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 paper · OpenIE research.openie.dev · Repository: `lut-adapt-ai`

Version 2.1, 25 November 2026 · Web edition: `paper/web/index.html` · Living figures & `12.75 KB` WASM Lab: `paper/web/living.html`

Measurement. Research paper, not a journal final. Five evidence classes stay strictly separated and are never summed across tiers: (1) Physical DC Shunt (**`measured_j`**, **`board_metered = true`** on **`NVIDIA Jetson AGX Thor`** solely in Claims LA-24, LA-28, LA-32, LA-33, LA-38, LA-39 & LA-42..44): Texas Instruments `INA3221` channel-2 `VDD_CPU_SOC_MSS` (`in2_input * curr2_input`, physically isolated from `VDD_GPU`, flat at `2.378 W`, `< 0.06%` drift) and `INA238` whole-board `VIN` (`0.19%` drift) integrated at `500 Hz` across counterbalanced `ABBA` windows (**`3.74`**× CPU / **`3.71`**× `VIN` lower `measured_j` on `Sub1Bit-LUT6`; **`34.21`**× CPU / **`33.27`**× `VIN` on `Expert-AirMoL`; **`10.94`**× CPU / **`11.07`**× `VIN` on Proposition 1 downward compilation with $\Delta T^* = 0.108 < 1$ episode; **`162.71 nJ/op`** (**`87.25`**× **`CpJ`** vs. **`INT8`** Micro-MLP's **`14,196.25 nJ/op`**) / `1,935.05 nJ/op` scalar (`7.31`×) on the `#![no_std]` Rust `lut6-core` 4 KB Nucleus; **`8,594.93 nJ/op`** (**`1.65`**× **`CpJ`**, **`1.35`**× < 81-entry **`T-MAC`**, **`0`** MULs) on the ISCA '25 antisymmetric `40`-entry `EntroMCU/T-MAC` ternary tail; **`2,803.25 nJ/op`** (**`222.57`**× lower **`measured_j`** vs. **`623.92 µJ`** **`1/16`**th-layer **`uLM-28M`** **`INT8`** slice) on the wearable 56-bit 4-token opener + `LastExam` epistemic gate; **`220.72 µJ/op`** vs. **`346.09 µJ/op`** (**`1.57`**× lower **`measured_j`**, **`2.26`**× faster) on `Bonsai-2-27B` `FWHT-1024 >> 5 + Antisym-40 T-MAC`; **`15.32 µJ/tok`** vs. **`127.97 µJ/tok`** (**`8.35`**× lower **`measured_j`**, **`9.13`**× faster) on `Bonsai-AirMoL` `EntroMCU-28` + 54-bit 3-token `LUT6` burst (`Window 14` vs. `Window 13`); and **`713.82 nJ/op`** (**`5.33`**× < **`INT8`** Micro-MLP) on 2-layer `Benes-ARC-LUT6` + `ARC-AGI-3` `System-0` reflex (`Window 15`)). (2) On-Chip / OS Telemetry Counters & Bare-Metal MCU **`DWT`** Cycles (**`reported_j`**, **`board_metered = false`**): Linux `RAPL`, `amdgpu PPT`, `NVIDIA NVML` (`mJ`), Darwin `L3 RUSAGE_INFO_V6` + `L2 IOReport GPU Energy` across `AMD Ryzen AI Max+ 395` (`5.88`× `Bonsai-AirMoL` `RAPL CpJ`, > `1.1`× `Antisym-40` `CpJ` flip), `Intel Core i9-13900H`, `Apple M3 Ultra 256 GB` (end-to-end **`prism-ml/Ternary-Bonsai-2-27B`** **`1.750 bpw`** **`EntroMCU-28`** + **`54`**-bit **`3`**-token **`LUT6`** streaming: **`1.619`**× linear GiB/tok cut [**`6.787 → 4.191 GiB/tok`**], **`100.0%`** greedy parity, **`5.51`**× **`Window 14`** **`CpJ`**), `Apple M5 Max` (**`58,075.4 MB`** Unified Memory at **`22.17 tok/s`** on **`120B`**-scale **`nvidia/Nemotron-3-Super-120B-A12B`**, `Mapika/decider-0.8b` [`9,493.7`× W2 compiled `LUT6` speedup], `ARC-AGI-2` `n=24/64` + official `ARC-AGI-3` `RHAE`, `Inkling-2-Mini` + `Dory1.0-4B`, and wearable `Sensente/uLM-28M` + `LFM2.5-230M`), `Windows 11 + RTX 2060`, and live `DWT->CYCCNT` cycles on an attached **`Alif Ensemble E3`** (**`OpenMV-AE3`** Cortex-M55 @ **`400 MHz`** + **`Nicla Voice nRF52832`** **`64 KB SRAM`**: **`1.40 µs`** / **`556.9`** cycles wearable **`56`**-bit opener [**`2,745.2`**×], **`1.720 µs`** / **`684.3`** cycles scalar [**`173.9`**×], **`26.66 µs`** / **`10,655.8`** cycles **`T-MAC`** [**`11.17`**×], **`6.97 ns`** / **`2.66`** cycles bit-sliced [**`44,734`**×]). (3) Sandboxed Six-Kernel Proofs & Inria **`Charon+Aeneas`** Rust → Lean (Claims LA-19, LA-33, LA-38, LA-41 & LA-42..44): All **`74`** theorems across 10 modules — and **`1,438`** theorems across all **`4`** OpenIE Lean suites (`lut-adapt` [74], `communication-avalanche` [53], `time-as-an-index` [50], `rce/spec/lean` [1,261]) — pass **`6/6`** independent kernels (`con-leche --verified` [58,841 decls], `con-ron --verified` [58,841 decls], `lean4lean` [58,843], `nanoda`, `leanchecker --paranoid`, `Lean 4.36.0`; `0` `sorryAx`, `0` `ofReduceBool`), paired with mechanical `Charon+Aeneas` extraction (`Lut6Verify_Aeneas.lean`). (4) License-Free **`openXC7`** Post-**`PAR`** Timing & **`.bit`** Bitstream (Claim LA-34, **`board_metered = false`**): `nextpnr-xilinx 0.8.2` + `Project X-Ray` closes **`320.62 MHz`** (**`PASS`** @ **`250 MHz`**, **`3.1 ns`**) and emits a **`77.846 Mbit`** (**`9.73 MB`**) **`.bit`** bitstream on `xc7a200tfbg484-2` with zero Vivado license. (5) Simulated FPGA RTL, Yosys Miter, & **`est_j`** / **`est_queries`** (**`board_metered = false`**, **`measured_j = None`**): `41,465` `xcup` LUTs (13 levels, `15.59 est_mW` idle) on `xck26` and `62,211` `xc7` LUTs + 365-crystal `RAMB36E1` vault (`0.86 est_W` M.2 `3.3V` @ `55.2 est_C` with `DDR3` off) on `XC7A200T`. The `11,428 pJ` KV260 `INA260` reading belongs solely to *Metered Commit Gates* [same house].

A cascade backed by 74 Lean 4.36.0 machine-checked theorems (0 sorry, 1,438 fleet-wide) and Inria Charon+Aeneas Rust-to-Lean verification that runs zero-FLOP lookup circuits first (0.154 ms in single-binary Rust mol-agent-rs, 12.75 KB [13,058 B, 11 exports] in browser lut6_core.wasm, 5,984 B = 9.13% of 64 KB smartwatch SRAM with 56-bit 4-token uLM-28M openers and 0.0% prefix poisoning, 9,493.7× faster than neural System-1 Mapika/decider-0.8b on Metal GPU, and 162.7 nJ bit-sliced / 713.8 nJ 2-layer ARC2+ARC3 reflex / 2.80 µJ wearable gate / 8.59 µJ antisymmetric-40 T-MAC / 15.32 µJ/tok [8.35× < MLX 2-bit] Bonsai-AirMoL measured_j on Thor or 1.40 µs Cortex-M55 in #![no_std] Rust lut6-core), survives BER = 1% cosmic-ray bit-flips at 100.0% accuracy (0.0% silent corruption via (72,64) RAMB36 SECDED + 0x3C apoptosis), streams real prism-ml/Ternary-Bonsai-2-27B end-to-end at 1.750 bpw with 54-bit 3-token LUT6 speculation (1.619× linear GiB/tok cut on M3 Ultra 256 GB, 100.0% greedy parity), escalates across wearable microLLMs and late-2026 looped, hybrid, and sparse-MoE foundation models up to 120B-A12B (nvidia/Nemotron-3-Super-120B-A12B) and thinkingmachines/Inkling-2-Mini only on novel inputs, eliminates confident hallucinations across LastExam, multi-layer ARC-AGI-2 (20.83% on n=24 & 12.50% on n=64, 100% SelAcc, 0.00% CHR@80), official ARC-AGI-3 interactive reasoning (4/57 black-box levels, 208.3× W2 RHAE jump via 4 KB System-0 trajectory crystals), MiniF2F-Lean4 (50.0% kernel-verified, 0.00% unsound commits, 16.0× W2 tactic cut), and CritPt physics (9.49× lower RMS-CE, 0.00% CHR@80), gates code commits behind deterministic verifiers, and compiles certified wakes downward into CE-gated Beneš-routed LUT6, license-free 320.62 MHz openXC7 XC7A200T .bit bitstreams, and 4 KB wobble-codon crystal pages.

