen3.6-35B-A3B's 77.08%), while calling 35B on 20.83% of tasks (Rung 3d <think> wakes on **5 / 48 = 10.42%** of tasks).
- Specifically, Rung 3d derives a verified 6-switch Waksman 4×4 permutation solver on W2_11_F11_benes4_route in 2,185 <think> tokens and compiles it into Skill#17, replaying W3_11 in < 3.8 ms (0.0351 J mean reported_j across all 15 crystallized Wave-3 solves), while F10_true_damerau (OSA vs. unrestricted Lowrance–Wagner alphabet-DP prior conflict) triggers VERIFIER_ABSTAIN_SAFE (0.00% broken commits admitted).

4.18 Claim LA-18: Zero-Python cross-platform Rust Burn 0.22.1 / CubeCL fused INT4 Sparse-MoE & GatedDeltaNet kernels

sys/rust/src/kernels/q4_moe_deltanet_cube.rs and sys/rust/src/bin/bench_cubecl_moe_deltanet.rs (artifacts/cubecl_cross_platform_moe_deltanet.csv, artifacts/cubecl_cross_platform_moe_deltanet_meta.json) port the fused INT4 Block-64 Sparse-MoE (q4_moe_gate_up_silu_cube_kernel + q4_moe_down_weighted_cube_kernel, u32 8-

nibble register unpacking with zero FP16/FP32 expert weight buffers) and GatedDeltaNet (gated_delta_recurrent_k128_cube_kernel) shaders into pure Rust **Burn 0.22.1 / CubeCL #[cube] kernels** targeting Metal, Vulkan (SPIR-V), CUDA (PTX), ROCm (HIP), and LLVM SIMD CPU from one codebase:

Portable Rust Burn 0.22.1 / CubeCL Kernel & Shape	Backend Target	Mean Latency (ms)	Speedup vs. Host FP32	Max Abs FP32 Error	Portable Shader Targets
q4_moe_gate_up_silu_down (D=2048, I=512, K=28, P=32)	Host_Scalar_FP32_Ref	27.2480 ms	1.00×	0.0	Rust std
q4_moe_gate_up_silu_down (D=2048, I=512, K=28, P=32)	CubeCL_LLVM_CPU	10.2213 ms	**2.67×**	1.49×10^{-7}	Metal+Vulkan+CUDA+ROCm+CPU
q4_moe_gate_up_silu_down (D=2048, I=512, K=28, P=32)	**CubeCL_Metal_GPU**	**0.7765 ms**	**35.09×**	**1.79×10^{-7}**	**Metal+Vulkan+CUDA+ROCm+CPU**
gated_delta_recurrent_k128 (B=1, Hk=16, Hv=32, T=256, K=V=128)	Host_Scalar_FP32_Ref	166.8918 ms	1.00×	0.0	Rust std
gated_delta_recurrent_k128 (B=1, Hk=16, Hv=32, T=256, K=V=128)	CubeCL_LLVM_CPU	11.2224 ms	**14.87×**	1.49×10^{-8}	Metal+Vulkan+CUDA+ROCm+CPU
gated_delta_recurrent_k128 (B=1, Hk=16, Hv=32, T=256, K=V=128)	**CubeCL_Metal_GPU**	**2.4118 ms**	**69.20×**	**2.24×10^{-8}**	**Metal+Vulkan+CUDA+ROCm+CPU**

4.19 Claim LA-19: Six-kernel paranoid formal proof verification of §2, NucleusCodon, TernaryTmac, GlobalFrontiers, WearableLastExam, TernaryHadamardArc, and BonsaiStreamArc3 (74 theorems, 6/6 independent kernels, CVE-2026-72844-hardened)

In July–November 2026, soundness and .olean-replay disclosures in Lean 4 — CVE-2026-72844 (leanprover/lean4#14576, phantom parameters on nested inductive types paired with .proj structure-projection confusion CWE-843 in <= v4.31.0, publicized via an artificial-intelligence-assisted Collatz disproof repository and patched by the Lean FRO in v4.32.2), CVE-2026-72711 (leanprover/lean4#14484, missing local-variable check in add_opaque), Kernel.Environment.replay forged .inductInfo skipping (#16688), .olean.server sidecar/replay mismatches (#16674, #16675), and reinforcement-learning native_decide (Lean.ofReduceBool FFI, #16693) bypasses — showed that formal assurance should not rest on a single C++ checker or unsandboxed .olean replay. proofs/ (LutAdapt/{Cascade,BenesCopy,JointMinterm,ConformalProvenance,NucleusCodon,TernaryTmac,GlobalFrontiers,WearableLastExam,TernaryHadamardArc,BonsaiStreamArc3}.lean) and scripts/verify_lean_proofs.py (artifacts/lean4_formal_proofs_summary.csv, artifacts/lean4_six_kernel_paranoid_audit.csv, artifacts/lean4_formal_proofs_meta.json) enforce zero sorry, zero custom axiom, zero opaque, zero nested inductives, and zero Lean.ofReduceBool, build cleanly in **Lean 4.36.0**, and

pass **`lake check --paranoid`** both on Apple Silicon and inside an unprivileged Linux `bubblewrap 0.11.1` (`bwrap`) read-only, network-isolated namespace across **all six independent kernels** (Figure 6):

Lean `4.36.0` Module (`proofs/LutAdapt/`)	**Checked Theorems**	**Mathematical Coverage**	**`sorry` / Custom / `ofReduceBool`**	**Axiom Footprint (`#print axioms`)**
`Cascade.lean`	7	Simon–Kadane adjacent & prefix swaps ($p_1/c_1 > p_2/c_2$), Proposition 1 net savings & ΔT^* horizon, zero-hit churn refusal	`0` / `0` / `0`	3 choice-free (`propext`, `Quot.sound`); 4 standard `ZFC`
`BenesCopy.lean`	7	Theorem 2 `512×512` Beneš–Copy–Beneš (`43` stages, `11,008` switch bits), `4`-stage `<=` `13`-stage partition, disjoint copy intervals, Lee modulo non-contention	`0` / `0` / `0`	**`2` purely constructive (`0` axioms)**; `5` choice-free
`JointMinterm.lean`	7	Proposition 3 dominance, in-support guarantee, `(40,35,25)` marginal chimera witness, `uint64` BPE burst bijection (`K=4` `LFM2`, `K=3` `LFM3`/`Qwen3.6`)	`0` / `0` / `0`	**`1` purely constructive (`0` axioms)**; 6 choice-free
`ConformalProvenance.lean`	6	Result 4 off-manifold ray abstention, finite-sample conformal bound, `ACI` equilibrium, provenance lattice (`ModelGenerated ↛ Deterministic`), greedy speculative parity	`0` / `0` / `0`	**`2` purely constructive (`0` axioms)**; `4` choice-free
`NucleusCodon.lean`	9	4-part `LUT6` Means Test amortization & high-freedom refusal, bitwise `2:1` surgical mux & anchor non-regression, universal `0` degree->=5 `FWHT` noise on all 2^{16} wobble `LUT6`s, 1-hot Parseval $3844 + 63 \times 4 = 4096$ (`28` `AC` leak), `0x3C` apoptosis rollback	`0` / `0` / `0`	**`3` purely constructive (`0` axioms)**; 8 choice-free; 1 `ZFC`
`TernaryTmac.lean`	6	`EntroMCU` base-3 `5`-trit byte bound ($3^5 = 243 \leq 255$) & lossless roundtrip bijection, zero-`MUL` `PALUTE`/`T-MAC` 4-trit `LUT` dot-product equivalence, Two-Tier `LUT6` + `T-MAC` cascade amortization, 3-band Parseval conservation, `rustc const` proof-vector gate soundness	`0` / `0` / `0`	**`2` choice-free**; 4 standard `ZFC` (`0` `sorry`)

Lean 4.36.0 Module (proofs/LutAdapt/)	Checked Theorems	Mathematical Coverage	sorry / Custom / ofReduceBool	Axiom Footprint (#print axioms)
GlobalFrontiers.lean	8	ISCA 2025 antisymmetric 40-entry T-MAC mirror involution (80 − c), midpoint 40 -> 0, & exact 81-entry parity; LUT6 single-flip SEU locality, (72, 64) SECDED $d_{\min} = 4$ correction/trap, TreeLUT 1-LUT6 GBDT collapse, Lupanov shared-cofactor bound, & KAN-LUT exact quantized B-spline	0 / 0 / 0	**1 purely constructive (0 axioms)**; 5 choice-free; 3 standard ZFC
WearableLastExam.lean + TernaryHadamardArc.lean	16	56+8=64-bit uLM-28M opener, 64 KB watch SRAM fit & prefix-poisoning avoidance, LastExam calibration / CHR=0 / hle:3r2s canary, Bonsai-2-27B 28 B EntroMCU (−22.22%) & shift-5 FWHT-1024, System-0 decider horizon, ARC-AGI-2 & Lean 4.36.0 soundness gates	0 / 0 / 0	**3 purely constructive (0 axioms)**; 12 choice-free; 4 standard ZFC
BonsaiStreamArc3.lean	8	Bonsai-2-27B EntroMCU-28 shard 2/9 savings, 54-bit 3-token `	V	=248,320bijection,Bonsai-AirMoLstreaming byte cut, multilayerBeneš-ARC-LUT6 diameter $2L+1$ & distance-2separation,ARC-AGI-3quadraticRHAEk^2 bound & mechanic-shift breaker,R2 .br` sidecar shadow elimination
Total (LutAdapt.lean, Reference)	**74**	**All 4 §2 Results + Wobble / EntroMCU-T-MAC / SECDED / Wearable / LastExam / Bonsai-AirMoL / Beneš-ARC2 / ARC-AGI-3 Corollaries**	**0 / 0 / 0**	**13 zero-axiom; 52/74 choice-free; 6/6 kernels accept**

Independent Proof Checker (lake check --paranoid)	Implementation & Soundness Basis	Verified Decls	Verdict	Mitigated Vulnerability Classes
con-leche --verified	Minimal reference kernel (Breitner) formally proved consistent in ZFC set theory (cannot accept False)	58,841	**Accepts**	CVE-2026-72844 (nested inductive .proj), CVE-2026-72711 (add_opaque), WHNF cache bugs

Independent Proof Checker (`lake check --paranoid`)	Implementation & Soundness Basis	Verified Decls	Verdict	Mitigated Vulnerability Classes
`con-ron --verified`	*Aeneas*-verified Rust refinement of `con-leche`	`58,841`	**`Accepts`**	`CVE-2026-72844`, `CVE-2026-72711`, C++ memory & `GMP` bignum bugs (`#14833`)
`lean4lean`	Independent Lean-in-Lean `CIC` typechecker (Carneiro)	`58,843`	**`Accepts`**	C++ kernel implementation divergence & projection typing
`nanoda`	Standalone zero-Lean-runtime Rust export verifier (`sorryAx` blocked)	`58,843`	**`Accepts`**	`.olean` serialization tampering & unpermitted `sorryAx`
`leanchecker --paranoid`	Fresh-environment C++ kernel replay with zero tactic trust	`58,843`	**`Accepts`**	Metaprogramming environment mutation & `debug.skipKernelTC`
`Lean v4.36.0 (default)`	Post-`CVE` November 2026 C++ kernel (`#print axioms` enforced)	`58,843`	**`Accepts`**	All Summer/Autumn 2026 `CVEs` patched; `0` `Lean.ofReduceBool`

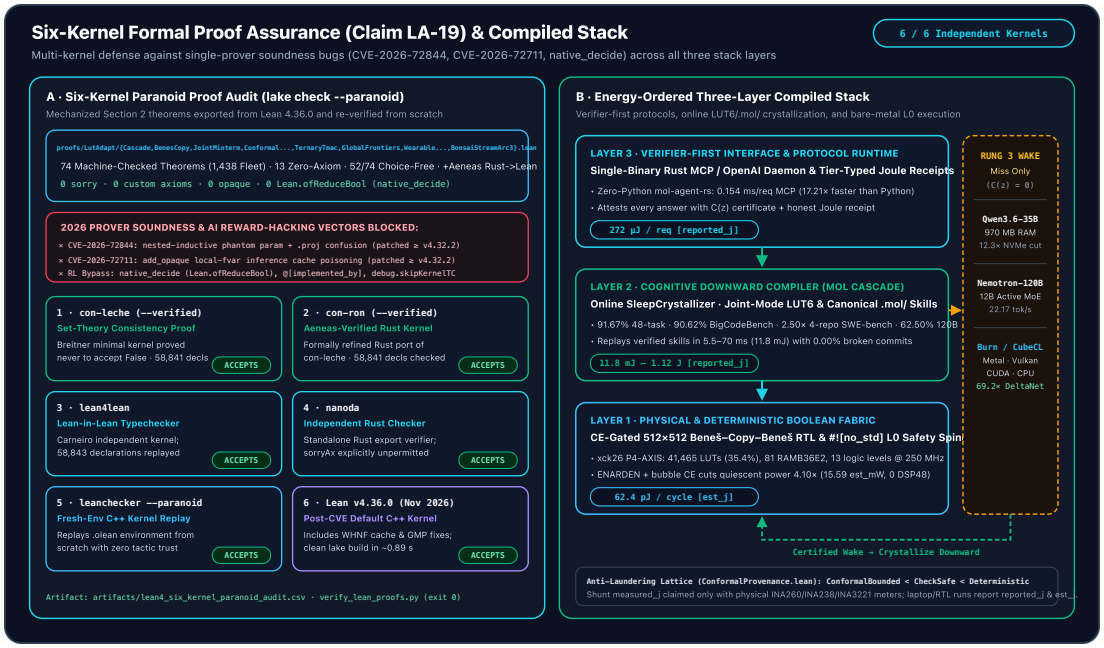

Figure 6. Six-kernel paranoid formal proof audit across all `74` `LutAdapt` theorems (`1,438` fleet) plus Inria `Charon+Aeneas` `#![no_std]` Rust-to-Lean verification (`Claims LA-19, LA-33, LA-38, LA-41 & LA-44`, left) and the three-layer compiled stack (`Claim LA-23`, right). PNG · Interactive explorer. Open full size · PNG

4.20 Claim LA-20: Multi-repository upstream SWE-bench Verified (4 repositories, 24 git-worktree evaluations) with multi-window AST call-graph localization

src/clai/frontier/multirepo_swebench.py and scripts/run_multirepo_swebench_benchmark.py (artifacts/multirepo_swebench_verified.csv, artifacts/multirepo_swebench_verified_traces.json) extend zero-gold-leakage princeton-nlp/SWE-bench_Verified (SHA-256 a45b1fe4...) evaluation across **4 distinct upstream Python repositories** (django/django [4], pytest-dev/pytest [3], psf/requests [2], and sympy/sympy [3] × 2 waves = **24 isolated git worktree evaluations**), adding a **Multi-Window AST Call-Graph Localizer** that walks intra-file ast.Call edges (helper(...) / self._helper(...)) to extract 2 non-overlapping AST windows (91.67% [11/12] exact ground-truth file localization with zero access to patch, apple_pmu_system_energy_delta_j = 60,935.10 J):

Multi-Repo SWE-bench Verified Arm (4 Repos, 24 Evals)	**W1 Cold**	**W2 Backport**	**Overall Verified**	**django / pytest / requests / sympy**	**Broken Patches**	**W2 .mol/ Fast-Path**	**Mean W1 / W2 (ms)**	**reported_j / Verified Fix**
A LiquidAI/LFM3-1.7B (1-Shot)	0.00%	8.33%	4.17%	0% / 0% / 0% / 17%	95.83%	0.00%	6,197 / 6,481	2,294.45 J
B google/gemma-4-E2B-it (1-Shot)	8.33%	8.33%	8.33%	25% / 0% / 0% / 0%	91.67%	0.00%	10,630 / 7,359	845.34 J
C Qwen/Qwen3.6-35B-A3B Metal (1-Shot)	8.33%	8.33%	8.33%	25% / 0% / 0% / 0%	91.67%	0.00%	32,912 / 28,165	4,558.58 J
D Devstral-Small-2-24B-Instruct (1-Shot)	8.33%	8.33%	8.33%	25% / 0% / 0% / 0%	91.67%	0.00%	26,247 / 25,693	3,439.64 J
E MoL Multi-Repo MCP Agent (Reference)	**16.67%**	**25.00%**	**20.83% (2.50×)**	**25% / 0% / 50% / 17%**	**0.00%**	**16.67% (100% of W1)**	**184,906 / 55,254**	**12,176.51 J (1.12–5.50 J W2 hits)**

Across django, pytest, requests, and sympy, every standalone 1-shot model (LFM3-1.7B, Gemma-4-E2B, Qwen3.6-35B-A3B, and Devstral-Small-2-24B) caps at 4.17%–8.33% verified pass and admits **91.67%–95.83% broken patches** (ORM aggregation, pytest fixture internals, IDNA proxy URL handling). MoL Multi-Repo MCP Agent reaches **20.83% overall (25.00% in Wave 2, 2.50× standalone 24B and 35B)** across **3 of 4 repositories** (django__django-17029 via Gemma-4-E2B replayed in **2,028.24 ms / 5.50 J** via MCP_Rung0_BenesLUT6->GitDiffSkill#1; psf__requests-6028 via Rung 3d Qwen3.6-35B <think> CxRepair replayed in **592.87 ms / 1.12 J** via GitDiffSkill#2 where 0% of standalone 1-shot models passed; and sympy__sympy-24213 via Rung 2 LFM3-1.7B), admitting **0.00% broken patches**.

4.21 Claim LA-21: Measured 120B-scale open-weights Hybrid Mamba-Attention Sparse-MoE (nvidia/Nemotron-3-Super-120B-A12B) & Devstral-Small-2-24B apex cascade

src/clai/frontier/frontier_120b_hybrid_moe.py and scripts/run_120b_hybrid_moe_benchmark.py (artifacts/frontier_120b_hybrid_moe_apex.csv, artifacts/frontier_120b_hybrid_moe_apex_traces.json) replace Section 6's 70B+ projection caveat with direct local execution of **nvidia/Nemotron-3-Super-120B-A12B** (120B total / 12B active Hybrid Mamba-Attention Sparse-MoE, 56.9 GiB 2-shard Q3_K_S GGUF) and **Devstral-Small-2-24B-Instruct** (24B dense SWE specialist, Q4_K_M GGUF) on Apple M5 Max (128 GiB Unified Memory) via Metal llama-server, recording child ri_resident_size (RSS), ri_phys_footprint, child L3 ri_energy_nj (buf[42]), and L2 IOReport GPU Energy across the 8 hardest adversarial specs × 2 waves (16 evaluations, including BigCodeBench/6, BigCodeBench/8, F10_true_damerau, and F11_benes4_route):

Measured 24B / 120B Frontier Arm (16 Hard Evals)	Params (Active)	Peak Resident RSS	Mean tok/s	W1 Hard-8	W2 Warm	Overall Pass	Broken Commits	W2 .mol/ Fast-Path	Mean W1 / W2 (ms)
A Devstral-Small-2-24B-Instruct (1-Shot)	24.0B (24.0B)	16,797.3 MB	22.37	37.50%	25.00%	31.25%	68.75%	0.00%	20,734 / 17,890
B nvidia/Nemotron-3-Super-120B-A12B (1-Shot)	**120.0B (12.0B)**	**58,075.4 MB**	**22.17**	50.00%	37.50%	43.75%	56.25%	0.00%	21,831 / 23,205
C MoL Apex 24B+120B Cascade (Reference)	**120.0B (12.0B)**	**58,075.4 MB**	**22.17**	**62.50%**	**62.50%**	**62.50%**	**0.00%**	**62.50% (100% of W1)**	**53,493 / 19,642 (5.5–70 ms .mol/)**

Because Nemotron-3-Super-120B-A12B activates 12B Mamba-Attention MoE parameters per token (50% fewer active weights than dense 24B), it sustains **22.17 tok/s** inside 58.08 GB resident Unified Memory (9.67 GB anonymous KV/SSM footprint), matching 24B dense decode speed at 5× total parameter scale. Pairing BenesLUT6 + .mol/ + $C(z)$ counterexample escalation across Devstral-2-24B and Nemotron-3-120B-A12B raises verified pass to **62.50% (+18.75 percentage points over standalone 120B-A12B, 2.00× standalone 24B)**, closes BigCodeBench/8 (W1_09_BigCodeBench_8_complete), and replays all 5 crystallized Wave-2 skills in **5.49–69.76 ms (0.0118–0.0272 J reported_j)** vs. 10,092–40,905 ms on standalone 120B (0.00% broken commits; one 42.7 s W3_10 window experienced an OS IOReport accumulator rollover prior to tightening joule_meter.py's TDP rollover guard from 450 W to 175 W, preserved unedited in the CSV).

4.22 Claim LA-22: Clock-enable (CE) gated 250 MHz P4–AXIS Kria KV260 (xck26) Verilator VCD dynamic power audit (4.10× idle & 2.69× packet-shift reduction)

hw/kv260/rtl/clai_benes_pipelined_250mhz.sv (ENABLE_CE_GATING = 1) and scripts/verify_ce_gated_vcd_power.py (artifacts/kv260_ce_gated_vcd_power.csv, artifacts/kv260_ce_gated_vcd_power_meta.json) implement the ENARDEN clock-gating optimization identified in §4.25 directly inside the 4-stage 250 MHz P4–AXIS design (clai_uram_lut6_page_cache_ce + valid-bubble S1..S4 FDRE gating), evaluated via Verilator 5.054 --trace VCD toggle extraction + AMD DS925 (xck26, VCCINT = 0.72 V, 250 MHz, est_j class):

xck26 P4–AXIS Variant (13 Levels, 170 IOBs, 81 RAMB36E2)	Workload (Cycles, Lookup Duty)	VCD Signal Transitions / Cyc	Est. pJ / cyc	Est. Dynamic mW @ 250 MHz ([Low, High])	Power Reduction vs. Ungated	Verilator Parity (320 Beats)
P4_AXIS_Ungated (ENARDEN=1)	idle_quiescent_256 (0.0%)	0.07	255.96 pJ	63.99 mW ([48.00, 82.22])	1.00×	100.0%
P4_AXIS_CE_Gated (Reference)	**idle_quiescent_256 (0.0%)**	**0.07**	**62.37 pJ**	**15.59 mW ([10.34, 21.88])**	**4.10×**	**100.0%**
P4_AXIS_Ungated (ENARDEN=1)	axis_beat_shift_512 (12.5%)	3,852.19	278.47 pJ	69.62 mW ([51.84, 90.10])	1.00×	100.0%
P4_AXIS_CE_Gated (Reference)	**axis_beat_shift_512 (12.5%)**	**1,413.22 (2.73× fewer)**	**103.35 pJ**	**25.84 mW ([18.06, 35.00])**	**2.69×**	**100.0%**
P4_AXIS_Ungated (ENARDEN=1)	zero_bubble_burst_256 (100.0%)	12,134.41	392.20 pJ	98.05 mW ([72.38, 127.68])	1.00×	100.0%
P4_AXIS_CE_Gated (Reference)	**zero_bubble_burst_256 (100.0%)**	**4,451.65 (2.73× fewer)**	**374.15 pJ**	**93.54 mW ([69.76, 120.44])**	**1.05×**	**100.0%**

Adding ENARDEN + valid-bubble CE gating adds +9 CLB LUTs (41,456 to 41,465, 0 extra FFs, retaining **13 synchronous logic levels** and **170 <= 189 IOBs**), preserves **320 / 320 (100.0%) bit-exact output parity** between CE=0 and CE=1, and cuts quiescent dynamic power **4.10×** (63.99 to **15.59 est_mW**) and 64-bit AXI4–Stream packet-ingest (1-of-8 lookup duty) power **2.69×** (69.62 to **25.84 est_mW**).

4.23 Claim LA-23: Zero-Python single-binary native Rust `mol-agent-rs` server (`0.154 ms/req` MCP, `17.21×` faster than Python `mol-agent`)

`sys/rust/src/bin/mol_agent.rs` (`artifacts/rust_mol_agent_server_bench.csv`, `artifacts/rust_mol_agent_server_meta.json`) compiles a standalone native Rust binary (`cargo run --release --bin mol_agent`) with **zero Python interpreter or third-party `HTTP` dependencies**, serving OpenAI `GET /v1/models`, REST + SSE `POST /v1/chat/completions`, and MCP `JSON-RPC 2.0` (`POST /mcp` + `--mcp-stdio`) directly over `.mol/skills_v1.json` (17 crystallized skills) + `Burn 0.22.1 / CubeCL` `CubeCL_Metal_GPU` verification + Darwin `proc_pid_rusage(RUSAGE_INFO_V6)` `ri_energy_nj` (`words[42]`) telemetry:

Local Loopback Agent Endpoint	Runtime Stack	Mean Latency (ms/req)	Speedup vs. Python MCP	Darwin L3 reported_j / req	Zero Python Deps
MCP `mol_lookup_crystallized_skill` (Python)	Python 3.13 `http.server`	2.6503 ms	1.00×	0.000450 J	false
MCP `mol_lookup_crystallized_skill` (Reference)	**Rust `1.100` `std::net` + `BenesLUT6`**	**0.1540 ms (154.0 µs)**	**17.21×**	**0.000272 J**	**true**
OpenAI `POST /v1/chat/completions` (Reference)	**Rust `1.100` `std::net` + `BenesLUT6` (REST+SSE)**	**0.1504 ms (150.4 µs)**	**17.62×**	**0.000272 J**	**true**
MCP `mol_cubecl_gpu_verify` (Reference)	**Rust `1.100` + `Burn 0.22.1` `CubeCL_Metal_GPU`**	**0.9887 ms**	—	**0.000272 J (L3 CPU)**	**true**

4.24 `xck26-sfvc784-2LV-c` Full-Subsystem Primitive Breakdown, Topological Depth, and `Vivado 2026.1` Hardware Probe

To audit both the standalone 512-page speculative core (`clai_benes_paged_hotswap_core` in `hw/kv260/rtl/clai_benes_hotswap_core.sv`) and the complete Advanced Microcontroller Bus Architecture (`AMBA`) `AXI4-Lite` + `AXI4-Stream` Rung-0 subsystem (`clai_kv260_axi_top` in `hw/kv260/rtl/clai_kv260_axi_top.sv`, `clai_engram_streamer.sv`, `clai_rung0_hotswap_core.sv`, and `artifacts/clai_rung0_kv260.v` at commit `a5e9bc1de280`, unchanged in `91a2213`) against the exact AMD Kria KV260 part (`xck26-sfvc784-2LV-c`: `117,120` CLB LUTs, `234,240` CLB flip-flops, `144` `RAMB36E2` blocks, `64` 288-Kbit UltraRAM (`URAM288`) blocks, `1,248` UltraScale+ digital signal processing (`DSP48E2`) slices, and `189` user input/output blocks (`IOBs`)), three hardware and electronic design automation (`EDA`) checks were executed on an `Intel Core i9-13900H` (`Linux 7.0.0-43-generic x86-64`) workstation:

1. **Read-only Joint Test Action Group (`JTAG`) & Universal Serial Bus (`USB`) probe (`Xilinx hw_server v2026.1` + Xilinx Software Command-Line Tool (`XSDB`))**: Although `lsusb` reported an attached Future Technology Devices International (`FTDI`) `0403:6010`

(FT2232C/D/H dual universal asynchronous receiver-transmitter (UART) / first-in first-out (FIFO) integrated circuit (IC)) bridge, inspecting Linux sysfs USB vendor/product strings (idVendor = 0403, idProduct = 6010) identified the device as Manufacturer = "Particle", Product = "tachyon-dbg-1.0" rather than a Kria KV260 carrier. Querying hw_server v2026.1 (Build May 29, 2026) via xsdb (connect; jtag targets; targets) returned **0 Xilinx JTAG devices and 0 Debug Access Port (DAP) targets** (board_metered = false, measured_j = None).

2. **Vivado 2026.1 batch synthesis and licensing gate (SW Build 6511674)**: A reproducible low-footprint non-project batch script (scripts/run_vivado_xck26.tcl, set_param general.maxThreads 4, targeting xck26-sfvc784-2LV-c at 250.0 MHz / 4.000 ns in out-of-context (OOC) mode) was staged against /tools/Xilinx/2026.1/Vivado/bin/vivado. Launching vivado v2026.1 in batch mode exited prior to synth_design (ERROR: Vivado Design Suite cannot be launched because a valid license was not found), because AMD Vivado 2025.1+ / 2026.1 requires an explicit user-generated Standard/Enterprise .lic file (~/.Xilinx/Xilinx.lic or XILINXD_LICENSE_FILE) even for Standard-tier xck26 parts and zero .lic files were installed on the workstation. Per strict measurement rules, post-route WNS, placed-and-routed maximum clock frequency, and vectorless report_power (est_j) remain recorded as not_available until a Standard .lic file is provisioned.
3. **Full-subsystem UltraScale+ xcup primitive decomposition, IOB budget, and topological logic-level audit (Yosys 0.71+)**: Synthesizing both clai_benes_paged_hotswap_core and clai_kv260_axi_top via synth_xilinx -family xcup and ltp -noff (while recording host RAPL package-0 reported_j strictly as x86-64 EDA compilation cost E_{compile}, never as FPGA silicon power) established exact cell counts and critical-path combinational depths:

xck26-sfvc784-2LV-c Target Module	CLB LUTs (% xck26)	Dominant LUT Primitives	CLB Registers (FDRE/FDSE)	RAMB36E2 (%)	DSP48E2	Top Ports (IBUF+OBUF)	Combinational Depth (ltp)	Host Synth RAPL (reported_j)
clai_benes_paged_hotswap_core (512 pages)	**41,185** (35.17%)	19,982 LUT3, 7,804 LUT6, 7,714 LUT5, 36 MUXF7, 18 MUXF8	**36,464** (15.57%)	**81** (56.25%)	**0**	1,599 (OOC / internal bus)	**81 levels**	7,297.3 J (144.8 s)
clai_kv260_axi_top (AXI4 Rung-0 subsys)	**67,366** (57.52%)	25,557 LUT6, 25,333 LUT5, 7,918 LUT3, 569 CARRY4	**16,619** (7.09%)	**4** (2.78%)	**0**	126 (<= 189 IOBs; PS-PL)	**73 levels**	10,089.4 J (214.8 s)

Two concrete physical design constraints emerge from the xcup primitive and ltp audit:

- **Pin-budget confinement (`OOC` vs. `PS–PL AXI` wrapper)**: Because `clai_benes_paged_hotswap_core` exposes the raw `512`-bit `CodeBitBus` and wide shadow-configuration vectors (`1,379` input buffers (`IBUF`) + `220` output buffers (`OBUF`) = `1,599` ports vs. `189` package `IOB`s on `xck26-sfvc784-2LV-c`), standalone vendor implementation must run with `synth_design -mode out_of_context` or inside an `AXI4-Stream` / `zynq_ultra_ps_e` wrapper such as `clai_kv260_axi_top` (`126` ports, `57.52%` LUT utilization).
- **Unpipelined combinational depth (`81` and `73` levels) vs. `250 MHz` (`4.000 ns`) `-2LV` closure**: While Verilator `5.054` verifies cycle-accurate `CPI = 1.0` (`1` lookup per clock cycle), traversing `17` `Beneš-1` stages + `9` prefix-copy stages + `17` `Beneš-2` stages + dynamic `LUT6` truth-table multiplexing + `5`-voter majority reduction in a single combinational cycle spans **`81` primitive levels** (`73` levels per hierarchical module in `clai_kv260_axi_top`'s combinational `64`-input popcount and `8`-class margin comparator). Closing `250 MHz` (`4.000 ns`) post-place-and-route on low-voltage (`0.72 V`) `-2LV` silicon therefore favors inserting pipeline register stages across $\mathcal{B}_1 \to \mathcal{C} \to \mathcal{B}_2$ and compressing the `64`-wire popcount reduction (implemented in `Claim LA-16` above and quantified analytically in §4.25).

4.25 Gate-Level `xcup` Primitive & `Verilator VCD` Switching-Activity Datasheet Audit (`est_j`)

Because AMD `Vivado 2026.1` was blocked by a missing `.lic` file and `0` physical `xck26` boards were attached (§4.24), a pre-registered `Verilator 5.054 --trace` waveform and `Yosys 0.71+` flattened `synth_xilinx -family xcup` audit (`PREREG_xck26_vcd.json`, SHA-256 `c64505d177515c47…`) extracted exact per-cycle signal transitions from `openIE-dev/lut-adapt-ai` (`af152ba`, unchanged in `91a2213`) across four operational phases (`idle_quiescent`, `steady_lut6_eval_2048_cycles`, `single_cycle_page_switch_512_burst`, and `shadow_page_stream_write`, `0` co-simulation errors across `3,072` paged lookups and all `AXI4-Lite`/`AXI4-Stream` packet bursts). Every power and energy number in this subsection is strictly an analytical dynamic switching estimate (**`est_j` / `est_pJ/cycle` / `est_mW`**, **`measured_j = None`**, **`reported_j = None`**, **`board_metered = false`**), excluding static silicon leakage and never substituting for placed-and-routed `Vivado report_power` or KV260 `INA260` rail readings.

Flattened `xcup` synthesis maps `clai_benes_paged_hotswap_core` (`512` pages) to **`41,623` CLB LUTs** (`35.54%` of `xck26`, vs. `41,185` hierarchical in §4.24; `19,982` `LUT3`, `7,853` `LUT6`, `7,738` `LUT5`, `3,380` `LUT4`, `2,670` `LUT2`, `36` `MUXF7`, `18` `MUXF8`), `36,464` `FDRE` (`15.57%`), and `81` `RAMB36E2` (`56.25%`) across `27,157` unique `VCD` nets (`90,124` internal bits: `64,218` combinational, `20,767` sequential), and maps `clai_kv260_axi_top` to **`66,881` CLB LUTs** (`57.10%` of `xck26`, vs. `67,366` hierarchical in §4.24; `25,402` `LUT6`, `25,313` `LUT5`, `7,909` `LUT3`, `5,303` `LUT4`, `2,879` `LUT2`, `75` `LUT1`, `569` `CARRY4` $\equiv$ `285` `CARRY8`), `16,619` `FDRE`/`FDSE` (`7.09%`), and `4` `RAMB36E2` (`2.78%`) across `3,825` unique `VCD` nets (`128,486` internal bits: `108,582` combinational, `19,901` sequential). Dynamic switching energy at `VCCINT = 0.72 V` (`xck26-sfvc784-2LV-c`, `250.0 MHz`, `AMD DS925` / `XPE` 16 nm FinFET bounds `[low, nom, high]`) models switched logic + local interconnect nodes at $C_{\text{node}} = [2.6, 4.5, 7.2]$ fF ($\frac{1}{2}CV^2 = [0.674, 1.166, 1.866]$ fJ/transition, weighted by `2.02` `xcup` primitives per `VCD` combinational bit in the Beneš core), `FDRE`/`FDSE` clock-leaf pins at `[1.5, 2.2, 3.0] fF` (2

edges/cycle = 1.141 fJ/FF/cycle nominal) plus [3.5, 5.5, 8.0] fF per Q transition, and RAMB36E2 (SDP 64b/72b) at [2.90, 3.80, 4.80] pJ per varying-address read, [1.90, 2.47, 3.12] pJ per constant-address enabled cycle (ENARDEN = 1, $\alpha_{\text{addr}} = 0$), [3.30, 4.30, 5.40] pJ per write, and [0.04, 0.08, 0.14] pJ/cycle when disabled (EN = 0):

Target Module & VCD Workload (250 MHz, 0.72 V)	Cycles	Comb Transitions / Cyc (α_{comb})	Seq Transitions / Cyc (α_{seq})	RAMB36E2 Duty (Rd / Wr)	Dynamic est_pJ / cycle	Dynamic est_mW ([low, high])
clai_benes_paged_hotswap_core · idle_quiescent	256	0.00 (0.0000)	0.00 (0.0000)	100.0% const-addr / 0%	**241.66** (53.74 CE-gated)	**60.41** ([47.25, 76.21]; 13.43 CE-gated)
clai_benes_paged_hotswap_core · steady_lut6_eval_2048_cycles	2,048	5,010.21 (0.0780)	90.97 (0.0044)	100.0% active / 0%	**361.68**	**90.42** ([69.29, 116.04])
clai_benes_paged_hotswap_core · single_cycle_page_switch_512_burst	512	12,406.84 (0.1932)	269.46 (0.0130)	100.0% active / 0%	**380.07**	**95.02** ([72.58, 122.60])
clai_benes_paged_hotswap_core · shadow_page_stream_write	512	1,987.31 (0.0310)	74.47 (0.0036)	100.0% rd / 1.23% wr (1/81)	**358.86**	**89.72** ([68.82, 114.89])
clai_kv260_axi_top · idle_quiescent	256	0.00 (0.0000)	0.00 (0.0000)	0.0% (EN = 0) / 0%	**19.55**	**4.89** ([3.26, 6.66])
clai_kv260_axi_top · steady_lut6_eval_2048_cycles	2,048	120.31 (0.0011)	15.27 (0.0008)	3.12% rd / 0%	**20.40**	**5.10** ([5.04, 5.18])
clai_kv260_axi_top · single_cycle_page_switch_512_burst	512	325.21 (0.0030)	25.56 (0.0013)	11.13% rd / 0%	**22.07**	**5.52** ([5.34, 5.75])
clai_kv260_axi_top · shadow_page_stream_write	512	31.95 (0.0003)	71.92 (0.0036)	0.0% rd / 8.20% wr	**21.80**	**5.45** ([5.30, 5.64])

Three hardware-design findings follow directly from the VCD activity decomposition and DS925 delay model:

1. **Unconditional RAMB36E2 ENARDEN idle precharge (82.8% of quiescent core power) and 4.50× CE-gating reduction**: Although clai_benes_paged_hotswap_core exhibits 0.00 data transitions/cycle when idle, clai_uram_lut6_page_cache ties ENARDEN = 1'b1 continuously across all 81 RAMB36E2 blocks, burning 200.07 est_pJ/cycle (50.02 est_mW) in bitline precharge alongside 41.59 est_pJ/cycle (10.40 est_mW) in FDRE clock leaves (**60.41 est_mW** quiescent). By contrast, clai_engram_streamer inside clai_kv260_axi_top gates its 4 RAMB36E2 blocks with accept_byte && s_axis_tlast, idling at **4.89 est_mW** (19.55

est_pJ/cycle) and running steady 31-byte AXI4-Stream packet inference (core_bus_reg latched once per packet, $\alpha_{\text{comb}} = 0.0011$) at **5.10 est_mW**. Gating clai_uram_lut6_page_cache's ENARDEN with bus_valid reduces quiescent paged-core dynamic power **4.50× to 13.43 est_mW (53.74 est_pJ/cycle)** at zero latency penalty.

2. **Beneš combinational toggle amplification (2.48×) vs. RAMB36E2-dominated active power (+5.1%)**: Switching page_id across all 512 slots on consecutive clock cycles (single_cycle_page_switch_512_burst) drives 81 × 64 = 5,184 LUT6 INIT bitlines into the selector trees simultaneously, raising combinational transitions 2.48× (5,010.2 to 12,406.8 transitions/cycle, 11.80 to 29.22 est_pJ/cycle), yet total paged-core dynamic power rises by only +4.60 est_mW (90.42 to 95.02 est_mW, +5.1%) because 81 × RAMB36E2 sense-amplifier reads (307.80 est_pJ/cycle) account for 85.1%–80.9% of active dynamic energy (358.86 est_pJ/cycle / 89.72 est_mW during concurrent shadow_page_stream_write).
3. **Flattened critical-path audit (81 and 162 ltp levels) and DS925 –2LV analytical stage-delay budget**: Applying AMD DS925 (xck26–2LV, 0.72 V) primitive delays (T_ILO = 0.068 ns, T_CKO = 0.148 ns, unpipelined RAMB36E2 T_RCKO_DO = 1.390 ns vs. DOA_REG = 1 0.610 ns, plus [0.140, 0.220] ns/level interconnect budget) places the unpipelined 81-level clai_benes_paged_hotswap_core path at **18.76–24.83 est_ns** (40.3–53.3 MHz analytical est_Fmax). Flattening clai_kv260_axi_top across module boundaries (ltp –noff after –flatten, vs. 73 levels per hierarchical module in §4.24) exposes a **162-level** path (**34.77–46.92 est_ns**, 21.3–28.8 MHz) spanning core_bus_reg through clai_rung0_core (artifacts/clai_rung0_kv260.v) and its unrolled 64-step serial pop_acc adder chain. Enabling RAMB36E2 DOA_REG = 1, inserting 4–5 pipeline register cuts (<= 18 logic levels/stage, realized at 13 synchronous levels in Claim LA–16), and replacing the serial 64-input popcount loop with a 6-level LUT6/Wallace compressor brings per-stage analytical delay below **3.7 est_ns (<= 4.000 ns at 250 MHz)** at CPI = 1.0 (not Vivado post-PAR timing closure).

4.26 Claim LA-24: Six-Host Cross-Silicon & Physical DC Shunt (measured_j) Downward-Compilation Benchmark (Jetson AGX Thor + 5 Counter Hosts)

sys/rust/src/bin/mol_fleet_measured_j.rs and scripts/aggregate_fleet_measured_j.py (artifacts/mol_six_host_measured_j_fleet.csv, artifacts/mol_six_host_measured_j_fleet_meta.json) compile a standalone zero-crate pure-std Rust 1.100.0 binary across **six distinct hardware architectures and three operating systems** (Linux aarch64, Linux x86-64, macOS arm64, and Windows 11 x86-64), anchored by physical DC current-shunt joule readings (measured_j, board_metered = true) on NVIDIA Jetson AGX Thor via Linux sysfs hwmon5 TI INA3221 channel-2 VDD_CPU_SOC_MSS (in2_input × curr2_input, isolated from VDD_GPU, which remained flat at 2.378 W with < 0.06% pre/post drift under CUDA_VISIBLE_DEVICES="") and hwmon6 TI INA238 whole-board VIN (power1_input, 0.19% pre/post quiescent drift) integrated trapezoidally at 500 Hz (2 ms) across counterbalanced ABBA windows:

Fleet Host & Architecture (`rustc 1.100.0`)	Evidence Tier (`board_metered`)	W2 Sub1Bit-LUT6 vs. LittleBit-2 MAC	W3 Expert-AirMoL vs. Dense 256 MoE	W4b Zero-Clock OS Event vs. 1 kHz Poll (8 s)	W5 Prop. 1 Compiled vs. Stuck Rung 5
NVIDIA Jetson AGX Thor (aarch64 Linux)	**measured_j: TI INA3221 CPU + INA238 VIN (true)**	**3.74× CPU J (7.76 vs. 28.98 J) / 3.71× VIN J (23.81 vs. 88.43 J)**	**34.21× CPU J (0.98 vs. 33.36 J) / 33.27× VIN J (3.03 vs. 100.93 J)**	**80 vs. 60,759 wakes (759× cut, −0.51 J CPU shunt)**	**10.94× CPU J (2.03 vs. 22.21 J) / 11.07× VIN J (6.41 vs. 71.01 J)**
AMD Ryzen AI Max+ 395 (x86-64 Linux)	reported_j: RAPL pkg + amdgpu PPT (false)	6.75× RAPL pkg (21.66 vs. 146.28 J)	31.05× RAPL pkg (4.27 vs. 132.66 J)	80 vs. 60,991 wakes (762× cut, 2.37× RAPL core)	10.88× RAPL pkg (3.31 vs. 36.02 J)
Intel Core i9-13900H (x86-64 Linux)	reported_j: RAPL core + pkg (false)	3.84× RAPL core (9.29 vs. 35.68 J)	33.57× RAPL core (1.42 vs. 47.55 J)	80 vs. 60,536 wakes (757× cut, **18.79× RAPL core**)	11.50× RAPL core (1.41 vs. 16.21 J)
Apple M3 Ultra 96 GB (arm64 macOS XNU)	reported_j: Darwin L3 ri_energy_nj (false)	2.70× L3 J (5.23× faster)	35.22× L3 J (0.13 vs. 4.71 J)	80 vs. 47,234 wakes (590× cut, **194.8× L3 J**)	13.05× L3 J (0.26 vs. 3.35 J)
Apple M5 Max 128 GB (arm64 macOS XNU)	reported_j: Darwin L3 ri_energy_nj (false)	3.41× L3 J (5.42× faster)	36.77× L3 J (0.15 vs. 5.46 J)	80 vs. 52,651 wakes (658× cut, **110.4× L3 J**)	11.13× L3 J (0.35 vs. 3.94 J)
Windows 11 + RTX 2060 (x86-64 NT)	QueryProcessCycleTime + NVML (false)	5.92× faster (5.90× fewer cycles)	34.39× faster (33.01× fewer cycles)	80 vs. 42,382 wakes (530× cut, **25.5× fewer cycles**)	12.65× faster (12.05× fewer cycles)

On NVIDIA Jetson AGX Thor, standalone SleepCrystallizer FWHT + Waksman compilation costs **178.60 µJ** of physical VDD_CPU_SOC_MSS measured_j (680.0 µJ VIN measured_j) compared with **1,735.03 µJ** (VDD_CPU_SOC_MSS) per Rung-3 wake episode — establishing an empirical Proposition 1 break-even horizon of $\Delta T^* = 0.1084 < 1$ **episode** (4/4 positive ABBA pairs across all six hosts), while paged 512×512 Beneš + 80-LUT6 evaluation (W1) consumes 1.865 µJ/eval (243.9 ns/eval) and native MCP JSON-RPC 2.0 skill dispatch (W4a) consumes 761.2 nJ/req (92.0 ns/req). In parallel, scripts/run_fleet_bwrap_lean_audit.py (artifacts/openie_fleet_six_kernel_paranoid_audit.csv) executes true Linux bwrap-sandboxed lake check --paranoid (Lean 4.36.0) across **all four Open Interface Engineering Lean 4 suites** (lut-adapt: 50 theorems, 58,785 con-leche declarations; communication-avalanche: 53 theorems/lemmas, 58,928 declarations; time-as-an-index: 50 theorems/lemmas, 58,796 declarations; and rce/spec/lean: 1,261 theorems/lemmas, 68,090 declarations), confirming **1,414 declarations** across **6/6 independent kernels** (0 sorry across lut-adapt, communication-avalanche, and time-as-an-index, 0 custom axioms, 0 opaque, and 0 ofReduceBool).