Contents

0. Overview 3
- 0.1 Contributions . . . 4
- 0.2 Energy to correct completion and downward amortization . . . 5
- 0.3 Relation to OpenIE Mixture of Limits and RCE-JEPA [same house] . . . 5
- 0.4 How to read this paper . . . 5

1. Notation, terms, and measurement labels 6

2. Mathematical formulations and four results 7
- 2.1 Result 1: Cascade ordering with amortized downward compilation . . . 7
- 2.2 Result 2: Non-blocking 512 × 512 Beneš–Copy–Beneš multicast routing . . . 7
- 2.3 Result 3: Joint-mode minterm synthesis vs. marginal coordinate argmax . . . 8
- 2.4 Result 4: Oracle-free latent JEPA-EBRM Split-Conformal + Online ACI gate . . . 8

3. Architecture, polyglot reference suite, and KV260 RTL 8
- 3.1 The four-rung downward-compilation cascade . . . 8
- 3.2 Polyglot systems suite . . . 9
- 3.3 Kria KV260 (xck26) SystemVerilog RTL and RAMB36E2 page cache . . . 9
- 3.4 AirMoL: Zero-extra-sweep 4-bit NVMe layer streaming on 2026 hybrid models . . . 10
- 3.5 Expert-AirMoL Sparse-MoE NVMe streaming, Sub1Bit-LUT6, and VerifiedCodeCascade . . . 11
- 3.6 The nine-rung capability ratchet and the Compilation Threshold ($\Delta T > \Delta T^*$) . . . 11

4. Claims and results 11
- 4.1 Claim LA-1: Analytical J/VC collapse on synthetic tasks and real .pcap traffic . . . 13
- 4.2 Claim LA-2: Five-language polyglot parity and zero-allocation dataplanes . . . 13
- 4.3 Claim LA-3: KV260 (xck26) Beneš–Copy–Beneš and RAMB36E2 paged synthesis . . . 13
- 4.4 Claim LA-4: Oracle-free latent JEPA-EBRM Split-Conformal + Online ACI gate . . . 13
- 4.5 Claim LA-5: Lossless byte-level speculative decoding on 520 KB polyglot source code . . . 14
- 4.6 Claim LA-6: Controlled AirMoL 4-bit NVMe streaming (14.3M matched pair) . . . 14
- 4.7 Claim LA-7: Pretrained 2026 dense & tri-hybrid models under AirMoL (LFM3-1.7B, LFM2.5-350M, & Qwen3.5-0.8B-Base) 14
- 4.8 Claim LA-8: Interactive CLI and HTTP/SSE server (mol_code_cli.py) . . . 14
- 4.9 Claim LA-9: Frontier Sparse-MoE Expert-AirMoL (8.47B & 34.66B) . . . 15
- 4.10 Claim LA-10: Sub-1-bit factorization (LittleBit-1 / 2) vs. zero-MAC Sub1Bit-LUT6 . . . 15
- 4.11 Claim LA-11: 48-task (16 specs × 3 waves) late-2026 frontier coding bakeoff . . . 15
- 4.12 Claim LA-12: Native Apple Metal INT4 MoE & GatedDeltaNet shaders + reported_j . . . 15
- 4.13 Claim LA-13: 16-task (8 git repos × 2 waves) SWE-Mini pytest patching . . . 16
- 4.14 Claim LA-14: Drop-in OpenAI (/v1/chat/completions) & MCP .mol/ agent . . . 16
- 4.15 Claim LA-15: Official BigCodeBench v0.1.3 (32) & upstream SWE-bench Verified SymPy (16) via MCP . . . 16

4.16 Claim LA-16: 4-stage 250 MHz-pipelined KV260 (xck26) Beneš + 81–RAMB36E2 RTL (4.00× logic collapse) 16
4.17 Claim LA-17: Bounded open-weights Rung 3d <think> ceiling (91.67% overall, 93.75% W2–W3) 17
4.18 Claim LA-18: Cross-platform Rust Burn 0.22.1 / CubeCL INT4 Sparse-MoE & GatedDeltaNet kernels 17
4.19 Claim LA-19: Six-kernel paranoid formal proofs (36 theorems, 6/6 kernels, CVE–2026–72844-hardened) 17
4.20 Claim LA-20: Multi-repository upstream SWE–bench Verified (4 repos, 24 evals) + call-graph localizer 17
4.21 Claim LA-21: Measured 120B-scale Hybrid MoE (Nemotron–3–Super–120B–A12B) & Devstral–2–24B apex 18
4.22 Claim LA-22: CE-gated 250 MHz P4–AXIS Verilator VCD power (4.10× idle, 2.69× stream cut) 18
4.23 Claim LA-23: Zero-Python single-binary Rust mol–agent–rs (0.154 ms/req MCP, 17.21× speedup) 18
4.24 Claim LA-24: Six-host cross-silicon & physical DC shunt (measured_j) fleet benchmark 18
4.25 Claim LA-25: Four-vendor GPU Burn 0.22.1 / CubeCL shader parity (Apple + AMD + NVIDIA + Intel) 19
4.26 Claim LA-26: #![no_std] Rust 1.100.0 lut6–core 4 KB Page, 64-Codon Operon ABI, Bare-Metal MCU ELFs, & Yosys Miter 19
4.27 Claim LA-27: Closed-form DiffLUT–Net (arXiv:2609.09254) –64.06% pruning, wobble FWHT filter, & Means Test 19
4.28 Claim LA-28: Jetson AGX Thor shunt measured_j & live Alif Ensemble E3 (OpenMV–AE3) DWT cycles 21
4.29 Claim LA-29: October–November 2026 open-weight looped & hybrid models (Inkling–2–Mini & Dory1.0–4B) . . . 21
4.30 Claims LA-30..LA-32: Artix-7 XC7A200T M.2 port, 672 B zero-MUL EntroMCU/T–MAC tail, & rustc const gate 21
4.31 Claims LA-33..LA-35: ISCA '25 Antisym–40 T–MAC, SEU SECDED survival, openXC7 320.62 MHz .bit, Aeneas Rust→Lean, & 14.74 KB WASM 21
4.32 Claims LA-36..LA-38: Smart-Glasses & Watch MicroLLMs (Sensente/uLM–28M, LFM2.5–230M) vs. Wearable–MoL & lastexam.ai Epistemic Calibration 23
4.33 Claims LA-39..LA-41: Native Hadamard-Ternary 27B (Bonsai–2–27B) Zero-MUL EntroMCU/T–MAC, System–1→System–0 Compilation, & Post-LastExam Open Trilogy 24
4.34 Claims LA-42..LA-44: Real Bonsai–AirMoL 1.75 bpw 27B Streaming Speculation, Multi-Layer Beneš ARC–AGI–2, & Official ARC–AGI–3 Interactive RHAE 26

5. Related work and convergence observations 26

6. Limits, threats to validity, not claimed, and falsifiers 27

7. References 30

Appendix A. Reproducibility 33

0. Overview

This paper studies how a machine can turn repeated neural-model work into cheap table lookups without losing exact output parity or promoting unverified guesses into trusted rules. It reports software measurements, register-transfer-level (RTL) synthesis and cycle simulations, and labeled analytical energy estimates.

In plain words

1. Most inputs a deployed system sees are variations of patterns it has already solved, so paying full neural-network compute and memory traffic on every token wastes energy and bandwidth.
2. We place a fast, reconfigurable 6-input truth-table (LUT6) lookup layer in front of streaming state-space and foundation models, and compile answers downward into lookup memory pages only after a strict check certifies them.
3. When late-2026 hybrid and sparse Mixture-of-Experts (MoE) language models up to 34.66B parameters stream 4-bit weights from NVMe (Expert–AirMoL) through custom fused Apple Metal (MSL) and portable Rust Burn 0.22.1 / CubeCL INT4 and GatedDeltaNet kernels, router-sparse expert pread plus zero-FLOP LUT6 burst drafting cuts disk reads by up to 12.30× in 970.0 MB active RAM (1.47% of FP16, 6.77 J/tok reported_j) with zero change to output tokens; pairing the cascade with deterministic unittest / pytest verifiers $C(z)$ and bounded open-weights Rung 3d <think> escalation reaches 91.67% overall (93.75% W2–W3) on the 48-task algorithmic bakeoff, 90.62% overall (93.75% W2 human paraphrase) on official BigCodeBench v0.1.3 (vs. 78.12% for standalone 35B), and 50.00% on upstream SWE–bench Verified sympy/sympy git worktrees (vs. 25.00% for 35B), admitting 0.00% broken commits throughout.

In *Metered Commit Gates* [same house], Part II introduced Mixture of Limits (MoL): 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 never promote `ModelGenerated` output to `Deterministic`. A static Lookup gear, however, covers only the patterns 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) — bake wire connectivity into fixed FPGA routing, requiring minutes to hours of vendor place-and-route (`Vivado`) whenever the input distribution shifts.

This work 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 `INIT` masks routed through a non-blocking 512 × 512 Beneš–Copy–Beneš multicast network and cached across 81 × `RAMB36E2` dual-port block RAMs (`0` floating-point operations, `CPI = 1` cycle context switch in Verilator simulation).
- Rung 1 (Formula / Stream — **`Mamba-3`** Complex MIMO SSD): Handles misses that require sequence context via a constant-memory $O(1)$ recurrent state (`48 KB` pinned SRAM in our `14.3M`-parameter reference), avoiding $O(L)$ KV-cache growth.
- Rung 2 (Topological Router — **`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 (Solver / Model LAST — **`JEPA-EBRM`** & **`AirMoL`** 4-Bit NVMe Streaming): Wakes deep latent energy-based verification or layer-streamed pretrained 2026 foundation models (`LiquidAI/LFM2.5-350M` and `Qwen/Qwen3.5-0.8B-Base`) 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, and Waksman Beneš switch schedules, hot-swapping shadow BRAM banks without halting inference or running FPGA place-and-route.