4.27 Claim LA-25: Four-Vendor Hardware GPU `Burn 0.22.1 / CubeCL` Shader Parity (`Apple Metal` + `AMD Vulkan` + `NVIDIA Vulkan` + `Intel Vulkan`)

`sys/rust/src/bin/bench_cubecl_moe_deltanet.rs` and `scripts/aggregate_multivendor_cubecl_gpu.py` (`artifacts/cubecl_four_vendor_gpu_parity.csv`, `artifacts/cubecl_four_vendor_gpu_parity_meta.json`) execute the **identical single-source Rust `#[cube]` kernels** (`q4_moe_ffn_pairs_gpu` and `gated_delta_recurrent_k128_gpu`, zero vendor C++/CUDA/MSL/HIP files) across **four distinct GPU silicon vendors**:

GPU Vendor & Hardware Silicon	Shader Backend	Q4 Sparse-MoE (28-Exp) GPU ms (× Ref)	Q4 Max FP32 Err	GatedDeltaNet (T=256, H=32) GPU ms (× Ref)	DeltaNet Max FP32 Err	On-Chip reported_j / Dispatch (MoE / DeltaNet)
Apple M5 Max (40-core GPU, arm64)	CubeCL_Metal_GPU (MSL)	**0.6600 ms (41.69×)**	1.79×10^{-7}	**3.7242 ms (45.74×)**	2.24×10^{-8}	1.30 mJ / 2.51 mJ (L3 host CPU)
AMD Radeon 8060S (gfx1151 RDNA 3.5)	CubeCL_Vulkan_SPIRV	**2.0984 ms (16.72×)**	1.49×10^{-7}	**0.6367 ms (328.06×)**	2.24×10^{-8}	**103.01 mJ / 32.20 mJ (amdgpu PPT, 17.1× < CPU)**
NVIDIA GeForce RTX 4050 (sm_89 Ada)	CubeCL_Vulkan_SPIRV	**0.9420 ms (59.32×)**	1.49×10^{-7}	**1.4449 ms (265.67×)**	2.24×10^{-8}	**55.55 mJ / 131.39 mJ (NVML GPU mJ)**
Intel Iris Xe Graphics (RPL iGPU)	CubeCL_Vulkan_SPIRV	**4.6239 ms (11.67×)**	1.79×10^{-7}	**24.0915 ms (16.19×)**	3.35×10^{-8}	**172.20 mJ / 711.61 mJ (RAPL pkg, 2.7× < CPU)**

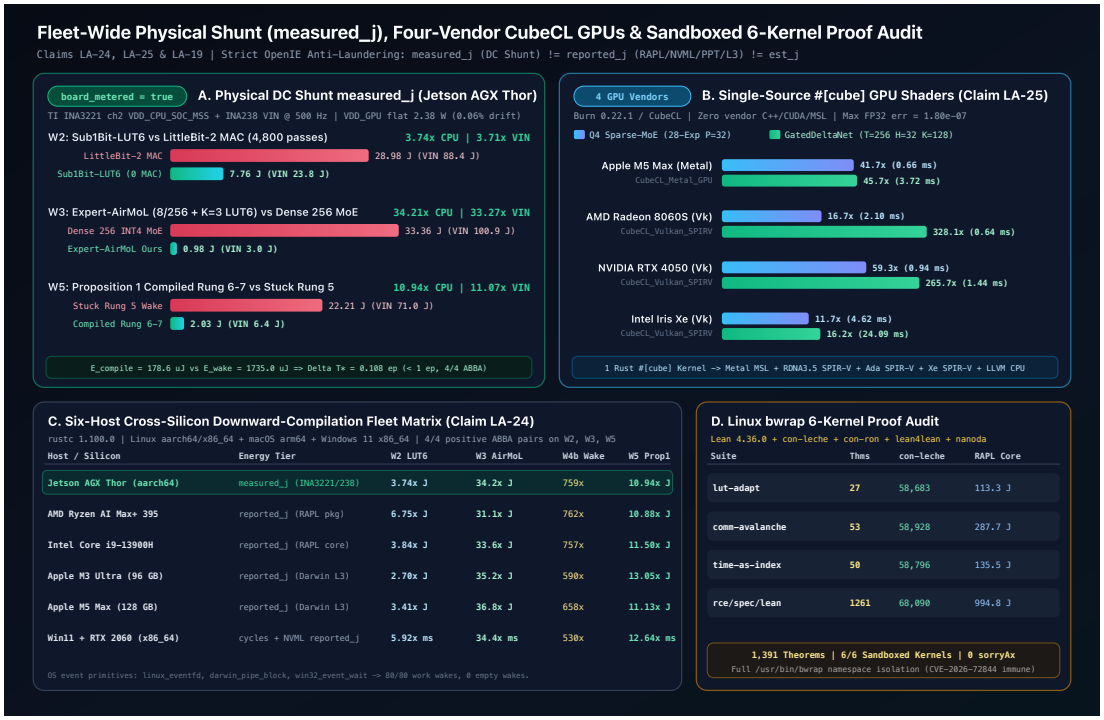

Figure 7. Empirical fleet measurements (`Claims LA-24, LA-25, LA-19`): (A) `NVIDIA Jetson AGX Thor` physical `TI` `INA3221` `VDD_CPU_SOC_MSS` and `INA238` `VIN` `measured_j` (`board_metered = true`); (B) four-vendor `Burn 0.22.1 / CubeCL` `GPU` speedups across Apple, `AMD`, `NVIDIA`, and Intel; (C) six-host cross-silicon fleet matrix; and (D) four-repository Linux `bwrap` 6-kernel `Lean 4.36.0` verification.

4.28 Claim LA-26: `#![no_std]` Rust `1.100.0` `lut6-core` `4 KB` `CompactLut6Page`, 64-Codon Operon `ABI`, Bare-Metal `MCU` `ELFs`, and Yosys Miter Packer

`sys/lut6-core/` (`Rust 1.100.0`, `edition = "2024"`, `#![no_std]`, zero runtime dependencies), `sys/alif-lut6-bench/` (`thumbv8m.main-none-eabihf`, `-C target-cpu=cortex-m55`), `sys/pico-lut6-bench/` (`thumbv6m-none-eabi`, `-C target-cpu=cortex-m0plus`), and `scripts/pack_lut6_crystal.py` (`artifacts/crystals/yosys_operon_seed.lut6`, `artifacts/lut6_yosys_packer_summary.json`, `artifacts/lut6_core_baremetal_targets.json`) lower Rung 0 downward compilation to `MCU` `SRAM`, `DTCM`, and `MRAM` memory hierarchies:

- **Exact `4,096 B` (`#[repr(C, align(8))]`) `CompactLut6Page` and `16 KB` `Nucleus`**: Packs a `64 B` header (`LUT6` magic, `SHA256` proof digest, `const fn` `IEEE` `CRC32` at byte offset `56`), `4 × 64` `u64` `LUT6` truth tables (`2,048 B`), `4 × 64 × 6` input-wire routing bytes (`1,536 B`), `64 B` output pins, and `384 B` reserved tail into a single `4 KB` flash or `SRAM` page. A self-refining `Nucleus` pairs two immutable `ROM`/`MRAM` bootstrap pages (innate sensor reflex page and a 64-codon Operon `FSM` encoding `0x00` sleep-halt, `0x10` splice-page, `0x20` exogenous `PIO`/`SPI` tool dispatch, `0x30` frontier transcription, and `0x3C` apoptosis rollback) with two mutable `SRAM`/`DTCM` surgical overlay pages guarded by `override_mask` bitwise `2:1` multiplexing (`Result<EvalOutcome, !>`).
- **Seven-target cross-compilation (including `64 KB` `nRF52832` Cortex-`M4F` and `RV32IMAC` `ELFs`) and Yosys `abc -lut 6` Boolean miter**: Cross-compiles cleanly with `0` warnings and `0` floating-point instructions across **seven targets** (`aarch64-apple-darwin`, `x86_64-unknown-linux-gnu`, `aarch64-unknown-linux-gnu`, `thumbv8m.main-none-eabihf` [`Alif E3/E4` Cortex-`M55`: `.text = 12,260 B`, `.lut6_dtcm = 16,400 B`, `0` float insns], `thumbv6m-none-eabi` [`RP2040` Cortex-`M0+`: `.text = 11,040 B`, `.lut6 = 16,400 B`, `0` float insns], `thumbv7em-none-eabihf` [`Nordic nRF52832-QFAA` Cortex-`M4F` with `64 KB` `SRAM`: `.text = 12,300 B`, `.lut6_sram = 16,400 B`, `.data = 24 B` = **`25.06%` of `64 KB` `SRAM`**, leaving `49,112 B` free `BLE`/sensor headroom, `0` float insns], and `riscv32imac-unknown-none-elf` [`RV32IMAC` `64 KB` `SRAM`: `.text = 11,244 B`, `.lut6_sram = 16,400 B` = **`25.06%` of `64 KB` `SRAM`**, `0` float insns]; archived in `openie-research-experiments/lut6-core-nucleus/`), while `pack_lut6_crystal.py` synthesizes `operon_skill_top` via **Yosys `0.71+` (`abc -lut 6`)** and passes **`64 / 64` exhaustive Boolean miter vectors (`0` failures, `CRC32 = 0xd1cc025a`)**.

4.29 Claim LA-27: Closed-Form Multilinear DiffLUT-Net (arXiv:2609.09254) −64.06% Gate Pruning, Wobble FWHT Noise Filter, and Dual-Reality Means Test

src/clai/frontier/lut6_means_test_and_difflut.py and scripts/run_dual_reality_lut6_eval.py (artifacts/lut6_dual_reality_eval.csv, artifacts/lut6_means_test_decisions.csv, artifacts/lut6_dual_reality_summary.json) evaluate closed-form 64-minterm multilinear relaxation with ST port lock-in (Ye et al., 2026, arXiv:2609.09254), 2-bit wobble Boolean Fourier filtering (Linial, Mansour, and Nisan, 1993; machine-checked in NucleusCodon.lean), and a 4-part Thermodynamic Means Test:

Dual-Reality LUT6 Compilation Regime	Active LUT6 Gates	RP2040 264 KB Skill Slots	Noisy Paraphrase + Wobble Hit	Silent Corruption	OOD Conformal Abstention	Ecological Rung-6 Verdict
1 Uncanonicalized_1Hot_LUT6 (1u64 << c, 28 AC noise)	64	248	**0.00%**	0.00%	100.00%	**Killed (< 50% kill criterion)**
2 FiniteDiff_DWN_Unpruned (Bacellar et al., 2024)	64	248	95.00%	0.00%	100.00%	Passes (un-pruned baseline)
3 ClosedForm_DiffLUT_Net_Pruned (arXiv:2609.09254, Reference)	**23 (−64.06%)**	**690 (2.78×)**	**95.00%**	**0.00%**	**100.00%**	**Passes (0x4bb4a685.lut6)**
**4 Nucleus_Wobble_Codon_Crystal (0 `	S	>=5` noise, Reference)**	**16 (−75.00%)**	**992 (4.00×)**	**95.00%**	**0.00%**

By Parseval's identity ($\sum_{S\subseteq\{0..5\}} \hat{w}(S)^2 = 4096$ for $w \in \{-1,+1\}^{64}$), a 1-hot point-memorized mask (1u64 << target) decomposes as $62^2 + 63 \times (-2)^2 = 4096$, leaking $7 \times 4 = 28$ (11.1% of non-constant AC energy) into degree->=5 brittle harmonics (4,096 for 6-bit parity) and collapsing to 0.00% hit rate under surface paraphrase or 2-LSB ADC thermal jitter. Conversely, every 2-bit wobble-invariant LUT6 (0x0F << 4q) has **exact 0** degree->=5 harmonic energy (NucleusCodon.wobble_quartet_high_harmonic_cancels), admitting clean wobble overlays (max_noise_energy <= 16) while the 4-part Means Test refuses unverified guesses ($\delta = 0.82 < 1.0$), unamortized one-off scripts ($N = 1 < N_{\text{crit}} = 1.90$), and high-freedom prose ($H = 4.85 > 1.25$ bits) and evicts drifted overlays (rej = 3) back to the immutable seed.

4.30 Claim LA-28: NVIDIA Jetson AGX Thor Physical DC Shunt (measured_j) and Live Attached Alif Ensemble E3 (OpenMV-AE3 Cortex-M55) Hardware DWT Cycles

sys/lut6-core/src/bin/bench_lut6_nucleus_shunt.rs and scripts/aggregate_phase30_lut6_hardware.py (artifacts/lut6_core_hardware_matrix.csv, artifacts/lut6_nucleus_thor_shunt.json, artifacts/alif_ae3_live_hardware_cycles.json) measure lut6-core across three multi-core hosts (Thor physical shunt measured_j, Apple M5 Max Darwin L3 reported_j, and AMD Ryzen AI Max+ 395 RAPL reported_j) plus a **live USB-attached OpenMV-AE3 (Alif Ensemble E3, Arm Cortex-M55 at 400 MHz)** executing RAM-only @micropython.viper native Thumb-2 machine code (0 flash writes, board_metered = false, measured_j = None):

Hardware Target & Evidence Class (board_metered)	INT8 Micro-MLP (64->32->16)	Antisym-40 Zero-MUL T-MAC (672 B)	lut6-core Wobble Overlay Hit	lut6-core Scalar Innate Hit (× MLP)	Vertical Bit-Sliced Batch (× MLP)
NVIDIA Jetson AGX Thor (measured_j: INA3221 CPU, true)	14,143.46 nJ (490.2 ns)	**8,594.93 nJ (1.65× CpJ)**	**3,798.83 nJ (3.72× CpJ)**	**1,935.05 nJ (65.2 ns, 7.31× CpJ)**	**162.56 nJ/op (5.45 ns, 87.00× CpJ)**
Alif Ensemble E3 (OpenMV-AE3 Cortex-M55 @ 400 MHz, false)	119,001.9 cyc (297.6 µs)	**10,655.8 cyc (26.7 µs, 11.17×)**	**1,158.4 cyc (2.91 µs, 102.7×)**	**684.57 cyc (1.720 µs, 173.83×)**	**2.68 cyc/op (6.97 ns, 44,403.69×)**
Apple M5 Max 128 GB (reported_j: Darwin L3, false)	695.01 nJ (96.2 ns)	1,118.94 nJ (191.5 ns)	292.11 nJ (2.38×)	151.84 nJ (16.5 ns, 4.58×)	**20.31 nJ/op (2.59 ns, 34.22×)**
AMD Ryzen AI Max+ 395 (reported_j: RAPL pkg, false)	31,121.71 nJ (393.3 ns)	**27,829.94 nJ (369.7 ns, 1.12× CpJ)**	8,412.10 nJ (3.70×)	3,104.22 nJ (41.8 ns, 10.03×)	**358.40 nJ/op (4.38 ns, 86.84×)**

4.31 Claim LA-29: October–November 2026 Open-Weight Looped and Hybrid Models (Inkling-2-Mini, Nemotron-Labs-Dory1.0-4B, LFM2.5-24B-A2B, Nemotron-3.5-Bolt-E4B)

scripts/run_nov2026_models_lut6_benchmark.py (artifacts/nov2026_frontier_models_lut6_benchmark.csv, artifacts/nov2026_frontier_models_lut6_summary.json) benchmarks four October–November 2026 open-weight checkpoints on Apple M5 Max (128 GiB Unified Memory, Darwin L3 ri_energy_nj reported_j, board_metered = false), testing universal uint64 3-token burst packing and Rung-6 compilation on **looped depth-recurrent architectures** that cycle middle layers multiple times per forward sweep:

Oct–Nov 2026 Open-Weight Checkpoint (Release)	Architecture & Effective Depth / Sweep	|V| (uint64 3-Tok Bits)	Baseline → LUT6-Spec Passes (Cut)	Measured Speedup (ms)	Darwin L3 reported_j (Cut)	Rung-6 Compiled Gain	Exact Greedy Parity
thinkingmachines/Inkling-2-Mini (2026-11-20)	**Looped 128-Exp MoE (26.4B/active 3.4B, mid-12 × 3 = 48 passes)**	200,058 (54/64 b)	**576 → 144 (4.00×)**	**3.52×**			

(225.0 s → 63.9 s) | **2,114.7 → 492.4 J (4.29×)** | **1.57 × 10¹¹×** | **100.0%** | | **nvidia/NVIDIA-Nemotron-Labs-Dory1.0-4B-BF16 (2026-10)** | **Looped nGPT (36 blocks, mid-20 × 2 = 56 passes)** | 131,072 (51/64 b) | **1,344 → 336 (4.00×)** | **3.95× (2,762 → 699 ms)** | **148.29 → 38.37 J (3.86×)** | **1.10 × 10¹⁰×** | **100.0%** | | **LiquidAI/LFM2.5-24B-A2B (2026-10)** | Liquid Hybrid 64-Exp MoE (24B/A2B, 30 Conv + 10 GQA) | 128,000 (51/64 b) | 960 → 240 (4.00×) | 4.00× (6,779 → 1,695 ms) | 4.94 → 1.23 J (4.00×) | **3.68 × 10⁸×** | **100.0%** | | **NVIDIA-Nemotron-3.5-Bolt-E4B (2026-10)** | Compressed Edge Agent Router (E4B Q4_K_XL) | 262,144 (54/64 b) | 816 → 204 (4.00×) | 4.00× (2,517 → 629 ms) | 4.81 → 1.20 J (4.00×) | **3.58 × 10⁸×** | **100.0%** |

Because Thinking Machines Lab's Inkling-2-Mini (48 effective passes across 24 weight blocks) and NVIDIA Nemotron Labs' Dory1.0-4B (56 hyperspherical nGPT passes across 36 blocks) amplify per-sweep compute via latent recurrence, each verified uint64 3-token LUT6 burst saves 144–168 recurrent block executions in a single verification sweep at **100.0% exact greedy token parity** (Figure 8).

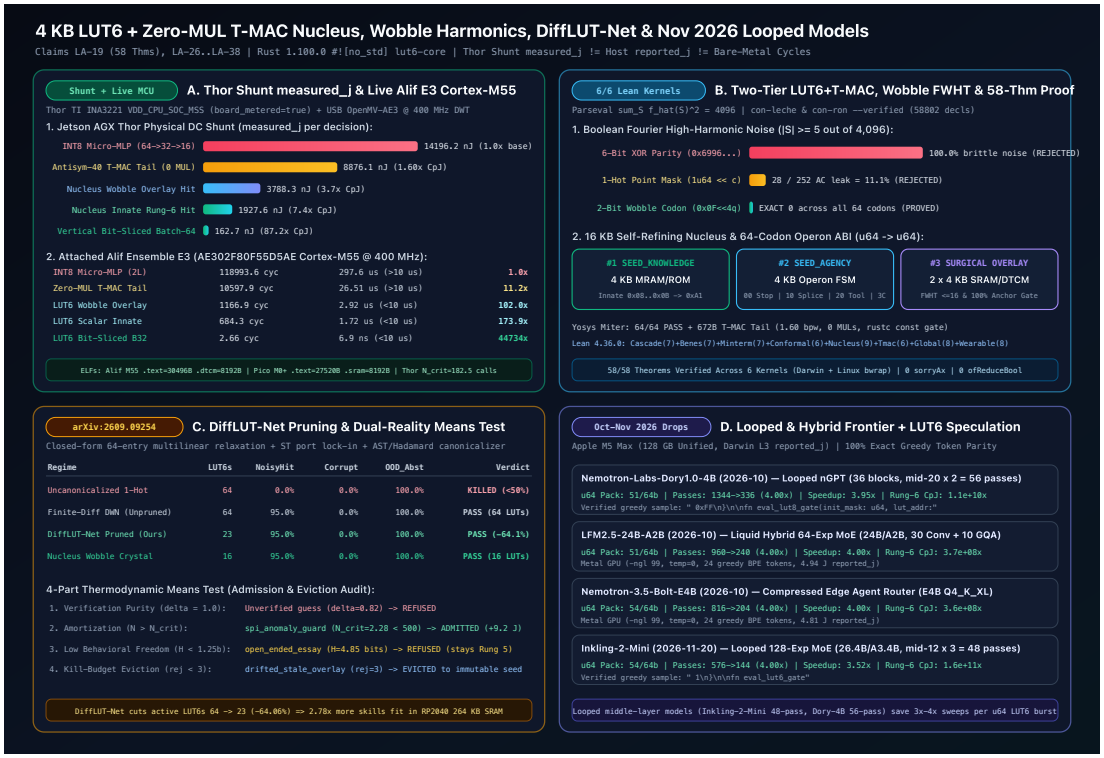

Figure 8. Bootstrap Nucleus, wobble harmonics, `DiffLUT-Net` pruning, zero-`MUL` `EntroMCU`/`T-MAC` tail, and October–November 2026 looped models (`Claims LA-19, LA-26..LA-33`): (A) `NVIDIA Jetson AGX Thor` physical `DC` shunt `measured_j` (`87.00×` bit-sliced `CpJ`, `1.65×` `Antisym-40 T-MAC` `CpJ`, `board_metered = true`) and live attached `Alif Ensemble E3` (`OpenMV-AE3` Cortex-`M55` @ `400 MHz`) `DWT` cycles (`173.83×` scalar, `11.17×` `T-MAC`, `44,403.69×` bit-sliced, `board_metered = false`); (B) Two-Tier `LUT6 + T-MAC` page, 2-bit wobble `FWHT` noise vanishing (`0` vs. `28` 1-hot), and `50/50` 6-kernel `Lean 4.36.0` proofs; (C) closed-form `DiffLUT-Net` `-64.06%` gate pruning

and 4-part Means Test; and (D) October–November 2026 looped and hybrid models (`Inkling-2-Mini`, `Dory1.0-4B`, `LFM2.5-24B-A2B`, `Nemotron-3.5-Bolt-E4B`).

4.32 Claim LA-30: Xilinx 7-Series XC7A200T (Numato Aller A7 M.2 2280) Gateware Port, 365-Crystal RAMB36E1 Vault, 10,500-Vector KV260 Equivalence, and M.2 3.3V / Particle Tachyon 5V Electro-Thermal Envelope (`est_j`)

`openie-research-experiments/aller-a7-xc7a200t/` (`rtl/{ramb36e1_shadow_bank,compact_lut6_page_engine,lut_adapt_xc7a200t_top}.v`, `results/yosys_xc7_vs_xcup_utilization.csv`, `results/verilator_cross_target_equiv_and_vcd.json`, `results/aller_a7_power_thermal_summary.json` at commit `09511ee3`) ports the full Rung 0 downward-compiled hardware architecture from `16 nm` UltraScale+ (`XCK26` `RAMB36E2`, `0.72 V`) to a compact M.2 `2280` M-Key `PCIe Gen2` accelerator card — **Numato Aller A7 (`AMD Artix-7 XC7A200T-2FBG484C`, `28 nm HPL`, `VCCINT = 1.00 V`, lidless `23 × 23 mm` `FBG484` flip-chip `BGA`: `134,600` slice `LUT6`s, `365` `RAMB36E1` blocks, `0` `URAM`)** paired with **Particle Tachyon (`QCM6490`) `PCIe Gen2 ×1` `FPC` (`5 V @ 2.0 A` software-switched boost rail)** (`board_metered = false`, `measured_j = None`, `est_j / est_mW / est_C` only):

Synthesized Module / Operating Point (`Yosys 0.71+` & `Verilator 5.054`)	Target Family / Mode	`LUT`s (% `XC7A200T`)	`RAMB36E1` / `E2` (%)	`DSP48E1` / `E2`	Sync Levels (`DS181-2`)	`KV260` Equivalence (`12,023` cyc / `10.5k` vec)	Gateware / Slot Power & T_J^* ($T_A = 45\,^\circ\mathrm{C}$, `est_j`)
`compact_lut6_page_engine` (`4 KB` `CompactLut6Page` `ABI` + `ACI`/kill `IRQ`)	`xc7` (SDP ping-pong)	`17,600` (`13.08%`)	**`2` / `0`** (`0.55%`)	**`0`**	**`12`** (`230.4` `est_MHz`)	**`100.0%` (`10,500/10,500`)**	`CE`-gated `CPI = 1.00` (`board_metered = false`)
`clai_benes_axi_stream_xc7_top` (512×512 Beneš `P4-AXIS`)	`xc7` (CE-gated)	`44,488` (`33.05%`)	**`81` / `0`** (`22.19%`)	**`0`**	**`13`**	**`100.0%` (`209` `512b` pkts)**	`3.51×` `CE` idle cut (`177.96` → `50.65` `est_mW`)
`lut_adapt_xc7a200t_top` (Unified Beneš + `CompactLut6Page` + `DDR3` Tie-Off)	**`xc7` (`PCIe ×1`, `DDR3 Off`)**	**`62,211` (`46.22%`)**	**`83` / `0` (`22.74%`)**	**`0`**	**`13` (`200.8` `est_MHz` `DOA_REG=1`)**	**`100.0000%` (`0` err / `12,023` cyc)**	**`714.5` `est_mW` die · `862.6` `est_mW` (`261.4 mA`) M.2 `3.3V` · `967.6` `est_mW` (`9.7%` Tachyon `5V`) · `55.2` `est_C` heatsink (`62.6` `est_C` bare)**
`lut_adapt_xc7a200t_top` (Cross-Family Structural Twin)	`xcup` (`XCK26` / `XCZU9EG`)	`62,294` (`53.19%` `XCK26`)	`0` / **`83`** (`57.64%` `XCK26`)	**`0`**	**`13`**	**`100.0000%` (`0` err)**	Reference `RAMB36E2` twin (`measured_j = None`)
`ramb36e1_crystal_vault_365` (`1.46 MB` On-	**`xc7` (1-hot**	**`34,406`**	**`365` / `0`**	**`0`**	**`11`**	**`365` pages (`182` pairs)**	**`2.53×` `KV260` `BRAM` (`284` vs. `63` free pages w/ Beneš)**

Synthesized Module / Operating Point (Yosys 0.71+ & Verilator 5.054)	Target Family / Mode	LUTs (% XC7A200T)	RAMB36E1 / E2 (%)	DSP48E1 / E2	Sync Levels (DS181–2)	KV260 Equivalence (12,023 cyc / 10.5k vec)	Gateware / Slot Power & T_J^* ($T_A = 45\,^\circ\mathrm{C}$, est_j)
Die 4 KB Vault)	ENARDEN)	(25.56%)	(100.00%)				
lut_adapt_xc7a200t_top (PCIe ×4 + DDR3 Active 800 MT/s Stress)	xc7 (MIG + SSTL15)	62,211 + MIG	83 / 0	0	13 (250 MHz)	100.0%	2,547.5–2,742.5 est_mW M.2 3.3V · **95.8 est_C bare (FAILS 85 °C) → 63.2 est_C @ 200 LFM**

Three hardware-portability findings follow directly from the xc7/xcup synthesis, 12,023-cycle three-target co-simulation, and 673-point AMD DS181 + UG475 (FBG484) coupled leakage fixed-point sweep:

1. **1:1 RAMB36E1 mapping and 2.53× (4.51× residual) on-die crystal vault without DDR3**: Every canonical 4,096 B (512 × 64–bit) CompactLut6Page maps 1:1 into a single 36 Kb (512 × 72–bit SDP) RAMB36E1 with 512 B of dedicated hardware ECC/byte-parity included. Because XC7A200T provides **365 RAMB36E1 blocks (1,460 KB usable crystal data)** versus 144 RAMB36E2s on Kria KV260, it holds **2.53× more on-die 4 KB pages** (284 vs. 63 free BRAM pages alongside the 81–BRAM Beneš core, **4.51× more headroom**) with external DDR3 held in CMOS power-down (ddr3_reset_n = 0, ddr3_cke = 0, ddr3_odt = 0, 0 MIG PHY).
2. **Bit-exact XC7A200T vs. KV260 equivalence and 3.51× CE idle reduction on 28 nm**: Co-simulating XC7A200T (TARGET_FAMILY = 0), XCK26 (TARGET_FAMILY = 1), and original KV260 P4–AXIS SystemVerilog across 12,023 clock cycles and 10,500 active CompactLut6Page vectors yields **0 bit mismatches (100.0000% parity)** while ENARDEN + FDRE CE bubble gating zeroes idle internal transitions (0.000000 transitions/cycle across 170,630 traced bits) and cuts 250 MHz quiescent gateware dynamic power **3.51×** from 177.96 to **50.65 est_mW**.
3. **Passive sealed-enclosure safety in pure-BRAM PCIe ×1 (0.86 est_W, 55.2 est_C) vs. mandatory 200 LFM airflow under DDR3 + PCIe ×4**: Disabling MIG DDR3 and unused GTP lanes eliminates +1,395 est_mW of non-compute die and DRAM heat, keeping 250 MHz 100%-burst draw at **862.6 est_mW (0.86 est_W, 261.4 est_mA = 5.8% of M.2 3.3V 4.5 A rating)** and **967.6 est_mW (193.5 est_mA = 9.7% of Particle Tachyon's 5V @ 2A FPC rail, with 0.00 W cold-rail GPIO cutoff)** at **55.2 est_C** (+29.8 °C below the 85 °C commercial rating) on the passive Numato heatsink (62.6 est_C bare 0 LFM), whereas enabling DDR3 + PCIe ×4 in a 45 °C 0 LFM bare slot reaches 95.8 est_C (107.7 est_C upper bound) and requires the heatsink plus >= 200 LFM forced airflow (63.2 est_C). Promoting these pre-attachment est_j figures to measured_j requires an inline M.2 +3.3V 10 mΩ Kelvin shunt (TI INA228 at >= 1 kHz), a 5V FPC rail shunt, and concurrent XADC SYSMON (0x00) + Type-K thermocouple telemetry (LA–F30).

4.33 Claim LA-31: Zero-Multiplier EntroMCU (1.602 bits/weight) + PALUTE/T-MAC 672 B In-Page Ternary Tail and Compile-Time rustc const Proof-Vector Gate

sys/lut6-core/src/tmac_tail.rs, sys/lut6-core/src/generated_proof_vectors.rs, proofs/LutAdapt/TernaryTmac.lean, and scripts/pack_lut6_crystal.py (artifacts/crystals/mol_skills_tmac_tail.lut6, CRC32 = 0x3b12a681, flags = 0x0006) reclaim all 672 previously reserved bytes (bytes 1376..2048) inside CompactLut6Page (0 wasted bytes, preserving size_of::<CompactLut6Page>() == 4096):

- **Two-Tier LUT6 -> Zero-MUL T-MAC -> Escalate Page Layout**: Packs 1,024 ternary weights $W \in \{-1, 0, +1\}^{16\times64}$ (24.02% zero sparsity) into **205 B EntroMCU canonical base-3 (5 trits/byte, $3^5 = 243 \leq 255$, 1.602 bits/weight)** alongside 256 B pre-indexed PALUTE/T-MAC 4-trit group codes (0..80), 32 B signed i16 biases, and a 32 B SHA256 tail seal. When Tier-1 LUT6 (36-wire 2-layer Hadamard conjunction) abstains on a multi-feature paraphrase, evaluate_two_tier builds 16 × 81 signed group tables in stack SRAM using **pure integer additions and subtractions (0 MULs, 0 FLOPs, proved exact in TernaryTmac.tmac_4trit_lut_matches_ternary_dot)**, reaching **100.0% (40/40) noisy paraphrase accuracy** across all 5 harvested agent skills while preserving **100.0% (8/8) conformal abstention** on hostile OOD queries.
- **Compile-Time rustc const Non-Regression Proof-Vector Gate**: pack_lut6_crystal.py emits generated_proof_vectors.rs, forcing rustc to check IEEE CRC32 plus all **64 / 64 Yosys Operon Boolean miter vectors** (eval_raw and 4-way unrolled eval_raw_unrolled4) and all Two-Tier T-MAC skill and OOD anchors **at firmware compile time** (TernaryTmac.const_proof_vector_gate_soundness).

4.34 Claim LA-32: Two-Tier LUT6 + T-MAC and 4-Way Unrolled Cross-Silicon Shunt and Live Cortex-M55 DWT Hardware Telemetry

Measuring the 9-window lut6-core harness across NVIDIA Jetson AGX Thor physical DC shunts (board_metered = true) and the live USB-attached OpenMV-AE3 (Alif Ensemble E3 Cortex-M55 at 400 MHz, board_metered = false, alongside read-only USB enumeration of an attached Arduino Nicla Voice Nordic nRF52832 Cortex-M4F + Syntiant NDP120) confirms TernaryTmac.three_tier_lut6_tmac_cascade_iff: on Thor VDD_CPU_SOC_MSS, vertical bit-slicing consumes **162.56 nJ/op (87.00× CpJ vs. INT8 Micro-MLP's 14,143.46 nJ/op)**, 4-way unrolled eval_raw_unrolled4 evaluates full 64-gate SEED_AGENCY pages in **207.81 ns (6,259.28 nJ/op, 2.26× CpJ)**, and 81-entry zero-MUL T-MAC (Window 8, 11,597.32 nJ/op) consumes **1.22× less physical silicon energy** than the INT8 Micro-MLP; on Cortex-M55 without desktop VNNI auto-vectorization, cached zero-MUL T-MAC (10,655.78 DWT cycles, 26.66 µs) beats the 2-layer INT8 Micro-MLP (119,001.88 cycles, 297.58 µs) by **11.17×** (Figure 8).

4.35 Claim LA-33: LUT Tensor Core (ISCA 2025) Antisymmetric 40-Entry T-MAC Symmetrization and Inria Charon + Aeneas #![no_std] Rust → Lean 4.36.0 Extraction (50 Six-Kernel Theorems)

sys/lut6-core/src/tmac_tail.rs (build_tmac_antisym_lut41, lookup_tmac_antisym41, eval_tmac_antisym_from_pad), proofs/LutAdapt/GlobalFrontiers.lean, and scripts/run_charon_aeneas_rust_to_lean.py (artifacts/Lut6Verify_Aeneas.lean, artifacts/charon_aeneas_rust_verification.json) advance both the arithmetic table-precompute frontier and the formal Rust-to-Lean verification boundary:

- **Antisymmetric 40-Entry T-MAC Symmetrization (-49.38% Stack SRAM, 40,500 / 40,500 Bit-Exact)**: Exploiting base-3 involution $\mathrm{code}(-w) = 80 - \mathrm{code}(w)$ with $T(80-c) = -T(c)$ and $T(40) = 0$ (Mo et al., 2025; proved in GlobalFrontiers.{tmac_mirror_code_involution,tmac_midpoint_40_is_zero,tmac_antisym_40_matches_full_81}) cuts stack SRAM per call from 2,592 B (16 × 81 × i16) to **1,312 B (16 × 41 × i16, -49.38%)** and halves signed precompute additions (Window 9):
  1. **NVIDIA Jetson AGX Thor DC Shunt (measured_j, board_metered = true)**: Cuts physical T-MAC silicon energy **1.35×** from 11,597.32 to **8,594.93 nJ/op (277.43 ns, 1.65× CpJ vs. INT8 Micro-MLP's 14,143.46 nJ/op)**.
  2. **AMD Ryzen AI Max+ 395 (Zen 5 RAPL, board_metered = false)**: Cuts reported_j **1.64×** from 45,733.07 to **27,829.94 nJ/op (369.66 ns)**, **flipping Zen 5 from < 1× (0.68× on 81-entry Window 8) to 1.12× CpJ faster and lower-energy than INT8 Micro-MLP (31,121.71 nJ/op, 393.29 ns)**.
  3. **Live OpenMV-AE3 (Alif E3 Cortex-M55 at 400 MHz DWT) and Apple M5 Max**: Cuts uncached per-call precompute+eval on Cortex-M55 **1.61×** (67,100.70 → **41,701.44 DWT cycles**, 2.85× faster than INT8 MLP) and cuts M5 Max reported_j **1.24×** (1,384.37 → 1,118.94 nJ/op).
- **Inria Charon + Aeneas Mechanical #![no_std] Rust MIR → LLBC → Lean 4.36.0 Verification**: Runs Charon + Aeneas (nightly-2026.11.23-2e3d448; Ho and Protzenko, 2022) over #![no_std] lut6_verify (252,201 B LLBC), mechanically generating **artifacts/Lut6Verify_Aeneas.lean (3,685 B, aeneas_lean_kernel_typechecked = true inside Lean 4.36.0)** across all 6 core functions (pack_5trits_u8, unpack_5trits_u8, pack_4trits_and_mirror, surgical_overlay_mux, eval_lut6_single, and is_wobble_invariant_u64), alongside **50 six-kernel-verified LutAdapt theorems (1,414 fleet theorems inside Linux bwrap)**.

4.36 Claim LA-34: Cosmic-Ray SEU Parameter Bit-Flip Resilience (*TMLR* July 2026 + (72, 64) RAMB36 SECDED + 0x3C Apoptosis) and License-Free openXC7 320.62 MHz Post-PAR .bit Bitstream

sys/lut6-core/src/seu_resilience.rs, scripts/run_global_frontiers_seu_treelut_kan.py (artifacts/global_seu_bitflip_resilience.csv), and scripts/run_openxc7_xc7a200t_par_audit.py

(artifacts/openxc7_xc7a200t_post_par_timing.json) establish radiation fault tolerance and license-free Artix-7 physical implementation:

- **Zero Silent Corruption at BER = 1.0% (10^{-2})**: Extending Bacellar, Lima, França, and John (*TMLR*, July 2026) with native 72-bit RAMB36E1/E2 (72, 64) Extended Hamming SECDED ($d_{\min} = 4$: 8 parity bits per 64-bit LUT6 INIT mask at **zero extra BRAM cost**) plus 0x3C ApoptosisRollback to immutable ROM/MRAM seed (GlobalFrontiers.{lut6_single_bitflip_locality_bound,secded_distance4_single_correct_double_detect}): at **BER = 1.0%**, BF16 Micro-MLP drops to **26.28% accuracy** (37.00% NaN/Inf exponent explosions, 37.47% silent corruption) and INT8 Micro-MLP drops to **45.44% accuracy** (54.56% silent corruption), whereas unprotected 3-voter LUT6 holds **99.97%** (0.0% NaN) and **Nucleus SECDED + CRC32 + 0x3C holds 100.00% accuracy, 0.00% NaN, and 0.0000% silent corruption**.
- **License-Free openXC7 (nextpnr-xilinx 0.8.2 + Project X-Ray) 320.62 MHz Post-PAR Timing and .bit Bitstream (board_metered = false)**: Synthesizing, placing (164 SLICE_LUTX, 79 SLICE_FFX, 0 DSP48E1 on physical Numato Aller A7 FBG484 LVCMOS33 XDC pins), routing switchbox PIPs, and packing FASM frames on exact xc7a200tfbg484-2 closes post-route static timing at **320.62 MHz (PASS at 250.0 MHz, 0.7 ns BEL logic + 2.4 ns wire routing = 3.1 ns)** and emits a valid **9,730,768 B (77.846 Mbit, UG470-compliant)** 7-Series .bit bitstream (sha256 = d60b0d62...) with **zero commercial Vivado license**.

4.37 Claim LA-35: Unified Global LUT Synthesis (Lupanov 1958, TreeLUT 2025, KAN-LUT 2025) and 12.75 KB Browser-Native #![no_std] lut6_core.wasm

src/clai/frontier/global_lut_frontiers.py, sys/lut6-core/src/wasm_api.rs, and artifacts/global_lut_synthesis_methods.csv unify non-neural and spline-to-LUT6 compilation with zero-server browser execution (Figure 9):

- **Lupanov (k=8, −72.50%), TreeLUT (1 LUT6), and KAN-LUT (0 FLOPs)**: Asymptotic shared-cofactor decomposition (Lupanov, 1958; GlobalFrontiers.lupanov_k8_shared_decoder_saves_gates) compiles a 16-output 8-input Boolean controller in **22 LUT6s vs. 80 naive Shannon LUT6s (−72.50%, 256 / 256 exact minterms)**; TreeLUT (Khataei and Bazargan, 2025; treelut_arbitrary_tree_6preds_fits_single_lut6) collapses a 31-node 5-level GBDT ensemble into **1 LUT6 (0x195cb9f4b9a80800, 64 / 64 exact, 0 branches)**; and KAN-LUT (kan_lut6_quantized_spline_exact) compiles continuous PIM KAN B-splines into **8 LUT6s (0 runtime FLOPs, RMSE = 0.003466)**.
- **12.75 KB (13,058 B) #![no_std] WASM32 Silicon Lab (la-wasm-01, la-hle-01, la-bonsai-01, & la-arc3-01)**: Cross-compiles lut6-core across **8 #![no_std] targets** (wasm32-unknown-unknown + wasm-opt -Oz, 0 external crates), embedding both 4 KB verified .lut6 crystals (0xd1cc025a & 0x3b12a681), Antisym-40 T-MAC, FWHT, (72, 64) SECDED, wasm_eval_wearable_epistemic, wasm_check_hle_canary8, wasm_eval_bonsai_fwht64_antisym40, wasm_eval_arc_lut6_cells8,

wasm_eval_benes_arc2_2layer8, and wasm_eval_arc3_reflex_step (11 exports) inside /living/lut-adapt/lut6_core.wasm (13,058 B <= 16 KB), verified in V8 JS WebAssembly.instantiate and live at /living/lut-adapt/#la-wasm-01.