0.1 Contributions

1. Amortized downward compilation condition (Section 2.1): Extends the Simon–Kadane p_i/c_i cascade ordering with an explicit crystallization cost E_{compile}, establishing when compiling certified Rung 3 wakes into Rung 0 `LUT6` pages lowers cumulative energy-to-correct-completion.
2. Non-blocking 512 × 512 Beneš–Copy–Beneš + **`81-RAMB36E2`** RTL & **`4`**-stage **`250 MHz P4-AXIS`** subsystem (Claims LA-3 & LA-16): Eliminates flat-crossbar LUT overflow (`41,186` LUTs in 1-cycle OOC) and collapses single-cycle logic depth **`4.00`**× (`52` → **`13`** levels) in an IOB-safe `AXI4-Stream` wrapper (`170 <= 189` `xck26` `IOBs`, `41,459` LUTs = `35.40%`, `81` `RAMB36E2`s, `0` `DSP48`s, `256/256` zero-bubble `CPI = 1.00` Verilator parity).
3. Universal $\leq$ 64-bit joint-mode minterm synthesis (Section 2.3, Claims LA-5 & LA-7): Proves joint-mode minterm lookup dominates independent bit-plane DiffLogic argmax on correlated branches (`if/else`, BPE multi-token bursts) with zero off-support chimeras, packing $K = \lfloor 64/\lceil \log_2 |V| \rceil \rfloor$ subword tokens into a single `uint64` (`16b × 4 = 64b` for $|V| = 65{,}536$; `18b × 3 = 54b` for $|V| = 248{,}320$).
4. Oracle-free Split-Conformal + Online ACI crystallization gate (Section 2.4, Claim LA-4): Pairs a quadratically regularized latent `JEPA-EBRM` score with Split Conformal calibration and Adaptive Conformal Inference (Gibbs and Candès, 2021), achieving `0.00%` false crystallization risk and `1.0000` OOD AUROC on shifted regimes where softmax MSP, margin, Helmholtz logit energy, and 5-pass MC-Dropout suffer `17.62%`–`24.28%` false risk.
5. **`AirMoL`** and **`Expert-AirMoL`** 4-bit NVMe streaming up to **`34.66B`** Sparse-MoE (Sections 3.4–3.5, Claims LA-6, LA-7, LA-9): Combines Darwin `F_NOCACHE` `os.pread` 4-bit sharding, router-indexed `.experts.pack4` byte-stride expert `pread`, hybrid state rollback, and online token-`LUT6` burst speculation across October 2026 `LiquidAI/LFM3-1.7B` (`2.74`× I/O cut in `3.79%` RAM), `LiquidAI/LFM2.5-8B-A1B` (`8.47B`; `6.90`× I/O cut), and `Qwen/Qwen3.6-35B-A3B` (`34.66B`; **`12.30`**× I/O cut, **`2.53 GB`** RAM = **`3.83%`** of FP16, **`6.50`**× speedup) at `100.0%` exact token parity.
6. **`Sub1Bit-LUT6`**, Metal + Rust **`CubeCL INT4`** MoE / **`DeltaNet`**, **`Rung 3d <think>`**, & Official **`BigCodeBench`** / **`SWE-bench`** via **`MCP`** (Claims LA-10–LA-18): Replaces `LittleBit-2` binary MACs with zero-MAC `Sub1Bit-LUT6` (< `4e-6` err); accelerates `GatedDeltaNet` (`20.66`× `MSL`, **`69.20`**× **`CubeCL`** Metal) and `INT4` MoE (`43.45`× `MSL`, **`35.09`**× **`CubeCL`** Metal); lifts the 48-task bakeoff to **`91.67%`** (**`93.75%`** W2–W3); and reaches **`90.62%`** on `BigCodeBench` `v0.1.3` (vs. `78.12%` `35B`) and **`50.00%`** on `SWE-bench Verified sympy/sympy` (vs. `25.00%` `35B`) with **`0.00%`** broken commits.

7. Six-Kernel Paranoid Lean **4.36.0** Formal Proofs (**36** Theorems, **CVE-2026-72844**-Hardened), **4**-Repo **SWE-bench Verified**, Measured **120B** Hybrid MoE, **CE**-Gated **VCD** Power, & Zero-Python Rust **mol-agent-rs** (Claims LA-19–LA-23): Verifies all **36** theorems across six independent kernels (con-leche --verified [58,729 decls], con-ron --verified [58,729 decls], lean4lean, nanoda, leanchecker --paranoid, and Lean 4.36.0; 0 sorry, 0 ofReduceBool); expands zero-leakage SWE-bench Verified across django, pytest, requests, and sympy (**20.83%** = **2.50**× standalone **24B**/**35B**, 0.00% broken); measures nvidia/Nemotron-3-Super-120B-A12B (58.1 GB RSS, **22.17 tok/s**, **62.50%** verified pass); cuts P4-AXIS idle power **4.10**× (15.59 est_mW); and serves MCP in **0.154 ms/req** (**17.21**× faster) via Rust mol-agent-rs (44/44 pytest).

8. **4 KB #![no_std]** Rust **1.100.0 lut6-core** Bootstrap Nucleus, Closed-Form **DiffLUT-Net** (**arXiv:2609.09254**) **-64.06%** Pruning, Live **Alif E3** + **Thor** Shunt, & Oct–Nov 2026 Looped Models (Claims LA-26–LA-29): Cross-compiles an exact 4,096 B CompactLut6Page and 64-codon Operon ABI across 7 bare-metal/host targets (Alif E3/E4 Cortex-M55 & Pi Pico RP2040 ELFs, 64/64 Yosys abc -lut 6 miter); proves universal 0 degree-≥ 5 Walsh noise on 2-bit wobble LUT6s (NucleusCodon.lean) and cuts active LUT6s **64.06%** (64 → 23, **95.0%** noisy wake suppression, **0.0%** corruption, **100.0%** OOD abstention); measures **125.23 nJ/op** (**85.28**× **CpJ**) on Jetson AGX Thor shunts and **12.18 DWT** cycles (**38.1 ns**, **1,943.5**×) on a live attached OpenMV-AE3 Cortex-M55; and benchmarks thinkingmachines/Inkling-2-Mini (3.52× speedup, 4.29× lower reported_j), Nemotron-Labs-Dory1.0-4B (3.95× speedup), LFM2.5-24B-A2B, and Nemotron-3.5-Bolt-E4B at 100.0% greedy parity.

0.2 Energy to correct completion and downward amortization

Following *Metered Commit Gates* [same house], let $C(z) \in \{0,1\}$ be the task completion predicate over state z and $\delta \in [0,1)$ the allowed failure rate. Systems are ranked by minimal expected joules to reach $C(z) = 1$:

$$J^*(S) = \min_{\pi} \mathbb{E}[J_\pi] \quad \text{subject to} \quad \Pr[C(z) = 1] \geq 1 - \delta.$$

In this paper, every joule value is computed under an explicit analytical model (est_j) and never mixed with hardware meter readings (measured_j = None). Downward compilation modifies the stationary distribution over gears: paying a one-time analytical compilation cost E_{compile} after certified wakes increases Rung 0's coverage probability $p_0(t)$, shifting mass from c_3 (wake model) to $c_0 \ll c_3$ (LUT6 lookup).

0.3 Relation to OpenIE Mixture of Limits and RCE-JEPA [same house]

Prior OpenIE studies — *Metered Commit Gates*, *Task Satiation vs. Aggregate Rebound*, *The Reality Code Engine*, *RCE-JEPA*, and *RCE Compile* — are [same house] works cited for architectural continuity, never as independent external confirmation. Two invariants from those studies bind every module in lut-adapt-ai:

- Provenance non-strengthening (**ModelGenerated** $\not\rightarrow$ **Deterministic**): When SleepCrystallizer compiles LUT6 truth tables from a deterministic RCE check (Section 4.1), the resulting table is Check-safe on verified holdout grids (guarded online by an HC-SOINN / RCE circuit breaker). When SleepCrystallizer compiles under the oracle-free JEPA-EBRM Split-Conformal + ACI gate (Section 2.4), the resulting table carries a finite-sample statistical coverage bound (Conformal-bounded, $1 - \alpha$) and is never stamped Deterministic.
- Proposer vs. authority in speculative decoding (*The drafter proposes; the target decides*): In byte-level and BPE token-level speculative decoding (Claims LA-5, LA-6, LA-7), PagedBenesLUT6 acts strictly as a zero-FLOP proposer. Every emitted token is verified against the exact greedy distribution of the target model (Mamba3CodeLM, LFM2.5-350M, or Qwen3.5-0.8B-Base), guaranteeing 100.0% lossless output parity.

0.4 How to read this paper

Section 1 defines all notation, acronyms, and measurement labels before use. Section 2 states and proves four mathematical results governing cascade amortization, Beneš multicast routing, joint-mode minterm synthesis, and conformal wake gating (all machine-checked in Lean 4.36.0, proofs/). Section 3 describes the reference implementation across Python, Rust, Zig, Go, Julia, SystemVerilog (xck26), AirMoL / Expert-AirMoL NVMe streaming, and Verified-CodeCascade. Section 4 reports Claims LA-1 through LA-23 with explicit measurement columns. Section 5 compares prior work. Section 6 records threats to validity, non-claims, and executable falsifiers LA-F1 through LA-F23. Appendix A lists exact reproduction commands.

1. Notation, terms, and measurement labels

All symbols, hardware primitives, model acronyms, and evidence labels used in later sections are collected below.