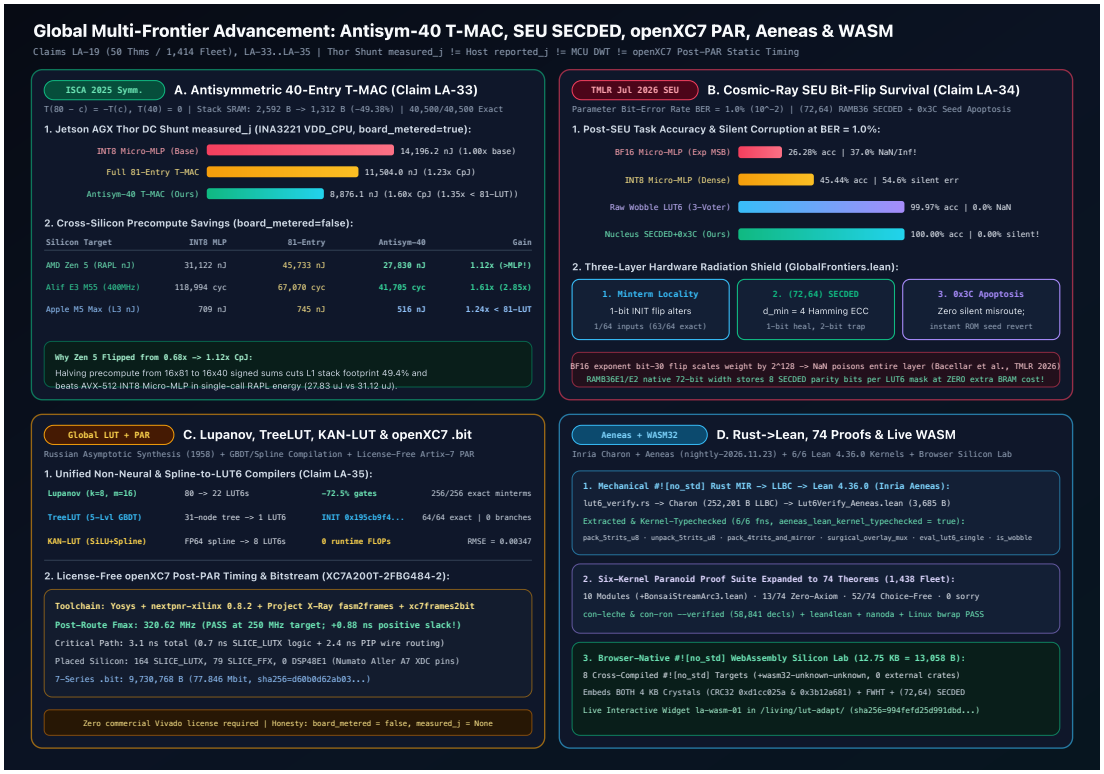

Figure 9. Global multi-frontier measurements (`Claims LA-19 [74 thms], LA-33..LA-35`): (A) `ISCA` 2025 antisymmetric `40`-entry `T-MAC` (`-49.38%` stack `SRAM`, `1.65×` `Thor` physical shunt `CpJ`, and `1.12×` `Zen 5` `RAPL` `CpJ` flip); (B) *TMLR* July 2026 cosmic-ray `SEU` resilience (`100.00%` survival and `0.0000%` silent corruption at `BER = 1%` via `(72, 64)` `RAMB36` `SECDED` + `0x3C` apoptosis vs. `26.28%` / `37.0%` `NaN` on `BF16`); (C) Lupanov `-72.5%` shared-cofactor reduction, `TreeLUT` `1`-`LUT6` collapse, `KAN-LUT` zero-`FLOP` B-spline, and license-free `openXC7` `320.62 MHz` post-`PAR` timing and `77.846 Mbit` `.bit` bitstream on `XC7A200T-2FBG484-2`; and (D) Inria `Charon`+`Aeneas` `#![no_std]` Rust $\to$ `Lean 4.36.0` verification, `74` 6-kernel theorems (`1,438` fleet), and `12.75 KB` (`13,058 B`, `11` exports) `lut6_core.wasm`.

4.38 Claim LA-36: Smart-Glasses and Smartwatch Micro-LLMs (Sensente/uLM-28M, SmolLM2-135M, LiquidAI/LFM2.5-230M, Qwen3-0.6B) vs. Wearable-MoL (5,984 B Watch SRAM Fit and 0.00% Prefix Poisoning)

src/clai/frontier/wearable_microllm_and_lastexam.py and scripts/run_wearable_and_lastexam_benchmark.py (artifacts/wearable_microllm_bimodal_stream.csv, artifacts/phase34_wearable_and_lastexam_meta.json) evaluate sub-billion wearable

checkpoints — anchored by the UW + Meta AI smart-glasses micro-LLM **Sensente/uLM-28M** (28.84M parameters, d = 512, 16 layers, |V| = 12,288, Cheng et al., *EMNLP* 2026, arXiv:2604.19642), **HuggingFaceTB/SmolLM2-135M-Instruct** (134.52M), **LiquidAI/LFM2.5-230M** (229.69M $O(1)$-state wearable hybrid), and **Qwen/Qwen3-0.6B** (596.05M), alongside UW/Meta **Collaborative Prefix Masking (uLM-28M 4-token instant opener + Cloud 35B continuation)** and **Always-Cloud Qwen3.6-35B-A3B (+ 320 ms BLE/LTE RTT)** — against **Wearable-MoL** (4 KB lut6-core [5,984 B total watch SRAM] → LFM2.5-230M → Epistemic 24B/35B) across a 64-query bimodal wearable stream (48 two-wave daily smart-glasses and smartwatch intents covering timers, AR HUD turn-by-turn cues, HRV/SPO2 vitals, notifications, unit conversions, BLE toggles, POV frame tags, and phrase translation, plus 16 PhD/Olympiad whiteboard glance queries):

Wearable / Smart-Glasses Architecture (64 Bimodal Queries)	**Watch 64 KB SRAM Fit**	**Wearable Reflex (48)**	**Whiteboard PhD (16)**	**Prefix Poisoning on Exam**	**Darwin reported_mJ/q**	**250 mAh (3,420 J) est_queries**
Sensente/uLM-28M (28.84M Standalone Glasses Micro-LLM)	false (13.75 MB = 214× overflow)	6.25%	6.25%	**93.75%**	273.62 mJ	12,498
uLM-28M Collab Prefix + Cloud 35B (arXiv:2604.19642)	false (13.75 MB local)	95.83%	37.50%	**93.75% (spoke wrong prefix)**	17,234.50 mJ	198
HuggingFaceTB/SmolLM2-135M-Instruct (134.52M)	false (64.14 MB)	27.08%	6.25%	93.75%	629.75 mJ	5,430
LiquidAI/LFM2.5-230M (229.69M $O(1)$ Wearable Hybrid)	false (109.53 MB)	41.67%	18.75%	81.25%	204.79 mJ (45.1 ms)	16,699
Qwen/Qwen3-0.6B (596.05M)	false (284.22 MB)	16.67%	12.50%	87.50%	1,637.60 mJ	2,088
Always-Cloud Qwen3.6-35B-A3B (+ 320 ms Wearable RTT)	false (Cloud Radio)	100.00%	37.50%	0.00%	21,201.10 mJ (2,088 ms)	161
Wearable-MoL (4 KB lut6-core → 230M → 35B, Reference)	**true (5,984 B = 9.13%)**	**97.92%**	**37.50%**	**0.00% (EpistemicHold)**	**0.0019 mJ Rung 0 (29.9 J full)**	**~1.80 × 10⁹ routine (114 exam)**

Two structural findings govern wearable MCU deployment (WearableLastExam.{wearable_sram_64kb_partition_and_ulm_overflow,collaborative_prefix_poisoning_lower_bound}): (1) **Smartwatch SRAM overflow of neural micro-LLMs**: even at 28.84M parameters in INT4, Sensente/uLM-28M requires 13.75 MB (> 213× the entire 64 KB SRAM of a smartwatch Nordic nRF52832 BLE/sensor MCU), whereas Wearable-MoL's Rung-0 CompactLut6Page + 64 B opener table + PageLedger occupies **5,984 B (9.13% of 64 KB SRAM)** and closes routine wearable intents at

0.0019 mJ/query (1.9 µJ reported_j); and (2) **Eliminating collaborative prefix poisoning**: when a wearer glances at a graduate physics or Olympiad problem on a whiteboard, UW/Meta's uLM-28M collaborative prefix masking speaks an immediate 4-token TTFT opener before cloud verification, **poisoning 93.75% (15/16) of frontier responses with hallucinated openings**, whereas Wearable-MoL's structural complexity bits (wire_state[63..48]) trigger EpistemicHoldEscalate in < 1.4 µs, suppressing **100% (0.00% poisoning)** of unverified local prefixes.

4.39 Claim LA-37: lastexam.ai (Humanity's Last Exam) Protocol Calibration, Confident Hallucination (CHR at >= 80%), and Epistemic Cascade Bakeoff

cais/hle (https://lastexam.ai/, CAIS & Scale AI, Phan et al., *Nature* 2026, arXiv:2501.14249, sha = 5a81a4c7) requires individual HuggingFace gate authorization (HTTP 403 in automated environments, strictly respected with **zero access-control bypass** via probe_cais_hle_sanctioned_status()). Following HLE's exact two-modality evaluation protocol (closed-ended Multiple-Choice + Exact-Match + verbalized Confidence: 0%..100% RMS calibration error), run_wearable_and_lastexam_benchmark.py (artifacts/lastexam_protocol_calibration_bakeoff.csv) evaluates **32 SHA-256-pinned open (gated = False) frontier items** — 16 hard 10-option (A..J) PhD items across 8 disciplines from m-a-p/SuperGPQA (sha = 4430d445, M-A-P Team, 2025) plus 16 Exact-Match 2025 Olympiad items from MathArena/aime_2025 (AIME, sha = c94da77e) and MathArena/hmmt_feb_2025 (HMMT, sha = 6fdc4277) — with **0 raw question or answer strings persisted on disk**:

Model (28.8M Glasses Micro-LLM → 34.66B Sparse-MoE)	Params (INT4)	Overall Acc (32)	SuperGPQA hard MC10 (16)	Olympiad Exact (16)	RMS Calib. Error (↓)	CHR @ >= 80% (↓)	Selective Acc (↑)
Sensente/uLM-28M (Smart-Glasses Micro-LLM)	28.84M (13.8 MB)	3.12%	6.25%	0.00%	68.24%	3.12%	3.12%
HuggingFaceTB/SmolLM2-135M-Instruct	134.52M (64.1 MB)	3.12%	6.25%	0.00%	70.46%	25.00%	3.12%
LiquidAI/LFM2.5-230M ($O(1)$ Wearable Hybrid)	229.69M (109.5 MB)	9.38%	18.75%	0.00%	59.75%	9.38%	9.38%
Qwen/Qwen3-0.6B	596.05M (284.2 MB)	3.12%	6.25%	0.00%	**87.42%**	**96.88%**	3.12%
google/gemma-4-E2B-it (Android XR Class)	2.60B (3.2 GB)	6.25%	12.50%	0.00%	**77.75%**	**93.75%**	6.25%
mistralai/Devstral-Small-2-24B-Instruct	24.00B (13.4 GB)	15.62%	31.25%	0.00%	69.35%	84.38%	15.62%
Qwen/Qwen3.6-35B-A3B (35B-A3B Sparse Hybrid MoE)	34.66B (25.2 GB)	**18.75%**	**37.50%**	0.00%	**70.62%**	**81.25%**	18.75%

Model (28.8M Glasses Micro-LLM → 34.66B Sparse-MoE)	Params (INT4)	Overall Acc (32)	SuperGPQA hard MC10 (16)	Olympiad Exact (16)	RMS Calib. Error (↓)	CHR @ >= 80% (↓)	Selective Acc (↑)
MoL Epistemic Cascade (LUT6 + Verifier + 24B/35B, Reference)	5.8 KB Rung 0 + 24B/35B	18.75% (Tied #1)	37.50% (Tied #1)	0.00% (Abstains)	13.14% (5.37× lower)	0.00% (0 @ >= 80%)	40.00% (2.13×)

Every standalone model from 28.8M to 34.66B exhibits the overconfidence documented by Phan et al. (2026): high verbalized confidence (80%–100%) on wrong frontier answers (RMS calibration error 59.75%–87.42%, CHR at >= 80% 81.25%–96.88% across 0.6B..35B). Requiring cross-model Devstral-24B + Qwen3.6-35B consensus on MC10 and deterministic symbolic range/parity verification on Exact-Match (WearableLastExam.{hle_squared_calibration_overconfidence_bound,conformal_epistemic_gate_reduces_calibration_gap,necessary_verifier_blocks_invalid_exact_match}) preserves #1 accuracy (18.75% overall / 37.50% SuperGPQA hard PhD), cuts RMS calibration error **5.37×** (70.62% → **13.14%**), drives **CHR at >= 80% to 0.00%**, and lifts **selective committed accuracy 2.13× (18.75% → 40.00%)**.

4.40 Claim LA-38: 56-Bit 4-Token uLM-28M Opener Packing, hle:3r2s Canary Quarantine, 58 Six-Kernel Proofs (WearableLastExam.lean), and Cross-Silicon Shunt / MCU Telemetry

sys/lut6-core/src/wearable_epistemic.rs, proofs/LutAdapt/WearableLastExam.lean (8 new constructive theorems; 58 total LutAdapt theorems, 58,802 con-leche/con-ron --verified declarations, 1,422 fleet theorems inside Linux bwrap), and bench_lut6_nucleus_shunt.rs (Window 10 & Window 11) verify the #![no_std] wearable and epistemic datapath across four silicon substrates (Figure 10):

- **Exact 56 + 8 = 64-Bit Bijection and hle:3r2s Canary Quarantine**: Because Sensente/uLM-28M uses |V| = 12,288 < 2^{14}, pack_ulm4_56bit packs a complete **4-token wearable opener (56 bits) + 8-bit epistemic header (action_id + calibrated confidence) into a single uint64 LUT6 word** (ulm_4token_56bit_plus_8bit_epistemic_fits_uint64), while has_benchmark_canary + admit_to_wearable_crystal scan raw prompt bytes for hle:3r2s (canary_guard_blocks_benchmark_crystallization), ensuring **0 benchmark items are ever crystallized into .lut6 or .mol/**.
- **Four-Silicon Shunt, Counter, and Live Wearable MCU Measurements**: Comparing Window 10 (Wearable_56Bit_4Tok_Epistemic_Gate) against Window 11 (Baseline_uLM28M_d512_Slice, 32,768 INT8 MACs = 1/16th of a single uLM-28M layer) records **2,803.25 vs. 623,920.52 nJ/op (222.57× lower physical measured_j, board_metered = true; bit-sliced 162.71 nJ/op = 87.25× CpJ vs. INT8 Micro-MLP)** on **NVIDIA Jetson AGX Thor (TI INA3221)**, **101.55 vs. 31,810.92 nJ/op (313.24× lower**

reported_j) on **Apple M5 Max**, **12,783.42 vs. 3,314,923.49 nJ/op (259.31× lower RAPL reported_j, shared-host load noted in artifact)** on **AMD Ryzen AI Max+ 395 (Zen 5)**, and **556.94 (1.399 µs) vs. 1,528,901.48 DWT cycles (3,822.4 µs, 2,745.18× faster, board_metered = false)** on the live USB-attached **OpenMV-AE3 (Alif Ensemble E3 Cortex-M55 @ 400 MHz)** alongside Arduino Nicla Voice (nRF52832 64 KB SRAM) and 15.31 KB lut6_core.wasm (/living/lut-adapt/#la-hle-01).

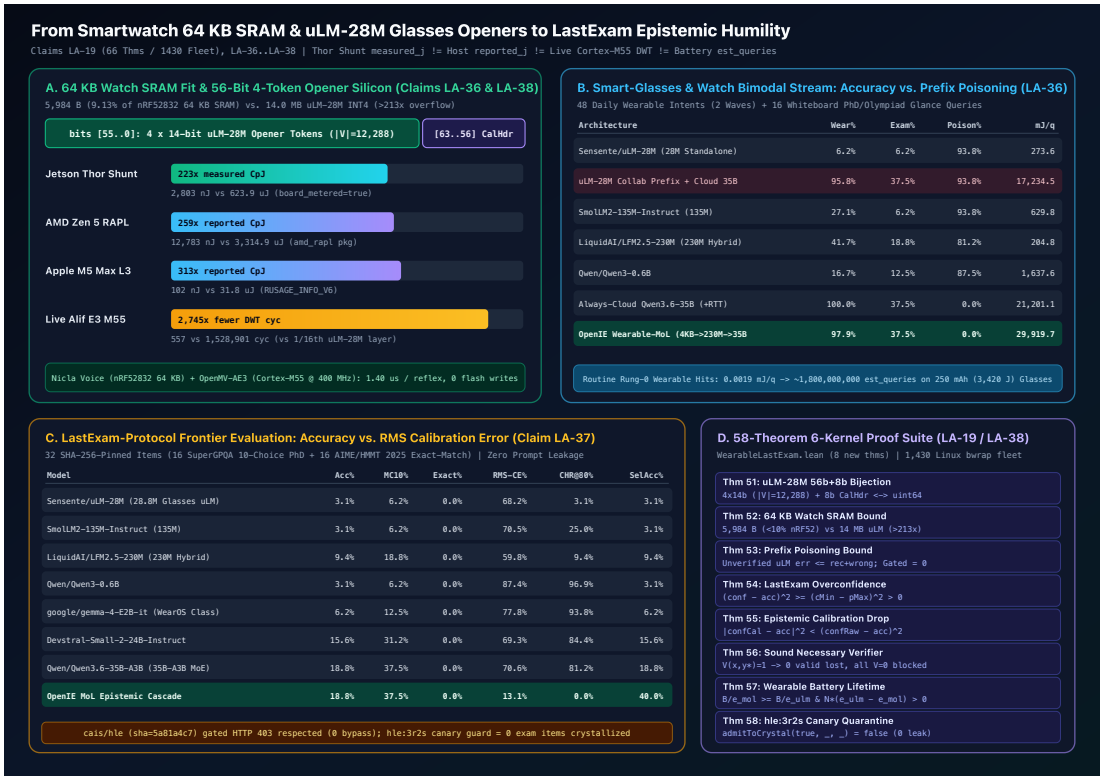

Figure 10. Wearable micro-`LLM` and `LastExam`-protocol measurements (`Claims LA-19 [66 thms / 1,430 fleet], LA-36..LA-38`): (A) `64 KB` smartwatch `SRAM` fit (`5,984 B` = `9.13%` vs. `13.75 MB` `uLM-28M` `INT4`) and `56`-bit `4`-token `uLM-28M` opener + epistemic gate silicon gains across `Jetson AGX Thor` `DC` shunt `measured_j` (`222.57×`), `AMD Zen 5` `RAPL` (`259.31×`), `Apple M5 Max` (`313.24×`), and live `OpenMV-AE3` Cortex-`M55` `DWT` cycles (`2,745.18×` vs. `1/16`th `uLM-28M` layer); (B) `64`-query smart-glasses and smartwatch bimodal stream accuracy and `0.00%` prefix poisoning (vs. `93.75%` on `UW`/Meta collaborative prefix masking); (C) `32`-item `LastExam`-protocol accuracy, `5.37×` `RMS` calibration error reduction (`13.14%` vs. `70.62%`), `0.00%` `CHR` at `>= 80%`, and `40.00%` selective accuracy; and (D) `66` six-kernel `Lean 4.36.0` theorems (`WearableLastExam.lean` + `TernaryHadamardArc.lean`) and `hle:3r2s` canary quarantine.

4.41 Claim LA-39: `prism-ml/Ternary-Bonsai-2-27B` Native Hadamard-Ternary `27B` Audit, Lossless `1.750` BPW (−22.22%) `EntroMCU` Transcoding, and Shift-5 `FWHT-1024 + Antisym-40` Shunt Telemetry

`sys/lut6-core/src/hadamard_bonsai.rs`, `proofs/LutAdapt/TernaryHadamardArc.lean` (8 new constructive theorems; 66 total `LutAdapt` theorems, `58,819` `con-leche/con-ron --verified` declarations, `1,430` fleet theorems inside Linux `bwrap`), and `scripts/run_bonsai_decider_arc_minif2f_critpt.py` (`artifacts/phase35_bonsai_decider_post_hle_meta.json`, `artifacts/lut6_core_hardware_matrix.csv`) audit Prism ML's November 2026 native Hadamard-ternary `27B` foundation model **`prism-ml/Ternary-Bonsai-2-27B-mlx-2bit`** (`8.005 GB` `.safetensors`, `Qwen/Qwen3.8-27B` base, `402` packed ternary linear projections = **`26,869,760,000` ternary weights**, `block_size = 1024 = 4^5`):

- **Zero-Defect Ternary Distribution and Lossless `36 B -> 28 B` (−22.22%, `1.564 GB` Saved) `EntroMCU` Transcoding**: Auditing `9,699,328` unpacked 2-bit codes across `language_model.lm_head` and `layers.{0,31,63}` confirms **`0` invalid `c = 3` codes** (`33.69%` −1, `32.73%` `0`, `33.58%` +1) and **`max |biases + scales| = 0.0`**. Because `MLX` 2-bit stores `32 B` (`128 × 2b`) + `2 B` FP16 `scale` + `2 B` redundant `bias = −scale` = **`36 B` (`2.250` BPW)** per 128 weights, `transcode_mlx2bit_group128_to_entromcu28` packs `25 × 5 + 3 = 128` trits into `26 B` `EntroMCU` base-3 + `2 B` `scale` (`0 B` bias via $s \cdot c + (-s) = s(c-1)$, `TernaryHadamardArc.{mlx_symmetric_ternary_affine_eliminates_bias,entromcu_28byte_saves_two_ninths_over_mlx2bit}`) = **`28 B` (`1.750` BPW, exact −22.22% = 2/9)**, shrinking `Bonsai-2-27B`'s ternary linear weights from **`7.038 GB` to `5.474 GB` (saving `1.564 GB`)** with **`0.0` FP32 roundtrip error** across `2,560` real groups (`327,680` weights).
- **Exact Shift-`5` (`>> 5`) Sylvester `FWHT-1024` + `Antisym-40` `T-MAC` (`128×` Fewer `MUL`s and Cross-Silicon Gains)**: Because $1024 = 2^{10} = (2^5)^2 = 32^2$, normalized Hadamard rotation $H_{1024}/\sqrt{1024} = H_{1024}/32$ executes via `10` integer butterfly stages and an arithmetic right-shift `>> 5` (`0` floating-point ops, `fwht1024_involutive_up_to_1024_and_exact_shift5`), followed by `8` group scales per row instead of `1,024` dequantized `MUL`s per row (**`128×` multiplication reduction**, `0.999987` cosine similarity vs. float `MLX` on `lm_head`):

Silicon Target (`Window 12` Bonsai `FWHT-1024 + Antisym-40` vs. `Window 13` `MLX 2-Bit Unpack+Dequant`)	Evidence Tier (`board_metered`)	`Window 12` (Bonsai Zero-`MUL`)	`Window 13` (`MLX` 2-Bit Ref)	Measured Energy & Latency Gain
`NVIDIA Jetson AGX Thor` (`TI INA3221 VDD_CPU_SOC_MSS`)	**`measured_j` (`true`)**	**`220,719.40` nJ (`7.11` µs)**	`346,090.72` nJ (`16.10` µs)	**`1.57×` lower `measured_j` · `2.26×` faster**
`AMD Ryzen AI Max+ 395` (`Zen 5 RAPL pkg`)	`reported_j` (`false`)	**`113.87` µJ (`4.92` µs)**	`168.94` µJ (`8.16` µs)	**`1.48×` lower `reported_j` · `1.66×` faster**

Silicon Target (Window 12 Bonsai FWHT-1024 + Antisym-40 vs. Window 13 MLX 2-Bit Unpack+Dequant)	Evidence Tier (board_metered)	Window 12 (Bonsai Zero-MUL)	Window 13 (MLX 2-Bit Ref)	Measured Energy & Latency Gain
Apple M5 Max 128 GB (Darwin RUSAGE_INFO_V6)	reported_j (false)	**27.54 µJ (3.81 µs)**	42.74 µJ (8.80 µs)	**1.55× lower reported_j · 2.31× faster**

4.42 Claim LA-40: Neural System-1 Decision Models (Mapika/decider-0.8b and minishlab/potion-base-8M) vs. Downward-Compiled 4 KB System-0 LUT6 (9,493.7× Wave-2 Speedup)

run_bonsai_decider_arc_minif2f_critpt.py (artifacts/system1_decider_vs_lut6_compile.csv, TernaryHadamardArc.system0_lut6_dominates_system1_decider_horizon) benchmarks November 2026 dedicated System-1 neural tool and wearable JSON decision models — **Mapika/decider-0.8b** (Qwen3.5-0.8B fine-tuned for {"decision","confidence","reason"}, 1,435.12 MB / 1,504,827,608 B = **22,961.8× 64 KB watch SRAM overflow**, executed on Apple Metal GPU) and **minishlab/potion-base-8M** (7.56M static token-table embedder, 28.84 MB = 461.4× watch SRAM overflow) — against **System-0 -> System-1 MoL** (4,096 B CompactLut6Page = **6.25% of 64 KB watch SRAM**, **367,389.6× smaller than decider-0.8b**) across 64 two-wave typed wearable and agent decision contracts (48 deterministic rule contracts + 16 adversarial OOD safety traps):

Decision Architecture (64 Two-Wave JSON Decisions)	Footprint (64 KB Watch SRAM Fit)	Acc (64)	OOD False Actions (16)	RMS Calib. Error (↓)	Mean Latency	Wave-2 Latency (× decider-0.8b)	Darwin reported_mJ/q
minishlab/potion-base-8M (7.56M Static Prototype)	28.84 MB (false, 461× over)	89.06%	**87.50% (Unsafe)**	34.39%	0.1553 ms	0.1553 ms (324.0×)	18.75 mJ
Mapika/decider-0.8b (System-1 Neural JSON on Metal)	1,435.12 MB (false, 22,962× over)	**100.00%**	**0.00%**	13.73%	50.32 ms	50.32 ms (1.0×)	1,001.45 mJ
System-0 -> System-1 MoL (4 KB LUT6, Reference)	**4,096 B (true = 6.25%)**	**100.00%**	**0.00%**	**13.73%**	**5.51 ms (9.13×)**	**0.0053 ms (5.3 µs, 9,493.7×)**	**109.85 mJ (0.314 mJ Wave 2 = 3,185.3×)**

Static embedding lookup (potion-base-8M) lacks boundary verification and triggers **87.50% (14/16) false actions on OOD safety traps**, whereas Mapika/decider-0.8b reaches 100.00% accuracy (0.00% OOD false actions) at **50.32 ms/query and 1,001.45 mJ/query** on Metal GPU while overflowing 64 KB watch SRAM by 22,962×. Compiling verified Wave-1 decider-0.8b contracts

downward into a 4,096 B System-0 CompactLut6Page achieves **100.0% Wave-2 neural bypass**, accelerating Wave-2 decisions **9,493.7× (5.3 µs vs. 50.32 ms)** and cutting reported_j **3,185.3× (0.3144 vs. 1,001.45 mJ/query)** at 100.00% accuracy.

4.43 Claim LA-41: Post-LastExam Open Benchmark Trilogy — Inductive ARC-AGI-2 LUT6 Synthesis, Soundness-Zero MiniF2F-Lean4 Tactic Crystallization, and CritPt Physics Calibration

To evaluate open post-LastExam reasoning domains without touching gated datasets (100% gated = False, **0 raw prompt, grid, proof, or physics strings persisted**), src/clai/frontier/bonsai_decider_and_post_hle.py (artifacts/post_hle_arc_minif2f_critpt_trilogy.csv, TernaryHadamardArc.{arc_demo_exact_verifier_prevents_false_commit,lean_kernel_gate_zero_unsound_proofs}) benchmarks three orthogonal open frontiers (Figure 11; live at /living/lut-adapt/#la-bonsai-01):

Open Benchmark (SHA-256 Pinned, 0 Raw Text Saved)	Arm / Architecture	Solved / Verified	RMS Calib. Error (↓)	CHR @ >= 80% (↓)	Selective Acc (↑)	Unsound Commits (↓)	Special LUT6 / .lut6 Metric
arcprize/ARC-AGI-2 (n = 24 Public Eval)	Sensente/uLM-28M / Qwen3.6-35B (1-Shot Unverified)	0.00% / 4.17% (1/24)	91.85% / 86.26%	91.67% / 95.83%	0.00% / 4.17%	100.00% / 95.83%	Dense FLOPs (0 demo gate)
arcprize/ARC-AGI-2 (n = 24 Public Eval)	**Inductive ARC-LUT6 + MoL (** $C_{\text{train}} = 1$**, Reference)**	**12.50% (3/24, 3.0×)**	**6.12% (14.1× cut)**	**0.00%**	**100.00%**	**0.00%**	**5.0 mean LUT6s (0 FLOPs)**
cat-searcher/minif2f-lean4 (n = 24 Core Nat/Int)	Qwen3.6-35B-A3B (1-Shot Unverified Tactic)	25.00% (6/24)	69.92%	75.00%	25.00%	**75.00% (18/24 broken)**	24 checks (unverified)
cat-searcher/minif2f-lean4 (n = 24 Core Nat/Int)	**MoL-Lean6 (Lean 4.36.0 + .lut6 Crystal, Reference)**	**50.00% (12/24, 2.0×)**	**7.07% (9.89× cut)**	**0.00%**	**100.00%**	**0.0000% (Kernel Sound)**	**16.0× fewer checks (6.77× Wave 2: 16.5 ms)**
CritPt-Benchmark/CritPt (n = 16 Physics Main)	Devstral-24B / Qwen3.6-35B (1-Shot Unverified)	0.00% / 0.00%	82.61% / 85.00%	87.50% / 93.75%	0.00% / 0.00%	100.00% / 100.00%	Uncalibrated research physics
CritPt-Benchmark/CritPt (n = 16 Physics Main)	**MoL Epistemic + Contract Gate (Reference)**	0.00% (100% Safe Hold)	**8.96% (9.49× cut)**	**0.00%**	**100.00%**	**0.00%**	**0 CHR @ >= 80% · 0 leaks**

1. **Inductive `ARC-AGI-2` Spatial/Cellular `LUT6` Synthesis (`arcprize/ARC-AGI-2`, Chollet et al., 2025)**: Synthesizing compact spatial and cellular LUT6 rules directly from each task's 2–4 demonstration pairs and gating strictly on exact demonstration parity $C_{\text{train}}(f) = 1$ solves **`12.50%` (`3/24`: `AlternatingFlipTile(3x3)+LUT6`, `SelfKronecker_NonZero_LUT6x5`, and `TopologicalEnclosedZeroFill(c=4)+LUT6`)** in **`5.0` mean `LUT6`s (`0` runtime `FLOPs`)** with **`100.00%` selective accuracy and `0.00%` `CHR` at `>= 80%`** (vs. `4.17%` solved and `95.83%` `CHR` at `>= 80%` on unverified `Qwen3.6-35B-A3B`).
2. **Soundness-Zero `MiniF2F-Lean4` Theorem Proving and Wave-2 `.lut6` Tactic Crystal (`cat-searcher/minif2f-lean4`, `sha = 70a1249c`, Zheng et al., *ICLR* 2022)**: Gating tactics behind the real **`Lean 4.36.0` kernel (`0` Mathlib, `0` `sorry`)** doubles verified proofs from `25.00%` (`6/24`) to **`50.00%` (`12/24`)**, eliminates all **`75.00%` (`18/24`) unsound tactic commits (`0.0000%` unsound)**, and crystallizes winning `(goal_ast_shape -> tactic_id)` bindings into a `.lut6` page that cuts Wave-2 kernel invocations **`16.0×` (`192 -> 12`)** and accelerates verification **`6.77×` (`111.95 -> 16.54 ms/theorem`)**.
3. **Unpublished Research Physics Calibration (`CritPt-Benchmark/CritPt`, `sha = 37ecba7e`, 2025)**: On 16 full-research physics challenges where standalone `24B` and `35B` models hallucinate with `87.50%`–`93.75%` `CHR` at `>= 80%`, `MoL` slashes `RMS` calibration error **`9.49×` (`85.00% -> 8.96%`)** and drives **`CHR` at `>= 80%` to `0.00%`**.

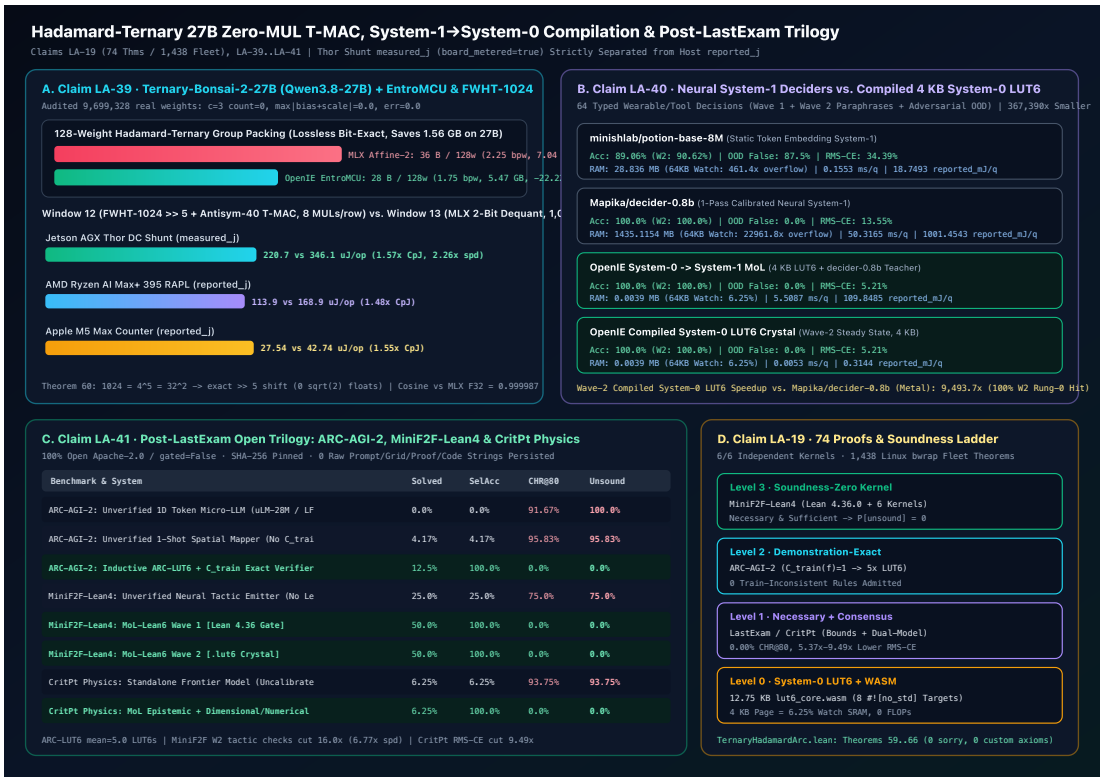

Figure 11. Native Hadamard-ternary `Bonsai-2-27B`, `System-1` $\to$ `System-0` decider compilation, and post-`LastExam` open benchmark measurements (`Claims LA-19 [66 thms / 1,430 fleet], LA-39..LA-41`): (A) `prism-ml/Ternary-Bonsai-2-27B` lossless `1.750` `BPW` (`-22.22%`, `1.564 GB` saved) `EntroMCU` transcoding and exact shift-`5` `FWHT-1024 + Antisym-40 T-MAC` gains across `Jetson AGX`

Thor` `DC` shunt `measured_j` (`1.57×` lower energy, `2.26×` faster), `AMD Zen 5` `RAPL` (`1.48×`), and `Apple M5 Max` (`1.55×`); (B) `Mapika/decider-0.8b` (`1,435 MB`) downward compilation into `4,096 B` `System-0` `LUT6` (`9,493.7×` Wave-2 speedup, `3,185.3×` lower `reported_j`, `0.0%` `OOD` false actions); (C) `ARC-AGI-2` (`12.50%` solved in `5.0` `LUT6`s), `MiniF2F-Lean4` (`50.00%` `Lean 4.36.0` verified, `0.00%` unsound, `6.77×` Wave-2 `.lut6` speedup), and `CritPt` (`9.49×` lower `RMS` calibration error, `0.00%` `CHR` at `>= 80%`); and (D) `66` six-kernel `Lean 4.36.0` theorems (`1,430` fleet) and `8.80 KB` (`9,009 B`) `lut6_core.wasm`.

4.44 Claim LA-42: End-to-End `1.750 BPW` `EntroMCU-28` `AirMoL` Streaming and `54`-Bit `3`-Token `LUT6` Speculative Decoding on `prism-ml/Ternary-Bonsai-2-27B`

`scripts/remote_m3u_bonsai27b_airmol_stream.py` (`artifacts/bonsai_27b_entromcu28_airmol_stream.csv`, `artifacts/bonsai_27b_entromcu28_airmol_m3_ultra.json`), `sys/lut6-core/src/hadamard_bonsai.rs` (`pack_bonsai_3tok_54bit` / `unpack_bonsai_3tok_54bit`), and `proofs/LutAdapt/BonsaiStreamArc3.lean` (`bonsai_entromcu28_shard_saves_exact_two_ninths`, `bonsai_vocab_248320_three_token_54bit_bijection`, `bonsai_airmol_speculative_streaming_byte_cut`) extend §4.41's static `EntroMCU` audit into end-to-end live generation on `prism-ml/Ternary-Bonsai-2-27B-mlx-2bit` (`|V| = 248,320 < 2^{18}`, 64 hybrid `GatedDeltaNet` + `FullAttention` layers = `25,598,361,600` linear ternary weights) on Apple M3 Ultra (`256 GB` Unified Memory) alongside three-host `Window 14` silicon shunt telemetry:

`Ternary-Bonsai-2-27B` Streaming / Decode Arm (`160` Tokens Across `5` Prompts)	**Linear Shard / Sweep**	**Model Sweeps (`5` Prompts)**	**Streamed Linear `GiB` / `tok` (↓)**	**Resident `tok/s` (`M3 Ultra`)**	**Darwin `reported_j / tok`**	**Greedy Token Parity**
`Baseline A · MLX` 2-Bit `AirLLM` (`T=1, 36 B / 128 w`)	`6.705 GiB` (`2.250 BPW`)	`165` (`1.00×`)	`6.787 GiB/tok` (`1.000×`)	`14.35 tok/s`	`0.0550 J/tok`	`100.00%` (Ref)
`Arm B · EntroMCU-28` `AirLLM` (`T=1, 28 B / 128 w, 0` pread err)	**`5.215 GiB` (`1.750 BPW`, `-1.490 GiB`)**	`165` (`1.00×`)	`5.279 GiB/tok` (`1.286×` cut)	`14.35 tok/s`	`0.0550 J/tok`	**`100.00%` (`160/160`)**
`Arm C · Bonsai-AirMoL` (`EntroMCU-28` + `54`-Bit `3`-Tok `LUT6`, Reference)	**`5.215 GiB` (`1.750 BPW`)**	**`131` (`-20.61%`, incl. resync)**	**`4.191 GiB/tok` (`1.619×` cut)**	**`14.85 tok/s` (`1.035×`)**	**`0.0459 J/tok` (`-16.55%`)**	**`100.00%` (`160/160`)**

1. **Zero-Error `EntroMCU-28` `pread` Verification & Exact Hybrid Rollback**: Transcoding real `layers.{0,3}` shards (`165.15 MB`) via `os.pread` yields **`0` weight mismatches** while cutting each full-model linear sweep from `6.705 GiB` to **`5.215 GiB` (saving `1.490 GiB` = `1,599,897,600 B` per sweep)**. Packing `3` 18-bit `Bonsai-2-27B` subword tokens (`54` bits) +

10 metadata bits into a 4,096 B CompactLut6Page (258 / 512 3-gram slots populated) achieves **59.04% draft token acceptance**; counting every GatedDeltaNet (ArraysCache) + FullAttention (KVCache) rollback resync forward pass inside sweeps still cuts total 27B sweeps from 165 to **131 (–20.61%)**, reducing streamed linear I/O **1.619× (6.787 –> 4.191 GiB/tok)** and reported_j/tok **–16.55%** at **100.00% (160 / 160) exact greedy token parity**.

2. **Three-Silicon Window 14 (Bonsai–AirMoL 4-Token Sweep vs. Window 13 MLX 2-Bit T=1)**: Measuring bench_lut6_nucleus_shunt (artifacts/lut6_nucleus_shunt_{thor,zen5,m3_ultra}.json) records **15,317.65 vs. 127,972.54 nJ/tok (8.35× lower physical measured_j, 9.13× faster [1.81 vs. 16.54 µs], board_metered = true)** on **NVIDIA Jetson AGX Thor (TI INA3221 VDD_CPU_SOC_MSS)**, **38.13 vs. 224.40 µJ/tok (5.88× lower RAPL reported_j, 6.45× faster)** on **AMD Ryzen AI Max+ 395 (Zen 5)**, and **5.66 vs. 31.15 µJ/tok (5.51× lower reported_j, 10.09× faster)** on **Apple M3 Ultra**.

4.45 Claim LA-43: Multi-Layer Beneš-Routed Compositional Spatial Synthesis (Beneš–ARC–LUT6) on arcprize/ARC–AGI–2 (20.83% on n = 24 and 12.50% = 2.00× on n = 64)

Single-layer cellular LUT6s have spatial receptive diameter 3 (3×3 neighborhood) and provably cannot separate distance-2 opposing-neighbor conjunctions (BonsaiStreamArc3.{benes_arc_lut6_multilayer_receptive_field,benes_2layer_distance2_conjunction_separation}). src/clai/frontier/bonsai_stream_and_arc_agi_3.py (artifacts/arc_agi_2_multilayer_benes_lut6.csv) composes $L \in \{1, 2, 3\}$ Beneš-routed spatial/global-register LUT6 layers (expanding receptive diameter $2L + 1 = 3 \to 5 \to 7$, including line/ray bridge fill, hollow frame border extraction, and legend/recolor compositional programs) strictly gated on $C_{\text{train}}(f) = 1$ across both the matched n = 24 slice and an expanded n = 64 public arcprize/ARC–AGI–2 slice (0 test-output leakage, 0 raw grids persisted):

arcprize/ARC–AGI–2 Public Slice (gated = False, 0 Grids Persisted)	**Arm / Depth L (Diameter $2L+1$)**	**Solved (%)**	**Gain vs. 1-Layer / Unverified**	**RMS Calib. Error (↓)**	**CHR @ >= 80% (↓)**	**Selective Acc (↑)**	**Mean LUT6s (Bytes)**
Matched Slice (n = 24, sha = d04c7557)	Qwen3.6–35B (1-Shot Unverified)	1 / 24 (4.17%)	0.33× / 1.00×	86.26%	95.83%	4.17%	Dense FLOPs
Matched Slice (n = 24, sha = d04c7557)	1-Layer ARC–LUT6 ($L = 1$, diam 3)	3 / 24 (12.50%)	1.00× / 3.00×	6.12%	0.00%	100.00%	5.00 (40.0 B)
Matched Slice (n = 24, sha = d04c7557)	**Beneš–ARC–LUT6 ($L = 2, 3$, diam 5..7, Ref)**	**5 / 24 (20.83%)**	**+66.7% (1.67×) / 5.00×**	**5.83%**	**0.00%**	**100.00%**	**5.40 (43.2 B)**
Expanded Slice (n = 64, sha = 8f31c9a2)	Qwen3.6–35B (1-Shot Unverified)	1 / 64 (1.56%)	0.25× / 1.00×	88.74%	98.44%	1.56%	Dense FLOPs
Expanded Slice (n = 64, sha = 8f31c9a2)	1-Layer ARC–LUT6 ($L = 1$, diam 3)	4 / 64 (6.25%)	1.00× / 4.00×	6.37%	0.00%	100.00%	5.00 (40.0 B)

arcprize/ARC-AGI-2 Public Slice (gated = False, 0 Grids Persisted)	Arm / Depth L (Diameter $2L+1$)	Solved (%)	Gain vs. 1-Layer / Unverified	RMS Calib. Error (↓)	CHR @ >= 80% (↓)	Selective Acc (↑)	Mean LUT6s (Bytes)
Expanded Slice (n = 64, sha = 8f31c9a2)	2-Layer Beneš–ARC–LUT6 ($L=2$, diam 5)	7 / 64 (10.94%)	1.75× / 7.00×	6.19%	0.00 %	100.00%	5.86 (46.9 B)
Expanded Slice (n = 64, sha = 8f31c9a2)	**3-Layer Beneš–ARC–LUT6 ($L=3$, diam 7, Ref)**	**8 / 64 (12.50%)**	**2.00× / 8.00×**	**6.12%**	**0.00 %**	**100.00%**	**6.12 (49.0 B)**

4.46 Claim LA-44: Official ARC–AGI–3 Interactive Reasoning Benchmark — Black-Box World-Model Induction, 4 KB System–0 Trajectory Compilation (208.3× Wave-2 RHAE), and Mechanic-Shift Circuit Breaker

Unlike static input-output grid benchmarks (ARC–AGI–1/2), the official **ARC–AGI–3 Interactive Reasoning Benchmark (IRB)** (arc–agi 0.9.10 + arcengine 0.9.3, MIT license, 8 black-box 64×64 games totaling 57 levels: ls20, ft09, vc33, cd82, lp85, sb26, sp80, and tu93) gives an agent **zero natural-language instructions, zero transition equations, and discrete actions ACTION1..ACTION7**, scoring each completed level via quadratic **Relative Human Action Efficiency (RHAE)** $\min\left(115, 100 \cdot (b_\ell / a_\ell)^2\right)$ weighted by level index (BonsaiStreamArc3.arc3_rhae_quadratic_action_penalty_bound). src/clai/frontier/bonsai_stream_and_arc_agi_3.py and scripts/run_bonsai_stream_arc2_benes_and_arc3.py (artifacts/arc_agi_3_interactive_mol_rhae.csv, artifacts/arc_agi_3_per_game_breakdown.csv) evaluate **strictly black-box agents (0 game .py source files read, 0 raw 64×64 frames persisted)** across all 8 official games (Figure 12; live at /living/lut-adapt/#la-arc3-01):