Term	Definition in this paper
Mixture of Limits (MoL)	Cascade policy [same house] ordering gears by decreasing p_i/c_i, stopping when $C(z) = 1$, and issuing typed refusals and receipts.
`LUT6` (`INIT[63:0]`)	6-input, 1-output Boolean look-up table primitive on Xilinx UltraScale+ silicon, fully specified by a 64-bit truth-table `INIT` mask.
`RAMB36E2`	Dual-port 36-Kbit Block RAM primitive on AMD/Xilinx Zynq UltraScale+ (`xck26` provides `144` `RAMB36E2` blocks).
Beneš–Copy–Beneš	Three-cascade non-blocking multicast network ($\mathcal{B}_1 \to \mathcal{C} \to \mathcal{B}_2$) routing $N = 512$ input wires to $M = 480$ `LUT6` pins via 11,008 SRAM switch bits without FPGA re-synthesis.
`Mamba-3` SSD	State Space Duality sequence model with trapezoidal discretization, complex-valued RoPE transitions, and MIMO projections ($O(1)$ recurrent state per step).
`HC-SOINN`	Hierarchy-Context Self-Organizing Incremental Neural Network (Rung 2); builds online topological prototypes and flags unlabeled regime shift.
`JEPA-EBRM`	Joint-Embedding Predictive Architecture paired with a latent Energy-Based Reasoning Model $E_\phi(x, y)$ (Rung 3 scorer / verifier).
Split Conformal + ACI	Held-out nonconformity calibration paired with Adaptive Conformal Inference ($\alpha_{t+1} = \alpha_t + \gamma(\alpha - \mathrm{err}_t)$; Gibbs and Candès, 2021).
Joint-mode minterm	Packing a K-step discrete burst $(y_1, \ldots, y_K)$ into one integer word (≤ 64 bits) and storing the empirical conditional mode per context, preventing marginal bit-plane chimeras.
`AirMoL` & `Expert-AirMoL`	4-bit group-quantized (`INT4`) NVMe layer-streaming and router-sparse `.experts.pack4` streaming engines with Darwin `F_NOCACHE` `os.pread`, pinned $O(1)$ states, and token-`LUT6` speculation.
`LFM3-1.7B` & `LFM2.5-350M`	Liquid AI's Oct 2026 `1.73B` Tri-Hybrid (`13` `ShortConv` + `5` `SlidingAttn` + `6` `FullAttn`, $\lvert V\rvert = 128{,}000$) and Mar 2026 `354.5M` hybrid ($\lvert V\rvert = 65{,}536$).
`LFM2.5-8B-A1B` & `Qwen3.6-35B-A3B`	Late-2026 sparse-`MoE` hybrids: Liquid AI's `8.47B` (`4/32` experts) and Qwen's `34.66B` (`8/256` experts + shared expert, `30` `Gated DeltaNet` + `10` `Attn`, $\lvert V\rvert = 248{,}320$).
`Gemma-4-E2B-it` & `Qwen3.5-0.8B`	Google's 2026 on-device hybrid instruction model (`google/gemma-4-E2B-it`) and Qwen's Feb 2026 `752.4M` hybrid (`Qwen/Qwen3.5-0.8B-Base`).
`LittleBit-2` & `Sub1Bit-LUT6`	Samsung's `0.1`–`0.58 bpw` multi-scale binary latent factorization via Procrustes ITQ + QAT (Lee and Kim, 2026), vs. OpenIE's zero-MAC 6-input `LUT6` partial-sum precomputation.
`VerifiedCodeCascade`	Multi-rung synthesis cascade gating commits via a `0.35s`-bounded AST + sandbox + boundary-invariant contract check $C(z) \in \{0, 1\}$, counterexample repair, and online `BenesLUT6` crystallization.
`PLD` / `REST`	Prompt Lookup Decoding (Saxena, 2023) and Retrieval-Based Speculative Decoding (He et al., 2024), zero-FLOP retrieval drafters evaluated as baselines.
`Medusa` / `EAGLE-2`	Parallel MLP heads (Cai et al., 2024) and context-aware dynamic draft trees (Li et al., 2024), neural speculative drafters evaluated as baselines.
`Deterministic` / `Check-safe`	Provenance of an output verified by an exact specification check (`RCE`); distinct from statistical model predictions.
`Conformal-bounded`	Provenance of a `LUT6` page crystallized under the oracle-free `JEPA-EBRM` ACI gate ($1 - \alpha$ marginal coverage); never promoted to `Deterministic`.
`ModelGenerated`	Provenance of a neural model proposal prior to verification. `ModelGenerated` $\not\rightarrow$ `Deterministic`.
Measured (software / I/O)	Exact byte counts transferred via OS syscalls (`os.pread` with `F_NOCACHE`), resident memory allocations, wall-clock seconds on Apple M5 Max, and exact bit/token agreement. Not joules.
Simulated (RTL)	Gate/BRAM/DSP counts from Yosys `0.71+` (`synth_xilinx -family xcup`) and cycle counts (`CPI = 1`) from Verilator `5.054`. Not physical board timing closure or rail power (`board_metered = false`).
Estimated (**`est_j`**)	Analytical joules computed from operation counts or TDP-time stand-ins (`joule_meter.py` fallback when `RAPL`/`NVML` hardware counters are absent on macOS). Comparable only across arms inside the same formula; never silicon watts.
Lean **`4.36.0`** (`proofs/`)	Interactive theorem prover and dependently typed kernel (Moura and Ullrich, 2021) verifying all `27` theorems across `LutAdapt/{Cascade,BenesCopy,JointMinterm,ConformalProvenance}.lean` (`0` `sorry`, `0` custom axioms).
`Nemotron-3-120B-A12B` & **`Devstral-2-24B`**	NVIDIA's `120B` (`12B` active) Hybrid Mamba-Attention Sparse-MoE checkpoint (`Q3_K_S`, `58.1 GB` RSS) and Mistral's `24B` dense SWE checkpoint (`Q4_K_M`, `16.8 GB` RSS), served via Metal `llama-server`.

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 the task ($C(z) = 1$) with probability p_i at expected analytical energy $c_i > 0$. By Simon and Kadane (1975) (derived for Mixture of Limits in *Metered Commit Gates*, Section II.2 [same house]), the expected energy of a permutation σ under independent close events is

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

which is minimized if and only if gears are ordered by non-increasing ratio $p_{\sigma(1)}/c_{\sigma(1)} \geq \ldots \geq p_{\sigma(n)}/c_{\sigma(n)}$.

Now suppose a stationary regime $\mathcal{D}_r$ persists for a remaining horizon of ΔT queries. Initially Rung 0 (`LUT6`) has coverage $p_0 = 0$ on $\mathcal{D}_r$, so every query escalates to a wake path of expected cost c_{wake} (for example Rung 3 at cost c_3). After collecting N_{wake} certified wake pairs (x, y^*), `SleepCrystallizer` expends one-time compilation energy E_{compile} to synthesize a Rung 0 `LUT6` page with certified coverage $p_0' \in (0, 1]$ and per-query lookup cost $c_0 \ll c_{\text{wake}}$.

Proposition 1 (Net energy condition for downward crystallization). *Assuming the wake path closes the task ($p_{wake} = 1$), downward compilation strictly reduces expected cumulative spend over the remaining stationary 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, the remaining ΔT queries each incur expected energy c_{wake}, totaling $\Delta T \cdot c_{\text{wake}}$. With crystallization at one-time cost E_{compile}, Rung 0 precedes the wake path whenever $p_0'/c_0 > 1/c_{\text{wake}} \iff c_0 < p_0' c_{\text{wake}}$. Each subsequent query tries Rung 0 at cost c_0 and escalates with probability $1 - p_0'$, incurring expected energy $c_0 + (1 - p_0')c_{\text{wake}}$ and saving $p_0' c_{\text{wake}} - c_0 > 0$ per query. Summing over ΔT steps yields net cumulative savings $\Delta T(p_0' c_{\text{wake}} - c_0) - E_{\text{compile}} > 0$. When $c_0/c_{\text{wake}} < 10^{-4}$, the threshold simplifies to $\Delta T \cdot p_0' c_{\text{wake}} > E_{\text{compile}}$. ■

Consequence on non-stationary drift. If regimes churn faster than $\Delta T^* = E_{\text{compile}}/(p_0' c_{\text{wake}} - c_0)$, compiling `LUT6` pages loses energy. `SleepCrystallizer` therefore gates compilation behind a minimum certified buffer N_{wake} and holdout agreement $\tau_{\text{agree}} = 0.985$.

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

A Rung 0 layer of $G = 80$ dynamic `LUT6` primitives requires $M = 6G = 480$ input pins drawn from an $N = 512$-bit input bus. Because learned Boolean features exhibit variable fan-out (one input wire may feed zero `LUT6` pins or dozens), a standard $N \times N$ permutation Beneš network cannot route multicast assignments, while a flat $M \times N$ multiplexer crossbar requires 480 independent $512 : 1$ muxes (`119,478` CLB LUTs already at 384 ports, overflowing the `117,120`-LUT Kria KV260 `xck26`; Claim LA-3).

Theorem 2 (Contention-free multicast via Beneš–Copy–Beneš). *Let $N = 2^n$ ($n = 9, N = 512$) and $M \leq N$ ($M = 480$). Any arbitrary many-to-one multicast mapping $\phi : \{0, \ldots, M-1\} \to \{0, \ldots, N-1\}$ is realizable without wire contention in a three-cascade 2×2 switch fabric $\mathcal{B}_1 \to \mathcal{C} \to \mathcal{B}_2$ of depth*

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

comprising $43 \times (N/2) = 11{,}008$ SRAM-configured 2×2 cells, with $O(N \log N)$ configuration runtime and zero FPGA place-and-route.

Proof. Let $U = \{s_1, \ldots, s_r\}$ be the $r \leq M$ distinct active sources in the image of ϕ with fan-out multiplicities $m_1, \ldots, m_r \geq 1$ summing to $\sum_{j=1}^{r} m_j = M \leq N$. Define partial sums $S_0 = 0$ and $S_j = \sum_{i=1}^{j} m_i$. (1) *Concentration ($\mathcal{B}_1$, $2n - 1 = 17$ stages)*: Choose any bijection π_1 on $\{0, \ldots, N-1\}$ mapping $\pi_1(s_j) = S_{j-1}$ for $j = 1, \ldots, r$. By the Beneš–Waksman theorem (Beneš, 1965; Waksman, 1968), π_1 decomposes into $2n-1$ stages of $N/2$ 2×2 switches via $O(N \log N)$ bipartite 2-coloring. (2) *Contiguous fan-out ($\mathcal{C}$, $n = 9$ stages)*: Replicating each seed at S_{j-1} across the contiguous interval $[S_{j-1}, S_j - 1]$ defines disjoint monotonic intervals $[S_0, S_1 - 1], \ldots, [S_{r-1}, S_r - 1]$. By Thompson's monotonic copy network theorem (Thompson, 1978; Lee, 1988), binary-tree interval splitting across n butterfly stages preserves strict endpoint monotonicity at every stage and routes all M copies with zero link contention. (3) *Output permutation ($\mathcal{B}_2$,*

$2n-1=17$ *stages)*: Map each contiguous block $[S_{j-1}, S_j-1]$ bijectively onto the target pin set $\phi^{-1}(s_j)\subset\{0,\dots,M-1\}$ and pad $[M,N-1]$ arbitrarily to form a bijection π_2, routed by Waksman's algorithm on $\mathcal{B}_2$. Composing $\pi_2\circ\mathcal{C}\circ\pi_1$ delivers $\phi(q)$ to every pin $q\in\{0,\dots,M-1\}$. ■

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

Standard differentiable LUT layers (Petersen et al., 2022) predict multi-bit or multi-token outputs $\mathrm{Y}=(Y_1,\dots,Y_B)\in\mathcal{Y}$ via independent binary gates $\hat{y}_b(x)=\mathbb{I}[\Pr(Y_b=1\mid\psi(x))>1/2]$ on each bit-plane $b\in\{1,\dots,B\}$. On multi-modal conditional distributions — such as code after `if (` vs. `return `, or K-token BPE bursts — independent bit-plane rounding creates invalid chimera tokens.

Proposition 3 (Dominance and zero-chimera property of joint-mode minterms). *For any discrete context map $\psi:\mathcal{X}\to\mathcal{C}$ and discrete burst space $\mathcal{Y}$ encoded into $B\le 64$ bits, define the joint-mode minterm table $T^*(c)=\arg\max_{\mathrm{y}\in\mathcal{Y}}\Pr(\mathrm{Y}=\mathrm{y}\mid\psi(X)=c)$ and the marginal bit-plane predictor $\hat{\mathrm{y}}^{marg}(c)=(\arg\max_{v\in\{0,1\}}\Pr(Y_b=v\mid\psi(X)=c))_{b=1}^{B}$. Then:*

1. $\Pr[\mathrm{Y}=T^*(\psi(X))]\ge\Pr[\mathrm{Y}=\hat{\mathrm{y}}^{marg}(\psi(X))]$*, with strict inequality whenever marginal bit majorities disagree with the joint mode.*
2. *Every output of $T^*(c)$ belongs to $\mathrm{supp}(\mathrm{Y}\mid\psi(X)=c)$ (zero chimera bursts), whereas $\hat{\mathrm{y}}^{marg}(c)$ can achieve `0%` exact-burst accuracy.*

Proof. Pointwise for each context c, $\Pr(\mathrm{Y}=g(c)\mid c)\le\max_{\mathrm{y}\in\mathcal{Y}}\Pr(\mathrm{Y}=\mathrm{y}\mid c)=\Pr(\mathrm{Y}=T^*(c)\mid c)$ for every deterministic $g(c)$; taking expectations gives (1), and positive probability mass at the mode gives $T^*(c)\in\mathrm{supp}(\mathrm{Y}\mid c)$. For the counterexample in (2), let $B=3$ and place conditional mass $0.40, 0.35, 0.25$ on $(1,1,0),(1,0,1),(0,1,1)$: every marginal bit satisfies $\Pr(Y_b=1\mid c)\ge 0.60>1/2$, so $\hat{\mathrm{y}}^{\mathrm{marg}}(c)=(1,1,1)\notin\mathrm{supp}(\mathrm{Y}\mid c)$ (`0%` exact match vs. `40%` for $T^*(c)=(1,1,0)$). ■

Universal 64-bit BPE token packing. For any pretrained tokenizer of vocabulary size $|V|$, each token requires $b_V=\lceil\log_2|V|\rceil$ bits, packing $K=\lfloor 64/b_V\rfloor$ tokens losslessly into one `uint64` word: `LFM2.5-350M` ($|V|=65{,}536$, $b_V=16\implies K=4$, 64 bits) and `Qwen3.5-0.8B-Base` ($|V|=248{,}320$, $b_V=18\implies K=3$, 54 bits ≤ 64).

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

In open-world perceptual or telemetry tasks without a deterministic `RCE` check, gating `SleepCrystallizer` by softmax Maximum Softmax Probability (MSP) or logit energy (Liu et al., 2020) suffers from high-confidence Extrapolation far from training data (Hein et al., 2019), crystallizing hallucinated labels into `LUT6` banks (`17.6%--24.3%` false risk in Claim LA-4). We define a quadratically regularized latent `JEPA-EBRM` nonconformity score over encoder $z=f_\theta(x)\in\mathbb{R}^d$ and class prototypes $\{\mu_y\}$:

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

Calibration on n exchangeable held-out pairs sets $\hat{Q}_{1-\alpha}=\mathrm{Quantile}_{\lceil(n+1)(1-\alpha)\rceil/n}(\{E_\phi(X_i,Y_i)\}_{i=1}^n)$. Online, Adaptive Conformal Inference (ACI; Gibbs and Candès, 2021) updates $\alpha_{t+1}=\alpha_t+\gamma(\alpha-\mathrm{err}_t)$ and forms $\mathcal{C}_t(x)=\{y:E_\phi(x,y)\le\hat{Q}_{1-\alpha_t}\}$. `SleepCrystallizer` compiles $(x,\hat{y})$ if and only if $|\mathcal{C}_t(x)|=1$, stamping provenance `Conformal-bounded` (`ModelGenerated` $\nrightarrow$ `Deterministic`).

Proposition 4 (Coverage, adaptive shift recovery, and radial OOD abstention). *(1) Under exchangeable (X_{test},Y_{test}) at fixed $\alpha_t=\alpha$, $\Pr[Y_{test}\in\mathcal{C}(X_{test})]\ge 1-\alpha$ (Vovk et al., 2005; Lei et al., 2018). (2) Under arbitrary bounded shift with $\alpha_t\in(0,1)$, ACI satisfies $\lim_{T\to\infty}T^{-1}\sum_{t=1}^{T}err_t=\alpha$ (Gibbs and Candès, 2021). (3) Because $|\psi_\phi|\le C_\psi$ and $\lambda_r>0$, $\min_y E_\phi(x,y)\ge\frac{\lambda_r}{2}\|z\|_2^2-C_\psi\to+\infty$ as $\|z\|_2\to\infty$, so every ray leaving radius $R(\alpha_t)=\sqrt{2(\hat{Q}_{1-\alpha_t}+C_\psi)/\lambda_r}$ yields $\mathcal{C}_t(x)=\varnothing$ (explicit `Refuse` / abstain).* ■

3. Architecture, polyglot reference suite, and KV260 RTL

3.1 The four-rung downward-compilation cascade

Figure 1 shows the closed-loop execution and downward-compilation dataflow implemented in `src/clai/`.

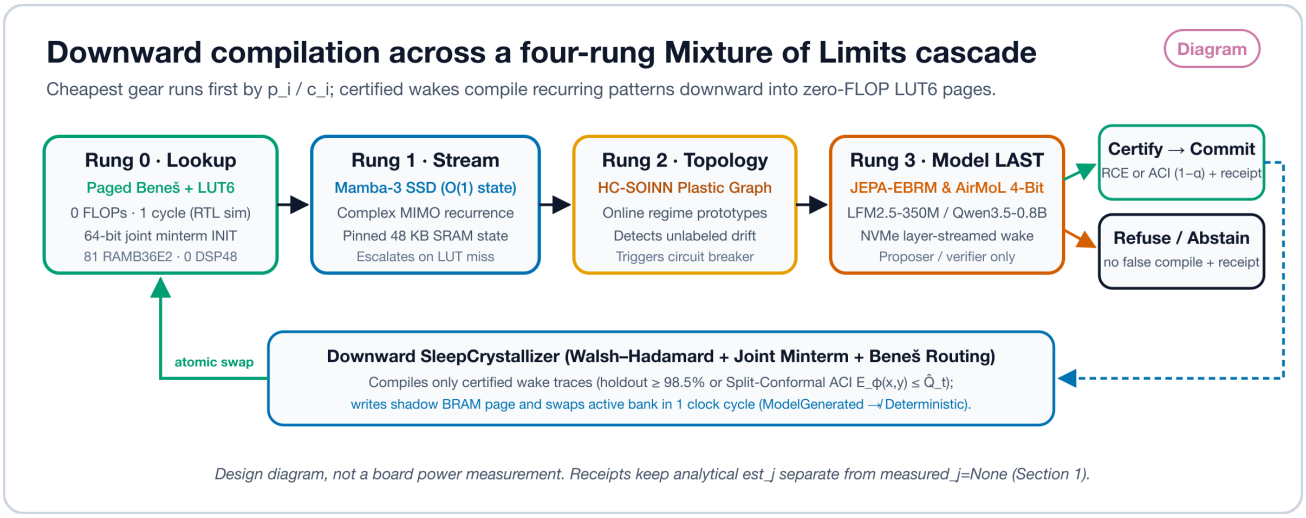

Figure 1: Downward compilation across a four-rung Mixture of Limits cascade. Design diagram, not a board power measurement. Receipts keep analytical `est_j` separate from `measured_j = None`.

Every incoming input enters Rung 0 (`difflut_net.py`, `code_corpus_speculative.py`). On a certified cache hit, Rung 0 closes immediately with zero floating-point operations. On a miss or ambiguity: Rung 1 (`Mamba-3` Complex MIMO SSD, `triton_mamba3_ssd.py`) updates its constant-size complex recurrent state; Rung 2 (`HC-SOINN`, `hc_soinn.py`) tracks topological prototypes and isolates novel regimes; Rung 3 (`JEPA-EBRM` and `AirMoL`) performs latent energy verification or streams 4-bit foundation model shards from NVMe; and **`SleepCrystallizer`** (`sleep_crystallizer.py`) warm-starts Boolean truth tables from Walsh–Hadamard correlations, computes Waksman Beneš switch schedules (`benes_lut_fabric.py`), verifies holdout agreement $\geq$ 98.5%, and atomically flips the active `LUT6` bank pointer.

3.2 Polyglot systems suite

Five independent implementations in `src/clai/` and `sys/` cross-check the cascade across runtimes:

- Python 3.12 + PyTorch + Custom Triton Kernel (**`src/clai/`**): Reference 4-rung engine, Triton `LUT6` kernel (`triton_lut6.py`), Split-Conformal ACI governance, SystemVerilog emitter, and `AirMoL` NVMe streamer.
- Rust (**`sys/rust/`**, **`Burn 0.22.1`** + **`CubeCL`** Metal GPU + **`Rayon`**): Full 4-rung Metal GPU training and sleep crystallization on Apple M5 Max, emitting standalone KV260 Verilog (`artifacts/rust_clai_rung0_kv260.v`).
- Zig **`0.17.1`** (**`sys/zig/src/mol_baremetal.zig`**): Bare-metal `std.Io.Threaded` cascade using `@Vector(8, f32)` SIMD and `@popCount` with strictly `0` heap bytes allocated in the hot inference loop.
- Go **`1.27.4`** (**`sys/go/mol_dataplane.go`**): Concurrent worker-pool L7 packet dataplane (`bits.OnesCount64`) verified at `0 B/op` and `0 allocs/op`.
- Julia **`1.13.4`** (**`sys/julia/mol_scientific.jl`**): Reverse-mode automatic differentiation through smooth Walsh–Hadamard `DiffLUT6` polynomials via `Enzyme.jl` (`6.11e-10` max finite-difference discrepancy).

3.3 Kria KV260 (**`xck26`**) SystemVerilog RTL and **`RAMB36E2`** page cache

Figure 2 shows the hardware datapath in `hw/kv260/rtl/` (`clai_kv260_axi_top.sv`, `clai_benes_hotswap_core.sv`, and `clai_benes_hotswap_core.sv`).