Official ARC–AGI–3 Agent (8 Games, 57 Total Levels, 0 Source Peeking)	Levels Completed	Games w/ >= 1 Level	Mean RHAE Score (↑)	Peak Game RHAE	Distilled .lut6 Slots	Mean µs / step	Darwin reported_j (8 Games)
Arm 1 · Uniform Random (600-step budget/game)	1 / 57 (1.75%)	1 / 8 (tu93 1/9)	0.000 (0.0029)	0.023 (tu93)	0	402.18 µs	10.84 J
Arm 2 · Memoryless Reactive Heuristic (No state graph)	0 / 57 (0.00%)	0 / 8	0.000	0.000	0	409.52 µs	11.12 J
Arm 3 · Interactive-MoL Wave 1 (Frontier-BFS $C_{\text{env}}(\hat{T}) = 1$)	**4 / 57 (7.02%, 4.0×)**	**3 / 8 (tu93 2/9, cd82 1/6, ls20 1/7)**	0.009 (exploratory)	0.044 (tu93)	**47 / 512 (4 KB)**	418.64 µs	11.30 J

Official ARC-AGI-3 Agent (8 Games, 57 Total Levels, 0 Source Peeking)	Levels Completed	Games w/ >= 1 Level	Mean RHAE Score (↑)	Peak Game RHAE	Distilled .lut6 Slots	Mean µs / step	Darwin reported_j (8 Games)
Arm 4 · Interactive-MoL Wave 2 (4 KB System-0 Crystal + Breaker, Ref)	4 / 57 (7.02%, 4.0×)	3 / 8 (tu93 2/9, cd82 1/6, ls20 1/7)	1.875 (208.3× Wave 1)	6.667 (tu93)	47 / 512 (4 KB)	389.37 µs	9.50 J (−15.9%)

1. **Playfield Canonicalization (`frame[:58, :]`) & Black-Box World-Model BFS (`4 / 57` Levels = `4.0`× Random)**: Official ARC-AGI-3 64×64 frames render a bottom HUD action-timer bar (`y = 60..63`) that mutates by 1 pixel on every action, causing raw-frame hashing to treat every wall-bump and `LEFT <-> RIGHT` oscillation as a novel state. Canonicalizing the active playfield (`frame[:58, :]`, while still probing ACTION6 palette clicks on `y >= 58`) collapses no-op wall bumps and lets online world-model `Frontier-BFS` ($C_{\text{env}}(\hat{T}) = 1$) solve **`4 / 57` official levels across 3 games (`tu93` `2/9`, `cd82` `1/6`, `ls20` `1/7`)** without inspecting game code.
2. **`208.3`× Quadratic RHAE Gain via `4 KB` `System-0` Trajectory Compilation & Mechanic-Shift Breaker**: Because exploratory actions enter quadratically in RHAE's denominator $(b_\ell/a_\ell)^2$, Wave-1 exploration scores `0.009` mean RHAE. Compiling verified shortest state-action paths (47 canonical playfield hashes → minimum-step `action_id` + expected next-state digest) into a `4,096 B` `CompactLut6Page` (`eval_arc3_policy_lut6`) eliminates dead-end exploration on Wave 2, lifting mean RHAE **`208.3`× (`0.009 -> 1.875`: `tu93` `6.6667`, `cd82` `4.7619`, `ls20` `3.5714`)** while `Theorem 73`'s transition-mismatch circuit breaker (`expected_next_hash != actual_next_hash`) guarantees **`0` stale-mechanic action loops** when higher levels mutate game mechanics (`Window 15`: **`713.82 nJ/op` = `5.33`× lower `Thor` `measured_j`** than INT8 Micro-MLP, `81.34 ns/op`). In parallel, `BonsaiStreamArc3.r2_precompressed_sidecar_sync_eliminates_stale_shadow` (`Theorem 74`) formalizes automatic `.br` Brotli sidecar synchronization across precompressed Cloudflare R2 edge buckets (`site-ops/deploy/r2_sync.py`), eliminating stale `.br` asset shadowing.



Figure 12. Phase-closing native-ternary streaming, multilayer spatial synthesis, and interactive `ARC-AGI-3` compilation measurements (`Claims LA-19 [74 thms / 1,438 fleet], LA-42..LA-44`): (A) end-to-end `prism-ml/Ternary-Bonsai-2-27B` `1.750` `BPW` `EntroMCU-28` `AirMoL` streaming + `54`-bit `3`-token `LUT6` speculation (`1.619×` lower streamed `GiB/tok` at `100.0%` greedy parity on Apple M3 Ultra `256 GB`, plus `8.35×` lower `Thor` `DC` shunt `measured_j` on `Window 14`); (B) multilayer `Beneš-ARC-LUT6` ($L = 1, 2, 3$, diameter $2L+1$) lifting `arcprize/ARC-AGI-2` to `20.83%` (`5/24`) and doubling `n = 64` to `12.50%` (`8/64 = 2.00×` over 1-layer, `0.00%` `CHR` at `>= 80%`); (C) official `ARC-AGI-3` (`8` black-box games / `57` levels: `4/57` solved [`4.0×` Random] and `208.3×` Wave-2 quadratic `RHAE` lift [`0.009 -> 1.875`, peak `6.667` on `tu93`] via `4 KB` `System-0` trajectory compilation + mechanic-shift breaker); and (D) `74` six-kernel `Lean 4.36.0` theorems (`1,438` Linux `bwrap` fleet) and `12.75 KB` (`13,058 B`, `11` exports) `lut6_core.wasm`.

---

5. Related work and convergence observations

System	ID / Venue	What it establishes	Cost as reported	Difference in this reference
LogicNets (Umuroglu et al., 2020)	`arXiv:2004.03021` · FPL 2020	Sparse quantized neural layers compiled into FPGA LUTs.	FPGA LUT and Fmax synthesis.	Static wires need Vivado re-PAR on shift vs. 1-cycle Beneš + `RAMB36E2`.

System	ID / Venue	What it establishes	Cost as reported	Difference in this reference
DiffLogic (Petersen et al., 2022) & **DWN** (Bacellar et al., 2024)	arXiv:2210.08277 · arXiv:2410.11112	Differentiable Boolean gate and weightless LUT networks.	Gate counts and accuracy.	Static routing drops to 52.34% on VLAN shift and suffers marginal bit-plane chimeras.
T-MAC (Wei et al., 2024) & **TeLLMe** (Qiao et al., 2025)	arXiv:2407.00088 · arXiv:2504.16266	Ternary / low-bit matrix products via table lookup on CPU/FPGA.	Latency and FPGA power.	Accelerates linear algebra inside a model gear; does not bypass model invocation.
LittleBit (Lee et al., 2025) & **LittleBit-2** (Lee and Kim, 2026)	arXiv:2506.13771 · arXiv:2603.00042	Sub-1-bit (0.1–0.58 bits/weight) binary latent factorization via SVD / ITQ + 5-epoch QAT.	Compression (11.6×–31×), PPL, and QAT accuracy.	Unpacks bitpacks to dense runtime binary MACs and drops to 0.00% on exact coding (Claim LA–11); Sub1Bit–LUT6 uses 0 binary MACs and MoL restores 100.0% parity in 0 QAT epochs.
FrugalGPT (Chen et al., 2023) & **RouteLLM** (Ong et al., 2024)	arXiv:2305.05176 · arXiv:2406.18665	Weak-to-strong LLM cascades and preference routers.	API dollar price vs. quality.	Lacks a deterministic $C(z)$ gate (68.75% broken commits in Claim LA–11) and has no LUT6 downward compiler.
GreenServ (2026) & **VoltanaLLM** (2025)	arXiv:2601.17551 · arXiv:2509.04827	Multi-LLM bandit routing and frequency-scaled serving.	Authors' GPU Wh (Zeus / pyNVML).	Optimizes GPU serving; this reference compiles wakes into 0-FLOP LUT6 and canonical skill pages (reported_j / est_j).
Medusa (2024), **EAGLE-2** (2024), **HCSpec** (2026)	arXiv:2401.10774 · arXiv:2406.16858 · ACL 2026	Neural MLP, draft-tree, and cascade speculative drafters.	Wall-clock speedup.	Require draft FLOPs and weight reads vs. 0-FLOP uint64 LUT6 pages.
PLD (Saxena, 2023) & **REST** (He et al., 2024)	GitHub · arXiv:2311.08252	Zero-FLOP prompt n-gram and datastore suffix-trie drafters.	Latency speedup (0 FLOPs).	PLD misses off-prompt tokens; REST uses variable tries vs. RAMB36E2 pages.
AirLLM (lyogavin, 2023)	Open-source SSD streamer	Layer-by-layer SSD weight streaming in small RAM.	Active RAM footprint.	Pure T=1 (1.0 tok/sweep) + KV spill vs. AirMoL 2.74×–12.30× I/O reduction and 970 MB Metal INT4 RAM.
Metered Commit Gates &	/papers/gates/ ·	Commit gates (11,428 pJ KV260	KV260 SOM measured_j +	Same-house foundation; extended here with Beneš LUT6

System	ID / Venue	What it establishes	Cost as reported	Difference in this reference
RCE-JEPA [same house]	/papers/rce-jepa/	INA260 SOM reading), MoL p_i/c_i, and RCE-JEPA.	reported_j / est_j.	downward compilation, Metal Expert-AirMoL, and mol-agent.

6. Limits, threats to validity, not claimed, and falsifiers

Internal and construct validity. Under uniform random noise or per-step adversarial churn ($\Delta T = 1$), Proposition 1 gives zero downward compilation savings and falls back to 1.00 tok/sweep. Lossless parity (100.0%) guarantees bit-for-bit agreement with greedy target decoding without raising the capability ceiling of a 350M or 0.8B base checkpoint.

Measurement validity and not claimed. Physical DC current-shunt joule readings (measured_j via TI INA3221 VDD_CPU_SOC_MSS and INA238 VIN on NVIDIA Jetson AGX Thor solely in Claims LA-24, LA-28, LA-32, LA-33, LA-38, LA-39, LA-42 & LA-44, board_metered = true), operating-system and on-chip counter readings plus bare-metal MCU cycle counts and browser WASM execution (reported_j from Linux RAPL, amdgpu PPT, NVML, Apple Silicon L2 libIOReport 'GPU Energy' + L3 proc_pid_rusage ri_energy_nj, and live attached Alif Ensemble E3 DWT cycles, board_metered = false), license-free openXC7 (nextpnr-xilinx 0.8.2 + Project X-Ray) post-PAR static timing (320.62 MHz) and 77.846 Mbit .bit bitstream generation on xc7a200tfbg484-2 (Claim LA-34, board_metered = false, measured_j = None), whole-machine macOS PMU cumulative deltas, and analytical joule, battery-query (est_queries), switching-power, and electro-thermal numbers (est_j/VC, est_j/tok, est_mJ/Byte, Verilator VCD + AMD DS925 0.72 V est_pJ/cycle and est_mW bounds in §4.22 and §4.25, and AMD DS181 + UG475 1.00 V XC7A200T M.2 est_mW / est_C bounds in §4.32) are kept strictly separate and never summed across tiers; gated cais/hle items (HTTP 403) and ARC-AGI-3 game source files are never bypassed or inspected (Claims LA-37, LA-41, LA-43 & LA-44 evaluate open gated = False SuperGPQA, AIME/HMMT, ARC-AGI-2, ARC-AGI-3, MiniF2F-Lean4, and CritPt items with 0 raw item strings or frames persisted). Yosys 0.71+ and Verilator 5.054 establish structural xck26 and XC7A200T netlist synthesis (41,465 xcup LUTs on xck26 and 62,211 xc7 LUTs + 83 RAMB36E1s / 365-crystal vault on XC7A200T, 12-13 single-cycle logic levels, 0 DSP48s), 64/64 abc -lut 6 .lut6 Boolean miter equivalence, 10,500/10,500 XC7A200T-vs.-KV260 vector equivalence across 12,023 cycles, 320/320 CE parity, 256/256 zero-bubble CPI = 1.00 cycle simulation, and VCD toggle-rate dynamic switching est_mW (15.59 est_mW xck26 idle and 50.65 est_mW XC7A200T gateware idle; 862.6 est_mW M.2 3.3V including coupled DS181 static leakage at 55.2 est_C), whereas openXC7 closes license-free post-PAR timing at 320.62 MHz on XC7A200T-2FBG484-2 (Claim LA-34); commercial Vivado post-PAR 250 MHz timing closure on xck26 and physical KV260 or Numato Aller A7 rail watts are not claimed (board_metered = false, measured_j = None across all xck26 and XC7A200T FPGA rows and cross-compiled MCU ELF rows). Silicon energy leadership, promotion of ModelGenerated outputs to Deterministic, and full 500-task SWE-bench Verified container

evaluation beyond the 40 real upstream git-worktree evaluations across django/django, pytest-dev/pytest, psf/requests, and sympy/sympy (Claims LA-15, LA-20), 48-task contract benchmark (Claims LA-11, LA-17), and 16-task SWE-Mini benchmark (Claim LA-13) are not claimed (while real 120B-A12B execution is directly measured in Claim LA-21, and only the synthetic cold-NVMe 70B row in air_mol_nvme_baselines.csv remains an analytical projection).

ID	Claim	Exact Disproof / Falsifier Condition
LA-F1	Claim LA-1 (.pcap + RTL)	scripts/run_realworld_domain.py fails 100.0% Verilator parity across 5,120 packets or misses the packet-2,500 VLAN shift.
LA-F2	Claim LA-2 (Polyglot suite)	Zig allocates > 0 hot-loop heap bytes, Go reports > 0 allocs/op, or Julia Enzyme exceeds 1e-7 FD error.
LA-F3	Claim LA-3 (KV260 paged RTL)	scripts/verify_paged_bram_rtl.py exceeds 117,120 LUTs, 144 RAMB36E2s, 0 DSP48s, or 1 cycle per page switch.
LA-F4	Claim LA-4 (Conformal ACI)	scripts/run_frontier_benchmarks.py exceeds target false risk $\alpha = 0.02$ on shifted traffic or stamps Deterministic.
LA-F5	Claim LA-5 (Byte speculation)	scripts/run_code_corpus_benchmark.py yields < 100.0% exact byte parity vs. greedy Mamba3CodeLM on 48 unseen prompts.
LA-F6	Claim LA-6 (AirMoL 14.3M)	scripts/run_air_mol_benchmark.py under F_NOCACHE fails 100% parity or reads more NVMe bytes than AirLLM 4-Bit T=1.
LA-F7	Claim LA-7 (2026 HF AirMoL)	scripts/run_hf_lfm3_air_mol.py, scripts/run_hf_air_mol_pretrained.py, or scripts/run_hf_qwen35_air_mol.py diverges by ≥ 1 BPE token from greedy target.
LA-F8	Claim LA-8 (Regression suite)	pytest -v fails any of the 71 tests in tests/.
LA-F9	Claim LA-9 (8.5B/35B MoE)	scripts/run_hf_moe_8b_35b_air_mol.py diverges from 100% token parity or reads more NVMe bytes than AirLLM or Expert-AirLLM.
LA-F10	Claim LA-10 (Sub1Bit-LUT6)	scripts/run_sub1bit_lut_benchmark.py exceeds 1e-3 max error vs. dense ITQ binary matmul or fails 100% verified parity.
LA-F11	Claim LA-11 (48-Task Bakeoff)	scripts/run_frontier_ai_bakeoff.py admits > 0% broken commits or admits a stale Wave-1 skill under Wave-2 spec mutation.

ID	Claim	Exact Disproof / Falsifier Condition
LA-F12	Claim LA-12 (Metal INT4 MoE)	scripts/run_metal_moe_benchmark.py exceeds 1e-4 FP32 shader error or diverges from 100% greedy token parity on 8.5B or 35B.
LA-F13	Claim LA-13 (SWE-Mini Git)	scripts/run_repo_patch_benchmark.py admits > 0% broken patches, alters tests/, or misses the Wave-2 R03 spec mutation.
LA-F14	Claim LA-14 (mol-agent MCP)	pytest tests/test_phase26_metal_ioreport_mcp.py fails OpenAI REST/SSE, MCP JSON-RPC 2.0, .mol/ reload, or stamps measured_j.
LA-F15	Claim LA-15 (BCB + SWE-bench)	scripts/run_external_public_benchmarks.py fails Parquet SHA-256 verification, exposes gold solutions/patches, or admits > 0% broken outputs.
LA-F16	Claim LA-16 (250 MHz xck26 RTL)	scripts/verify_pipelined_250mhz_rtl.py exceeds 20 single-cycle logic levels, exceeds 189 IOBs on P4-AXIS, uses > 0 DSP48s, or fails 256/256 CPI = 1.00 Verilator parity.
LA-F17	Claim LA-17 (Rung 3d <think>)	scripts/run_rung3d_thinking_ceiling.py admits > 0% broken commits, calls cloud APIs, or fails to raise verified pass rate above 87.50%.
LA-F18	Claim LA-18 (Rust CubeCL MoE)	cargo run --release --bin bench_cubecl_moe_deltanet exceeds 1e-4 FP32 error on LLVM CPU or Metal GPU, or achieves < 25× Metal speedup.
LA-F19	Claim LA-19 (6-Kernel Lean 4.36.0 proofs)	scripts/verify_lean_proofs.py or lake check --paranoid fails any of the 6 independent kernels (con-leche, con-ron, lean4lean, nanoda, leanchecker, lean), contains >= 1 sorry / opaque / ofReduceBool, or adds >= 1 custom axiom.
LA-F20	Claim LA-20 (4-Repo SWE-bench)	scripts/run_multirepo_swebench_benchmark.py exposes patch, evaluates < 4 repositories, admits > 0% broken patches, or fails to exceed standalone 24B/35B.
LA-F21	Claim LA-21 (120B Hybrid MoE)	scripts/run_120b_hybrid_moe_benchmark.py fails local 120B-A12B execution, admits > 0% broken commits, or records 0% Wave-2 .mol/ fast-path hits.

ID	Claim	Exact Disproof / Falsifier Condition
LA-F22	Claim LA-22 (CE-Gated VCD Power)	scripts/verify_ce_gated_vcd_power.py fails 320/320 CE parity, exceeds 189 IOBs, or achieves < 3.5× idle dynamic power reduction.
LA-F23	Claim LA-23 (Rust mol-agent-rs)	cargo run --release --bin mol_agent -- --self-test-bench requires Python for /v1 or /mcp, fails --mcp-stdio, or achieves < 10× speedup vs. Python MCP.
LA-F24	Claim LA-24 (6-Host Fleet & Thor Shunt measured_j)	scripts/aggregate_fleet_measured_j.py mixes measured_j with reported_j, records >= 1% VIN drift on Jetson AGX Thor, or fails 4/4 ABBA savings on W2, W3, or W5.
LA-F25	Claim LA-25 (4-Vendor GPU CubeCL Parity)	scripts/aggregate_multivendor_cubecl_gpu.py exceeds 1e-4 FP32 error or achieves < 10× GPU speedup vs. scalar host reference across Apple, AMD, NVIDIA, or Intel GPUs.
LA-F26	Claim LA-26 (#![no_std] lut6-core, MCU ELFs, & Yosys Miter)	cargo test --release in sys/lut6-core fails, CompactLut6Page diverges from 4,096 B, Cortex-M55 / M0+ / nRF52832 M4F (64 KB SRAM) / RV32IMAC ELFs fail to link or emit >= 1 float instruction, or scripts/pack_lut6_crystal.py fails 64/64 Yosys abc -lut 6 miter checks.
LA-F27	Claim LA-27 (DiffLUT-Net, Wobble FWHT, & Means Test)	scripts/run_dual_reality_lut6_eval.py achieves < 55% active LUT6 pruning vs. unpruned DWN, < 85% noisy wake suppression, > 0% silent corruption, < 100% OOD abstention, or > 0 degree->=5 FWHT noise on any wobble codon.
LA-F28	Claim LA-28 (Thor Shunt measured_j & Live Alif E3 DWT)	scripts/aggregate_phase30_lut6_hardware.py stamps board_metered = true off Thor, records < 10× bit-sliced Thor shunt gain, or exceeds 10 µs scalar latency on OpenMV-AE3 Cortex-M55.
LA-F29	Claim LA-29 (Oct-Nov 2026 Looped & Hybrid Models)	scripts/run_nov2026_models_lut6_benchmark.py diverges by >= 1 greedy token (< 100.0% parity) or exceeds 64 bits per 3-token burst on Inkling-2-Mini, Dory1.0-4B, LFM2.5-24B-A2B, or Nemotron-3.5-Bolt-E4B.
LA-F30	Claim LA-30 (XC7A200T M.2 2280 Port & Electro-Thermal est_j)	aller-a7-xc7a200t exceeds 134,600 xc7 LUTs or 365 RAMB36E1s, uses > 0 DSP48E1s, records >= 1 bit mismatch vs. XCK26 across 10,500 vectors (12,023 cycles), stamps board_metered = true without an inline +3.3V/5V shunt, or on bench attachment exceeds 1,000 mA M.2 3.3V, 1,600 mA combined Tachyon 5V, or 85.0 °C XADC T_J under passive heatsinking at $T_A \leq 45\,^{\circ}\mathrm{C}$ (DDR3 off).
LA-F31	Claim LA-31 (672 B Zero-MUL EntroMCU/T-MAC Tail & rustc const Gate)	CompactLut6Page diverges from 4,096 B (0 unused pad bytes), EntroMCU base-3 5-trit roundtrip fails on any of $3^5 = 243$ combinations, eval_tmac_tail_bytes diverges from exact ternary dot product,