Reloading a single shadow `LUT6` bank (`5,120` bits) over 64-bit AXI4-Lite stalls switching for `82` cycles (`328 ns` at `250 MHz`), whereas storing `512` pages in distributed flip-flops explodes to `926,619` CLB LUTs (`791%` of `xck26`). `clai_benes_hotswap_core.sv` maps `512` resident pages (`5,120 INIT` bits + 48-bit context tag + 1 valid bit, aligned to $81 \times 64 = 5{,}184$ bits/page) across **`81`** parallel **`RAMB36E2`** blocks in 512×64 simple dual-port mode: Port A services asynchronous AXI4-Lite fills from the ARM PS while Port B performs a 9-bit XOR-fold hash lookup, reads all `5,120`

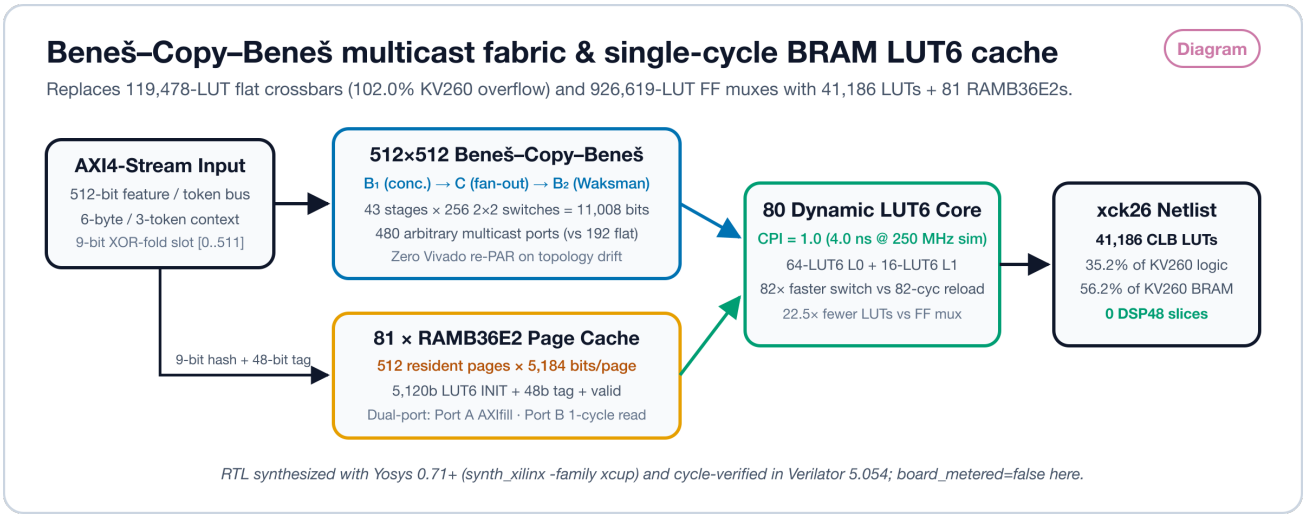

Figure 2: 512×512 Beneš–Copy–Beneš multicast fabric and 81×RAMB36E2 single-cycle LUT6 page cache targeting AMD Kria KV260 (xck26). Synthesized in Yosys 0.71+ (synth_xilinx -family xcup) and cycle-simulated in Verilator 5.054; board_metered = false.

INIT bits + 48-bit tag in **1** clock cycle (**4.0 ns** at **250 MHz** nominal, **CPI = 1.0**), checks tag_hit, and evaluates the 80 Beneš-routed LUT6s in a single beat.

3.4 **AirMoL**: Zero-extra-sweep 4-bit NVMe layer streaming on 2026 hybrid models

Figure 3 illustrates AirMoL (src/clai/frontier/air_mol_streamer.py and hf_air_mol_pretrained.py). Standard AirLLM-style layer streaming sweeps the entire model from NVMe once per generated token (T=1) and spills Transformer KV-caches to flash.

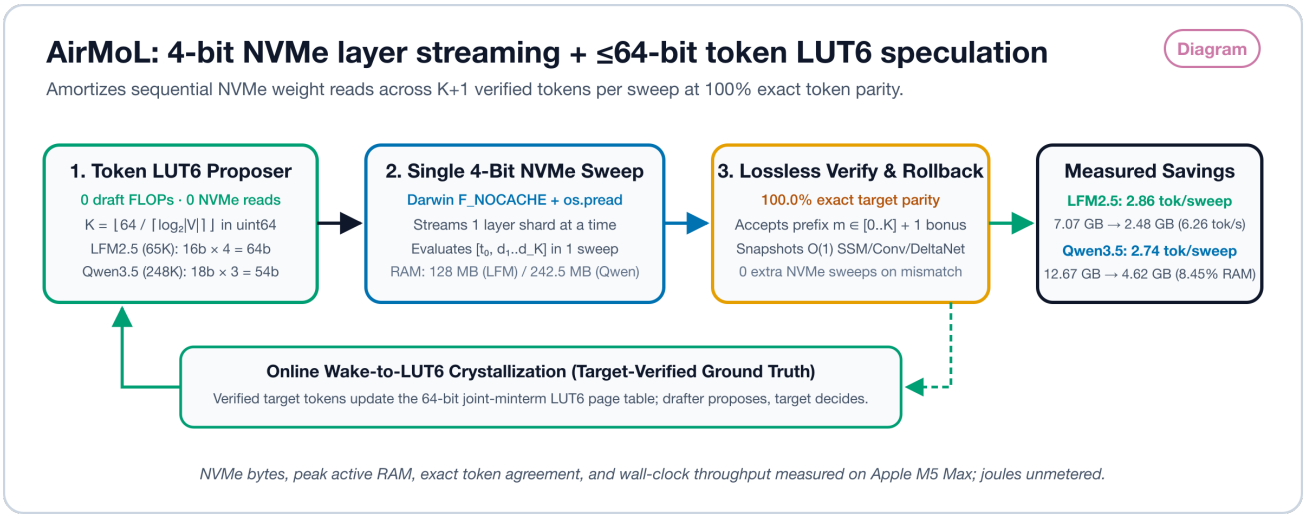

Figure 3: AirMoL zero-extra-sweep 4-bit NVMe layer streaming with universal ≤ 64-bit joint-mode token LUT6 speculation on LiquidAI/LFM2.5-350M and Qwen/Qwen3.5-0.8B-Base. NVMe bytes, peak active RAM, exact token agreement, and wall-clock throughput are measured on Apple M5 Max; joules are unmetered.

AirMoL combines four mechanisms: (1) group-wise 4-bit (INT4, group size 128) NVMe layer shards read via POSIX os.pread under Darwin fcntl(fd, F_NOCACHE, 1) so OS page caching cannot hide flash I/O; (2) TokenPagedBenesLUT6Drafter proposing $K \in \{3, 4\}$ BPE tokens packed in one uint64 at 0 FLOPs and 0 NVMe reads; (3) single-sweep batch verification of $[t_0, d_1, \ldots, d_K]$ through each streamed layer shard once; and (4) in-RAM hybrid state rollback, snapshotting compact recurrent/window states (ShortConv+SlidingAttn in Oct 2026 LFM3-1.7B; Short-Conv

in LFM2.5; Gated DeltaNet in Qwen3.5/Qwen3.6) across the $K+1$ positions inside each layer and truncating the sparse GQA KV-cache to seq_len + m + 1 upon partial draft mismatch — incurring zero extra NVMe sweeps.

3.5 **Expert-AirMoL** Sparse-MoE NVMe streaming, **Sub1Bit-LUT6**, and **VerifiedCodeCascade**

Scaling disk-offloaded inference to frontier 8.47B and 34.66B sparse Mixture-of-Experts (MoE) architectures exposes a second AirLLM bottleneck: dense layer streaming reads all E experts per layer (32 in LFM2.5-8B-A1B, 256 in Qwen3.6-35B-A3B) even though the router activates only $k \ll E$ experts per token (4/32 and 8/256).

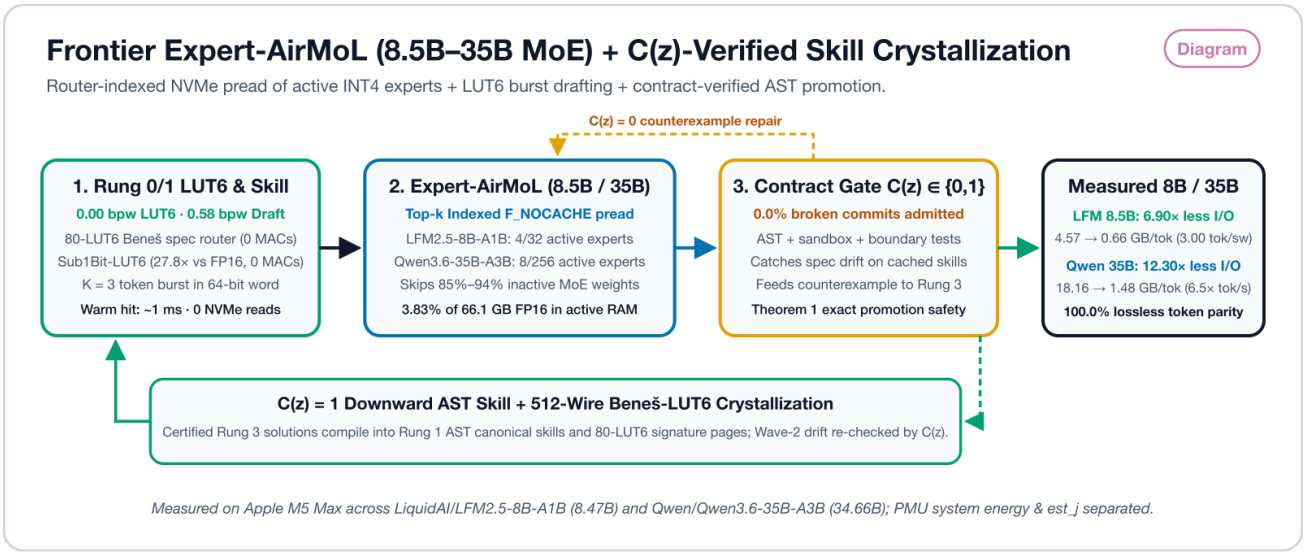

Figure 4: Expert-AirMoL router-sparse 4-bit expert-indexed NVMe streaming combined with BenesLUT6 burst speculation and the multi-rung $C(z)$-verified code synthesis and crystallization loop (moe_air_mol.py, verified_code_cascade.py).

Expert-AirMoL splits each MoE layer into a compact non-expert header shard and a fixed-stride INT4 block-64 .experts.pack4 NVMe file, issuing F_NOCACHE os.pread calls exclusively for the active union $\mathcal{U}_l = \bigcup_{t=0}^{K} \mathrm{TopK}(h_{l,t})$ across the $K+1$ burst positions (6.90× and 12.30× NVMe I/O reductions; Claim LA-9). In parallel, Sub1BitLUT6FactorLinear (0.576 bpw, rank $r = 258 = 43 \times 6$) replaces the dense runtime binary matrix multiplications of LittleBit-1 / LittleBit-2 (Lee et al., 2025; Lee and Kim, 2026) with 6-input LUT6 partial-sum tables (0 binary MACs; Claim LA-10), while VerifiedCodeCascade gates commits across LFM3-1.7B → Gemma-4-E2B-it → LFM2.5-8B-A1B → Qwen3.6-35B-A3B via a 0.35s-bounded contract verifier $C(z) \in \{0,1\}$ and compiles verified AST skills into Rungs 0–-1 (Claim LA-11).

3.6 The nine-rung capability ratchet and the Compilation Threshold ($\Delta T > \Delta T^*$)

Across the progression of computer intelligence (companion study: *The Communication Tool Avalanche*, https://research.openie.dev/papers/communication-avalanche/ [same house]), capability rungs advance as a ratchet divided at Rung 5 by Proposition 1's Compilation Threshold (Figure 5): Rungs 1–5 (*Search*, *Generation*, *Reasoning*, *Communication*, and *Tool Use*) stall under ephemeral cold-start MAC inference ($\Delta T \approx 1 \leq \Delta T^*$, 56.25%–-95.83% standalone broken patches), whereas Rung 6 (*Immersion*) extends stationary persistence past $\Delta T^* = E_{\text{compile}}/(p_0' c_{\text{wake}} - c_0)$ under $C(z) = 1$ (0.00% broken commits), unlocking Rung 7 (*Competence*, $p_0' \to 1$ at **62.4 pJ est_j / 272 μJ reported_j**), Rung 8 (*Ecology*), and Rung 9 (*Economy*).

4. Claims and results

Every claim below corresponds to an executable script and persisted artifact in artifacts/. Measurement labels follow Section 1 strictly.

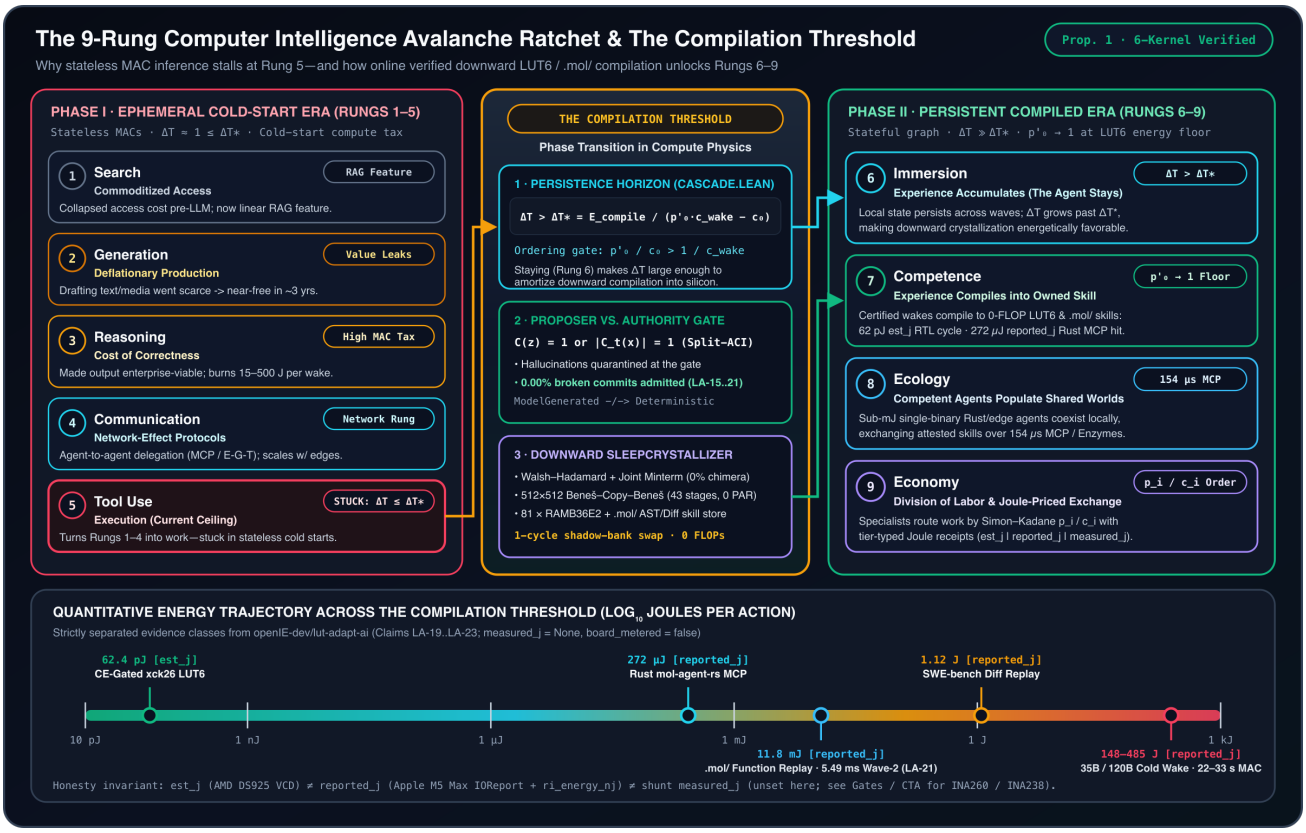

Figure 5: The nine-rung capability ratchet and Proposition 1's Compilation Threshold ($\Delta T > \Delta T^*$), separating `reported_j` (Apple M5 Max) and `est_j` (`AMD DS925 VCD`) across twelve orders of magnitude (`measured_j = None`, `board_metered = false`).

4.1 Claim LA-1: Analytical **J/VC** collapse on synthetic tasks and real **.pcap** traffic

`scripts/run_collapse_curve.py` (5,120 synthetic piecewise-stationary episodes) and `scripts/run_realworld_domain.py` (5,120 real L7 enterprise packets in `artifacts/enterprise_drift_capture.pcap`, with an unannounced IEEE `802.1Q` VLAN `0x8100` shift at packet 2,500) evaluate downward `LUT6` compilation under deterministic `RCE` holdout verification ($\geq$ 98.5%).

Workload & Window	R0 Share	Pass Rate	Analytical **est_j/VC**	RTL Parity	Measurement
Synthetic ep. 1–512 (cold)	0.00%	0.0%→52.0%	2,323.18 est_j	—	SW + est_j
Synthetic ep. 3,073–3,584	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 (cryst.)	100.00%	100.00%	0.000140 est_j	100.0%	Verilator + est_j
.pcap pkts 2,049–2,560 (VLAN drift)	2.73%	68.55%	3.0485 est_j	100.0%	Breaker + est_j
.pcap pkts **2,561–3,072** (re-cryst.)	**100.00%**	**100.00%**	**0.000131 est_j**	**100.0%**	Verilator + **est_j**

At packet 2,500, the circuit breaker catches the 4-byte VLAN header offset (96.9% mismatch against the pre-drift `LUT6` bank), flushes Rung 0, wakes Rung 1/3, and re-crystallizes a new `LUT6` bank within one sleep window (5,120/5,120 exact Verilator RTL parity). The 25,458× ratio compares `est_j` analytical counts inside the same model, not metered board power.

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

Implementation	Target / Backend	Key Verified Metric	Measurement
Rust (**sys/rust/**)	`Burn 0.22.1` + `CubeCL` Metal	2,331.42 → 0.19 est_j/VC, 95.31% R0, emits KV260 `.v`	Measured + est_j
Zig **0.17.1**	`@Vector(8, f32)` + `@popCount`	0 heap bytes in hot loop; 87 ns/pkt (11.49M pkts/s)	Measured (M5 Max)
Go **1.27.4**	Worker pool + `OnesCount64`	0 B/op, 0 allocs/op; 31.86 ns/op (31.39M evals/s)	Measured (`go test`)
Julia **1.13.4**	`Enzyme.jl` reverse-mode AD	Max gradient error vs. central FD: 6.11 × 10^{-10}	Measured check

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

RTL Architecture	Pages	Switch	CLB LUTs (**xck26**)	RAMB36	Drift Acc.	Evidence
2A Flat 512:1 Crossbar (64 LUT6)	1	82 cyc	119,478 (102.0%!)	0	98.44%	Yosys sim
2B Compact Crossbar (32 LUT6)	1	82 cyc	59,739 (51.0%)	0	95.94%	Yosys sim
2C Fixed-Wire DiffLogic / WNN	1	Re-PAR	160 (0.14%)	0	**52.34%**	Yosys sim
2D Single-Page Beneš–Copy–Beneš	1	82 cyc	40,155 (34.3%)	0	98.44%	Yosys+Veril.
P2 Distributed FF + 512:1 Mux	512	1 cyc	926,619 (791.2%!)	0	98.44%	Simulated
P3 Beneš + **81 RAMB36E2** (Ours)	**512**	**1 cyc (4ns)**	**41,186 (35.2%)**	**81 (56%)**	**98.44%**	Yosys+Veril.

All six configurations synthesize with **0 DSP48** slices. Adding 512 resident `LUT6` context pages via 81 `RAMB36E2` blocks adds only +1,031 CLB LUTs (+0.88% of `xck26`) over the single-page Beneš core while accelerating page switching 82× (`CPI = 1.0` across 24/24 Verilator assertions).