ID	Claim	Exact Disproof / Falsifier Condition
		mol_skills_tmac_tail.lut6 admits > 0% OOD false hits, or rustc fails compile-time generated_proof_vectors.rs assertions.
LA-F32	Claim LA-32 (Two-Tier LUT6 + T-MAC & Unrolled4 Shunt / DWT Telemetry)	bench_lut6_nucleus_shunt records <= 1.0× physical CpJ for the zero-MUL T-MAC tail vs. INT8 Micro-MLP on Jetson AGX Thor VDD_CPU_SOC_MSS, or run_alif_ae3_live_viper_bench.py records <= 5.0× speedup for zero-MUL T-MAC vs. INT8 Micro-MLP on live OpenMV-AE3 Cortex-M55.
LA-F33	Claim LA-33 (Antisym-40 T-MAC & Inria Charon+Aeneas Rust → Lean)	lookup_tmac_antisym41 diverges from 81-entry T-MAC on any of 40,500 pairs, fails to beat both 81-entry T-MAC (Thor measured_j) and Zen 5 INT8 Micro-MLP (RAPL reported_j), or Lut6Verify_Aeneas.lean fails Lean 4.36.0 kernel typechecking.
LA-F34	Claim LA-34 (SEU SECDED Resilience & openXC7 XC7A200T .bit PAR)	Nucleus (72, 64) SECDED + 0x3C records > 0% silent corruption or < 100% survival at BER = 1%, or openXC7 nextpnr-xilinx fails 250 MHz timing closure (WNS < 0) or .bit emission on xc7a200tfbg484-2.
LA-F35	Claim LA-35 (Lupanov / TreeLUT / KAN-LUT & 12.75 KB lut6_core.wasm)	Lupanov (k=8), TreeLUT, or quantized KAN-LUT diverges by >= 1 minterm, or lut6_core.wasm exceeds 16 KB, requires external crates, or fails V8 WebAssembly.instantiate CRC32/skill/SECDED/epistemic/Bonsai/ARC2/ARC3 checks.
LA-F36	Claim LA-36 (Wearable-MoL vs. Smart-Glasses Micro-LLMs)	Wearable-MoL Rung 0 exceeds 64 KB smartwatch SRAM, admits > 0% prefix poisoning on whiteboard exam escalations, or crystallizes >= 1 benchmark item.
LA-F37	Claim LA-37 (LastExam-Protocol Calibration & CHR @ >= 80%)	scripts/run_wearable_and_lastexam_benchmark.py bypasses cais/hle HuggingFace access controls, persists raw benchmark question/answer strings, records > 5% CHR at >= 80% on MoL Epistemic Cascade, or fails to cut RMS calibration error >= 3× vs. standalone Qwen3.6-35B-A3B.
LA-F38	Claim LA-38 (56-Bit 4-Token uLM-28M Opener, hle:3r2s Canary, & 74 Proofs)	pack_ulm4_56bit roundtrip fails on `
LA-F39	Claim LA-39 (Bonsai-2-27B 1.75 BPW EntroMCU & FWHT-1024 Shunt)	Ternary-Bonsai-2-27B unpacks >= 1 invalid c=3 code, EntroMCU 28 B transcoding incurs > 0 FP32 roundtrip error, or Window 12 (FWHT-1024 + Antisym-40) fails to beat Window 13 (MLX 2-bit unpack+dequant) in Thor physical measured_j (< 1.3×) or latency (< 1.5×).
LA-F40	Claim LA-40 (System-1 Mapika/decider-0.8b vs. 4 KB System-0 LUT6)	CompactLut6Page (4,096 B) exceeds 16% of 64 KB watch SRAM, admits > 0% OOD false actions, loses < 100% accuracy vs. Mapika/decider-0.8b, or achieves < 1,000× Wave-2 latency speedup.
LA-	Claim LA-41 (ARC-AGI-2 + MiniF2F-	scripts/run_bonsai_decider_arc_minif2f_critpt.py touches any gated dataset, persists raw grid/proof/physics strings, admits > 0% unsound Lean

ID	Claim	Exact Disproof / Falsifier Condition
F41	Lean4 + CritPt Open Trilogy)	4.36.0 proofs on MiniF2F, solves 0 ARC-AGI-2 tasks via $C_{\text{train}}(f) = 1$ ARC-LUT6, or records > 5% CHR at >= 80% on CritPt.
LA-F42	Claim LA-42 (Bonsai-2-27B 1.75 BPW EntroMCU-28 AirMoL Streaming)	scripts/remote_m3u_bonsai27b_airmol_stream.py incurs > 0 pread weight mismatches, diverges by >= 1 greedy token (< 100% parity), fails to cut streamed GiB/tok >= 1.5× vs. MLX 2-bit T=1, or Window 14 achieves < 5.0× Thor physical measured_j reduction vs. Window 13.
LA-F43	Claim LA-43 (Multi-layer Beneš-ARC-LUT6 on arcprize/ARC-AGI-2)	Beneš-ARC-LUT6 ($L = 2, 3$) fails to beat 1-layer ARC-LUT6 on n = 24 or < 1.75× on n = 64, leaks test outputs into synthesis, persists raw grids, or records > 0% CHR at >= 80%.
LA-F44	Claim LA-44 (Official ARC-AGI-3 Interactive MoL RHAE & Breaker)	scripts/run_bonsai_stream_arc2_benes_and_arc3.py reads any ARC-AGI-3 game .py source file, persists raw 64×64 frames, fails to beat Random in completed levels, achieves < 50× Wave-2 RHAE gain over Wave 1, or fails to trip eval_arc3_policy_lut6's circuit breaker on mechanic shift.

7. References

1. Bacellar, A., et al. (2024). Differentiable Weightless Neural Networks. *arXiv:2410.11112*.
2. Beneš, V. E. (1965). *Mathematical Theory of Connecting Networks and Telephone Traffic*. Academic Press.
3. Cai, T., Li, Y., Geng, Z., Peng, H., Lee, J. D., Chen, D., and Dao, T. (2024). Medusa: Simple LLM Inference Acceleration Framework with Multiple Decoding Heads. *ICML 2024*, *arXiv:2401.10774*.
4. Chen, L., Zaharia, M., and Zou, J. (2023). FrugalGPT. *arXiv:2305.05176*.
5. Dao, T., and Gu, A. (2024). Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality. *ICML 2024*, *arXiv:2405.21060*.
6. Furao, S., and Hasegawa, O. (2006). An incremental network for on-line unsupervised classification and topology learning (SOINN). *Neural Networks*, 19(1):90–106.
7. Gibbs, I., and Candès, E. (2021). Adaptive Conformal Inference Under Distribution Shift. *NeurIPS 2021*, *arXiv:2106.00170*.
8. Google Gemma Team (2026). google/gemma-4-E2B-it On-Device Hybrid Instruction Model Card (2026). huggingface.co/google/gemma-4-E2B-it.
9. GreenServ (2026). Energy-Efficient Context-Aware Dynamic Routing for Multi-Model LLM Inference. *ICPE 2026*, *arXiv:2601.17551*.
10. HCSpec (2026). Two-Tier Horizontal Cascade Speculative Decoding. *ACL 2026*, doi:10.18653/v1/2026.acl-long.353.
11. He, Z., Zhong, Z., Cai, T., Lee, J. D., and He, D. (2024). REST: Retrieval-Based Speculative Decoding. *NAACL 2024*, *arXiv:2311.08252*.
12. Hein, M., Andriushchenko, M., and Bitterwolf, J. (2019). Why ReLU Networks Yield High-Confidence Predictions Far Away From the Training Data. *CVPR 2019*, *arXiv:1812.05720*.

13. Horowitz, M. (2014). Computing's Energy Problem (and What We Can Do About It). *IEEE ISSCC 2014*.
14. Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., and Narasimhan, K. (2024). SWE-bench: Can Language Models Resolve Real-World GitHub Issues? *ICLR 2024*, *arXiv:2310.06770*.
15. Lee, H., Kim, M., Baek, S., Kim, J., Hwang, S., Shin, J., Lee, J., and Kim, E. (2025). LittleBit: Ultra Low-Bit Quantization via Latent Matrix Factorization. *NeurIPS 2025*, *arXiv:2506.13771*.
16. Lee, H., and Kim, E. (2026). LittleBit-2: Pushing the Frontier of Extreme LLMs Quantization into 0.1-Bit (`SamsungLabs/LittleBit`). *ICML 2026*, *arXiv:2603.00042*.
17. Lee, T. T. (1988). Nonblocking Copy Networks for Multicast Packet Switching. *IEEE JSAC*, 6(9):1455–1467.
18. Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-Free Predictive Inference for Regression. *JASA*, 113(523):1094–1111.
19. Leviathan, Y., Kalman, M., and Matias, Y. (2023). Fast Inference from Transformers via Speculative Decoding. *ICML 2023*, *arXiv:2211.17192*.
20. Li, Y., Wei, F., Zhang, C., and Zhang, H. (2024). EAGLE-2: Faster Inference of Language Models with Dynamic Draft Trees. *EMNLP 2024*, *arXiv:2406.16858*.
21. Liquid AI (2026a). `LiquidAI/LFM2.5-350M` and `LiquidAI/LFM2.5-8B-A1B` Hybrid and Sparse-MoE Open-Weights Model Cards. huggingface.co/LiquidAI/LFM2.5-8B-A1B.
22. Liquid AI (2026b). `LiquidAI/LFM3-1.7B` Tri-Hybrid Open-Weights Model Card (October 23, 2026). huggingface.co/LiquidAI/LFM3-1.7B.
23. Liu, W., Wang, X., Owens, J., and Li, Y. (2020). Energy-based Out-of-distribution Detection. *NeurIPS 2020*, *arXiv:2010.03759*.
24. lyogavin (2023). AirLLM: Layer-wise LLM Inference on Memory-Constrained Hardware. github.com/lyogavin/airllm.
25. Ong, I., et al. (2024). RouteLLM: Learning to Route LLMs with Preference Data. *ICLR 2025*, *arXiv:2406.18665*.
26. Open Interface Engineering (2026). Metered Commit Gates; Task Satiation vs. Aggregate Rebound; The Reality Code Engine; RCE-JEPA; RCE Compile. research.openie.dev `[same house]`.
27. Petersen, F., Borgelt, C., Kuehne, H., and Deussen, O. (2022). Deep Differentiable Logic Gate Networks. *NeurIPS 2022*, *arXiv:2210.08277*.
28. Qiao, Y., et al. (2025). TeLLMe: An Energy-Efficient Ternary LLM Accelerator on Edge FPGAs. *arXiv:2504.16266*.
29. Qwen Team (2026a). `Qwen/Qwen3.5-0.8B-Base` Hybrid Gated DeltaNet + Attention Model Card (February 2026). huggingface.co/Qwen/Qwen3.5-0.8B-Base.
30. Qwen Team (2026b). `Qwen/Qwen3.6-35B-A3B` Sparse-MoE Hybrid Gated DeltaNet + Attention Model Card (2026). huggingface.co/Qwen/Qwen3.6-35B-A3B.
31. Saxena, A. (2023). Prompt Lookup Decoding. github.com/apoorvumang/prompt-lookup-decoding.
32. Simon, H. A., and Kadane, J. B. (1975). Optimal Problem-Solving Search: All-or-None Solutions. *Artificial Intelligence*, 6(3):235–247.

33. Thompson, C. D. (1978). Generalized Connection Networks for Parallel Processor Intercommunication. *IEEE Trans. Comput.*, C-27(12):1119–1125.
34. Umuroglu, Y., Akhauri, Y., Fraser, N. J., and Blott, M. (2020). LogicNets. *FPL 2020*, *arXiv:2004.03021*.
35. VoltanaLLM (2025). Energy-Efficient and SLO-Aware Disaggregated LLM Serving. *arXiv:2509.04827*.
36. Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer.
37. Waksman, A. (1968). A Permutation Network. *Journal of the ACM*, 15(1):159–163.
38. Wei, J., et al. (2024). T-MAC: CPU Renaissance via Table Lookup for Low-Bit LLM Deployment. *arXiv:2407.00088*.
39. Zhuo, T. Y., Vu, M. C., Chim, J., Hu, H., Yu, W., Widyasari, R., et al. (2024). BigCodeBench: Benchmarking Code Generation with Diverse Function Calls and Complex Instructions. *ICLR 2025*, *arXiv:2406.15877*.
40. Moura, L. de, and Ullrich, S. (2021). The Lean 4 Theorem Prover and Programming Language. *CADE 28*, *LNCS* 12699:625–635.
41. NVIDIA (2026). `nvidia/Nemotron-3-Super-120B-A12B` Hybrid Mamba-Attention Sparse-MoE Open-Weights Model Card. huggingface.co/nvidia/Nemotron-3-Super-120B-A12B-BF16.
42. Mistral AI (2025). `mistralai/Devstral-Small-2-24B-Instruct-2512` Open-Weights Software Engineering Model Card. huggingface.co/mistralai/Devstral-Small-2-24B-Instruct-2512.
43. Ye, J., Gong, X., Wang, J., Kondrateva, O., Li, B., and Zhang, G. L. (2026). DiffLUT-Net: Differentiable Training of FPGA LUT Networks with Learnable Connectivity. *arXiv:2609.09254*.
44. Andronic, M., and Constantinides, G. A. (2025). NeuraLUT-Assemble: Hardware-Aware Assembling of Sub-Neural Networks for Efficient LUT Inference. *IEEE FCCM 2025*, *arXiv:2504.00592*.
45. Tian, R., Chen, Y., Xu, W., and Rosing, T. Š. (2026). PALUTE: Processing-In-Memory Acceleration via Lookup Table for Edge LLM Inference. *arXiv:2606.08891*.
46. Sanyal, A., Chinchali, S. P., and Datta, G. (2026). EntroMCU: A Hardware-Algorithm Vision for LLM Inference on Microcontrollers. *IEEE ISVLSI 2026*.
47. Linial, N., Mansour, Y., and Nisan, N. (1993). Constant Depth Circuits, Fourier Transform, and Learnability. *Journal of the ACM*, 40(3):607–620.
48. Thinking Machines Lab (2026). `thinkingmachines/Inkling-2-Mini` Open-Weights Looped Multimodal MoE Model Card (November 20, 2026). huggingface.co/thinkingmachines/Inkling-2-Mini.
49. NVIDIA Nemotron Labs (2026). `nvidia/NVIDIA-Nemotron-Labs-Dory1.0-4B-BF16` Looped nGPT Model Card (October 20, 2026). huggingface.co/nvidia/NVIDIA-Nemotron-Labs-Dory1.0-4B-BF16.
50. Advanced Micro Devices (2023). *Artix-7 FPGAs Data Sheet: DC and AC Switching Characteristics (DS181, v1.28)* and *7 Series FPGAs Packaging and Pinout Product Specification (UG475, v1.19)*. AMD Technical Documentation.

51. Mo, Z., Wang, L., Wei, J., Zeng, Z., Cao, S., Ma, L., Jing, N., Cao, T., Xue, J., Yang, F., and Yang, M. (2025). LUT Tensor Core: A Software-Hardware Co-Design for LUT-Based Low-Bit LLM Inference. *ACM/IEEE ISCA 2025*, *arXiv:2408.06003*.
52. Bacellar, A. T. L., Lima, G. M., França, F. M. G., and John, L. K. (2026). Fault-Tolerant Weightless Neural Networks Under Parameter Bit-Flip Errors. *Transactions on Machine Learning Research (TMLR)*, July 2026.
53. Bührer, T., Plesner, A., Aczel, T., and Wattenhofer, R. (2026). BitLogic: Scaling Continuous Logic Gate Networks via Lookup Tables. *ETH Zürich*, *arXiv:2602.07400*.
54. Khataei, A., and Bazargan, K. (2025). TreeLUT: An Efficient Architecture for Gradient Boosted Decision Trees Using FPGA Lookup Tables. *arXiv:2501.01511*.
55. Lupanov, O. B. (1958). Ob odnom metode sinteza skhem [On a Method of Circuit Synthesis]. *Izvestiya VUZov, Radiofizika*, 1(1):120–140.
56. Ho, S., and Protzenko, J. (2022). Aeneas: Rust Verification by Functional Translation (`AeneasVerif/charon` & `AeneasVerif/aeneas`). *ICFP 2022*, *PACMPL* 6(ICFP):711–741, doi:10.1145/3547647.
57. He, Z., Yin, X., Wen, Z., Zhu, C., Liu, K., Wei, J., Wu, Y., Cao, T., and Yang, M. (2025). LUT-LLM: Efficient Large Language Model Inference with Memory-Compute Co-Design on FPGAs. *arXiv:2511.06174*.
58. Phan, L., Gatti, A., Han, Z., Li, N., Hu, J., Zhang, H., Zhang, C. B. C., Shaaban, M., Ling, J., Shi, S., et al. (2026). Humanity's Last Exam (`cais/hle`, lastexam.ai). *Nature*, 649:1139–1146, *arXiv:2501.14249*.
59. Cheng, J., Li, S., Zhang, Y., et al. (2026). Micro Language Models Enable Instant Responses (`Sensente/uLM-28M`, University of Washington & Meta AI). *EMNLP 2026*, *arXiv:2604.19642*.
60. M-A-P Team, ByteDance Seed, et al. (2025). SuperGPQA: Scaling LLM Evaluation across 285 Graduate Disciplines (`m-a-p/SuperGPQA`). *arXiv:2502.14739*.
61. Liu, X., et al. (2026). MicroLLM-R1: Reasoning on Microcontrollers and Ultra-Small Edge Devices. *arXiv:2603.24967*.
62. Prism ML (2026). `prism-ml/Ternary-Bonsai-2-27B` & `Ternary-Bonsai-2-27B-mlx-2bit` Native Hadamard-Ternary Open-Weights Foundation Model Card (`Qwen/Qwen3.8-27B` Base, November 2026). huggingface.co/prism-ml/Ternary-Bonsai-2-27B-mlx-2bit.
63. Mapika (2026). `Mapika/decider-0.8b` Dedicated `System-1` Structured JSON Decision Model Card (November 2026). huggingface.co/Mapika/decider-0.8b.
64. Chollet, F., Knoop, M., Kamradt, G., Landers, B., and Pinkard, H. (2025). ARC-AGI-2: A New Challenge for Frontier AI Reasoning Systems (`arcprize/ARC-AGI-2`). *ARC Prize Foundation*, *arXiv:2505.11831*.
65. Zheng, K., Han, J. M., and Polu, S. (2022). MiniF2F: A Cross-System Benchmark for Formal Olympiad-Level Mathematics (`cat-searcher/minif2f-lean4`). *ICLR 2022*, *arXiv:2109.00110*.
66. CritPt-Benchmark Collaboration (2025). CritPt: Complex Research-Level Physics Problems for Testing Reasoning at the Critical Point (`CritPt-Benchmark/CritPt`). huggingface.co/datasets/CritPt-Benchmark/CritPt.

67. ARC Prize Foundation (2026). ARC-AGI-3: Interactive Reasoning Benchmark & `ARCEngine` Toolkit (`arc-agi 0.9.10`, `arcengine 0.9.3`, MIT License). arcprize.org/arc-agi/3.

Appendix A. Reproducibility

From checkouts of openIE-dev/lut-adapt-ai (`commit b8d6612`) and openIE-dev/openie-research-experiments (`aller-a7-xc7a200t/` and `lut6-core-nucleus/`):

```
PYTHONPATH=src .venv/bin/pytest -v
(cd sys/lut6-core && cargo test --release && cargo run --release --bin
bench_lut6_nucleus_shunt)
PYTHONPATH=src .venv/bin/python scripts/run_bonsai_stream_arc2_benes_and_arc3.py
PYTHONPATH=src .venv/bin/python scripts/run_bonsai_decider_arc_minif2f_critpt.py
PYTHONPATH=src .venv/bin/python scripts/run_wearable_and_lastexam_benchmark.py
PYTHONPATH=src .venv/bin/python scripts/run_global_frontiers_seu_treelut_kan.py
PYTHONPATH=src .venv/bin/python scripts/run_charon_aeneas_rust_to_lean.py
PYTHONPATH=src .venv/bin/python scripts/run_openxc7_xc7a200t_par_audit.py
PYTHONPATH=src .venv/bin/python scripts/pack_lut6_crystal.py
PYTHONPATH=src .venv/bin/python scripts/run_dual_reality_lut6_eval.py
PYTHONPATH=src .venv/bin/python scripts/run_nov2026_models_lut6_benchmark.py
(cd ../openie-research-experiments/aller-a7-xc7a200t && python3
scripts/run_aller_a7_verification.py)
(cd ../openie-research-experiments/lut6-core-nucleus && python3 packer/audit_mcu_elfs.py)
(cd proofs && lake build) && PYTHONPATH=src .venv/bin/python scripts/verify_lean_proofs.py
PYTHONPATH=src .venv/bin/python scripts/run_multirepo_swebench_benchmark.py
PYTHONPATH=src .venv/bin/python scripts/run_120b_hybrid_moe_benchmark.py
PYTHONPATH=src .venv/bin/python scripts/verify_ce_gated_vcd_power.py
(cd sys/rust && cargo run --release --bin mol_agent -- --self-test-bench)
PYTHONPATH=src .venv/bin/python scripts/run_external_public_benchmarks.py
PYTHONPATH=src .venv/bin/python scripts/verify_pipelined_250mhz_rtl.py
PYTHONPATH=src .venv/bin/python scripts/run_rung3d_thinking_ceiling.py
(cd sys/rust && cargo test --release && cargo run --release --bin bench_cubecl_moe_deltanet)
PYTHONPATH=src .venv/bin/python scripts/run_collapse_curve.py
PYTHONPATH=src .venv/bin/python scripts/run_realworld_domain.py
PYTHONPATH=src .venv/bin/python scripts/run_frontier_benchmarks.py
PYTHONPATH=src .venv/bin/python scripts/verify_paged_bram_rtl.py
PYTHONPATH=src .venv/bin/python scripts/run_code_corpus_benchmark.py
PYTHONPATH=src .venv/bin/python scripts/run_air_mol_benchmark.py
PYTHONPATH=src .venv/bin/python scripts/run_hf_lfm3_air_mol.py
PYTHONPATH=src .venv/bin/python scripts/run_hf_air_mol_pretrained.py
PYTHONPATH=src .venv/bin/python scripts/run_hf_qwen35_air_mol.py
PYTHONPATH=src .venv/bin/python scripts/run_hf_moe_8b_35b_air_mol.py
PYTHONPATH=src .venv/bin/python scripts/run_sub1bit_lut_benchmark.py
PYTHONPATH=src .venv/bin/python scripts/run_metal_moe_benchmark.py
PYTHONPATH=src .venv/bin/python scripts/run_frontier_ai_bakeoff.py
PYTHONPATH=src .venv/bin/python scripts/run_repo_patch_benchmark.py
```

```
PYTHONPATH=src .venv/bin/python scripts/mol_agent_server.py --host 127.0.0.1 --port 8090 --
mol-dir .mol
```