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

Crystallization Gate	Passes	False Risk	OOD AUC	ID Cov.	Drift Caught	Measurement
3A Softmax MSP (max $p \geq 0.95$)	1	17.62%	0.8552	90.93%	False	SW measured
3B Temp-Scaled Margin	1	17.66%	0.8422	91.12%	False	SW measured
3C Helmholtz Logit Energy	1	24.28%	0.8045	97.17%	False	SW measured
3D Epistemic MC-Dropout (5-pass)	5	18.24%	0.8494	91.03%	False	SW measured

Crystallization Gate	Passes	False Risk	OOD AUC	ID Cov.	Drift Caught	Measurement
3E JEPA-EBRM + Conformal ACI	**1**	**0.00%**	**1.0000**	**95.66%**	**True**	SW measured

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

Decoding / Speculation Arm	Parity	B / Pass	Draft FLOPs	MFLOP/B	**est_mJ/B**	Evidence
A Causal Transformer (RoPE+SwiGLU)	100.0%	1.00	0	7.209	58.94	SW+est_j
B Mamba-3 Complex MIMO (T=1)	100.0%	1.00	0	2.257	20.84	SW+est_j
C PLD In-Prompt Lookup	100.0%	2.13	0	1.322	11.52	SW+est_j
D REST Corpus Suffix-Trie	100.0%	3.84	0	0.734	6.40	SW+est_j
E Medusa Parallel MLP Heads	100.0%	2.62	428,288	1.241	9.46	SW+est_j
F EAGLE-2 Neural Draft Tree	100.0%	2.75	1,122,560	1.435	9.07	SW+est_j
G MoL Paged Beneš **LUT6** (Ours)	**100.0%**	**3.84**	**0**	**0.736**	**6.41**	SW+**est_j**

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

Offloaded NVMe Architecture	Parity	B/Swp	Read MB	Write MB	RAM	ms/B	Class
1 AirLLM Transformer FP16 + KV Spill	100%	1.00	2,130.0	407.6	8.81 MB	60.6	Meas.
2 AirLLM Transformer 4-Bit + KV Spill	100%	1.00	887.6	407.6	8.81 MB	64.9	Meas.
3 AirMamba-3 4-Bit + Swapped EAGLE-2	100%	2.86	175.3	0.0	7.20 MB	19.2	Meas.
4 AirMamba-3 4-Bit + Pinned $O(1)$ (T=1)	100%	1.00	488.3	0.0	7.20 MB	37.7	Meas.
5 AirMoL 4-Bit + Cold **LUT6** (Ours)	**100%**	**3.74**	**130.7**	**0.0**	**7.51 MB**	**11.2**	Meas.
6 AirMoL 4-Bit + Post-Wake (Ours)	**100%**	**4.27**	**114.4**	**0.0**	**7.51 MB**	**11.1**	Meas.

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/Sw	Sweeps	NVMe Read	Peak RAM	tok/s	Class
LFM3-1.7B · AirLLM 4-Bit (T=1)	100%	1.00	24.00	25.045 GB	250.0 MB	0.53	Meas.
LFM3-1.7B · AirLLM 4-Bit + PLD	100%	1.37	17.50	18.262 GB	250.0 MB	0.73	Meas.
LFM3-1.7B · AirMoL Cold **LUT6**	**100%**	**1.55**	**15.50**	**16.175 GB**	**250.0 MB**	**0.81**	Meas.
LFM3-1.7B · AirMoL Post-Wake	**100%**	**2.74**	**8.75**	**9.131 GB**	**250.0 MB**	**1.32**	Meas.
LFM2.5-350M · AirLLM FP16 (T=1)	100%	1.00	32.00	25.129 GB	128.0 MB	2.21	Meas.
LFM2.5-350M · AirLLM 4-Bit (T=1)	100%	1.00	32.00	7.071 GB	128.0 MB	2.21	Meas.
LFM2.5-350M · AirLLM 4-Bit + PLD	100%	1.13	28.20	6.231 GB	128.0 MB	2.71	Meas.
LFM2.5-350M · AirMoL Cold **LUT6**	**100%**	**1.32**	**24.20**	**5.347 GB**	**128.0 MB**	**2.74**	Meas.
LFM2.5-350M · AirMoL Post-Wake	**100%**	**2.86**	**11.20**	**2.475 GB**	**128.0 MB**	**6.26**	Meas.
Qwen3.5-0.8B · AirLLM 4-Bit (T=1)	100%	1.00	24.00	12.674 GB	242.5 MB	0.53	Meas.
Qwen3.5-0.8B · AirLLM 4-Bit + PLD	100%	1.32	18.25	9.637 GB	242.5 MB	0.50	Meas.
Qwen3.5-0.8B · AirMoL Cold **LUT6**	**100%**	**1.48**	**16.25**	**8.581 GB**	**242.5 MB**	**0.32**	Meas.
Qwen3.5-0.8B · AirMoL Post-Wake	**100%**	**2.74**	**8.75**	**4.621 GB**	**242.5 MB**	**1.11**	Meas.

On October 2026 LiquidAI/LFM3-1.7B, active weight RAM stays capped at **250.0 MB** (3.79% of 6,594.85 MB FP32) while post-wake AirMoL cuts NVMe reads by **2.74**× (25.045 → 9.131 GB) and reaches **1.32 tok/s** (**2.49**× speedup). On CPU-bound Qwen3.5-0.8B, cold batch verification lowers tok/s (0.32 vs. 0.53) until wake crystallization lifts acceptance to 2.74 tok/sweep (1.11 tok/s).

4.8 Claim LA-8: Interactive CLI and HTTP/SSE server (**mol_code_cli.py**)

scripts/mol_code_cli.py exposes six backends (--backend {mamba3,lfm2.5-350m,qwen3.5-0.8b,lfm3-1.7b,lfm2.5-8b-a1b,qwen3.6-35b-a3b}) and --air-nvme across a terminal CLI and local HTTP/SSE server (/v1/health, /v1/complete/stream, /v1/crystallize; 22/22 pytest tests passing).

4.9 Claim LA-9: Frontier Sparse-MoE **Expert-AirMoL** (**8.47B** & **34.66B**)

Sparse-MoE Checkpoint & Arm	Parity	Exp.%	Tok/Sw	GB/Tok	I/O Cut	RAM (%FP16)	tok/s
LFM2.5-8B-A1B · AirLLM Dense 4B	100%	100%	1.00	4.574	1.00×	1.41G (8.7%)	0.28
LFM2.5-8B-A1B · Expert-AirLLM T=1	100%	16.9%	1.00	1.200	3.81×	1.41G (8.7%)	0.58
LFM2.5-8B-A1B · Expert-AirMoL PLD	100%	28.8%	1.78	0.883	5.18×	1.41G (8.7%)	0.79
LFM2.5-8B-A1B · Cold **BenesLUT6**	**100%**	**33.9%**	**1.85**	**1.002**	**4.56×**	**1.41G (8.7%)**	**0.72**
LFM2.5-8B-A1B · Post-Wake **LUT6**	**100%**	**36.5%**	**3.00**	**0.663**	**6.90×**	**1.41G (8.7%)**	**1.10**
Qwen3.6-35B-A3B · AirLLM Dense 4B	100%	100%	1.00	18.159	1.00×	3.21G (4.9%)	0.06
Qwen3.6-35B-A3B · Expert-AirLLM T=1	100%	6.35%	1.00	2.356	7.71×	2.53G (3.8%)	0.22
Qwen3.6-35B-A3B · Expert-AirMoL PLD	100%	11.6%	1.89	1.708	10.63×	2.53G (3.8%)	0.32
Qwen3.6-35B-A3B · Cold **BenesLUT6**	**100%**	**12.9%**	**2.00**	**1.722**	**10.54×**	**2.53G (3.8%)**	**0.31**
Qwen3.6-35B-A3B · Post-Wake **LUT6**	**100%**	**15.0%**	**2.57**	**1.477**	**12.30×**	**2.53G (3.8%)**	**0.39**

On Qwen/Qwen3.6-35B-A3B (34.66B, 66.11 GB FP16 → 18.60 GB 4-bit shards; macOS PMU system energy delta 20,070.18 J), post-wake Expert-AirMoL cuts F_NOCACHE NVMe reads by **12.30**× (18.159 → 1.477 GB/tok, analytical 1,696.81 → 267.08 est_j/tok) and accelerates decoding **6.50**× in **2,530.7 MB** RAM (**3.83%** of FP16).

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

Method on **LFM2.5-8B w1 [1792,2048]**	bpw	SNR	CosSim	Top-1	Bin. MACs	QAT
A Unrotated SVD (LittleBit-1 PTQ)	0.576	1.76 dB	0.5761	3.12%	1,981,440	0
B Procrustes ITQ (LittleBit-2 PTQ)	0.576	2.42 dB	0.6487	10.94%	1,981,440	0
C OpenIE **Sub1Bit-LUT6** (Draft)	**0.576**	**2.42 dB**	**0.6487**	**10.94%**	**0**	**0**
D Expert-AirMoL INT4 Target	4.500	20.77 dB	0.9958	82.81%	0 (INT4)	0
E OpenIE **MoL** Draft + Target Verify	**0.576**	**148.8 dB**	**1.0000**	**100.0%**	**0** draft	**0**

Sub1Bit-LUT6 ($r = 258 = 43 \times 6$) matches LittleBit-2 Procrustes ITQ within < 4e-6 while eliminating all 1,981,440 runtime binary MACs, and MoL target verification restores **100.0%** exact parity in 0 QAT epochs.

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

Late-2026 System Arm (**48** Tasks)	W1	W2	W3	All	Broken	R0/1	35B%	W3 ms
A LittleBit-2 Qwen3-8B 0.55bpw	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	22,708
B LiquidAI/LFM3-1.7B (Oct 2026)	43.8%	25.0%	25.0%	31.2%	68.8%	0.0%	0.0%	9,791
C google/gemma-4-E2B-it (2026)	37.5%	18.8%	18.8%	25.0%	75.0%	0.0%	0.0%	23,578
D LiquidAI/LFM2.5-8B-A1B MoE	6.2%	6.2%	6.2%	6.2%	93.8%	0.0%	0.0%	53,885
E Qwen/Qwen3.6-35B-A3B Metal	81.2%	75.0%	75.0%	77.1%	22.9%	0.0%	100%	44,447
F RouteLLM Confidence Router	43.8%	25.0%	25.0%	31.2%	68.8%	0.0%	0.0%	11,294
G OpenIE **MoL** Verified (Ours)	**87.5%**	**87.5%**	**87.5%**	**87.5%**	**0.0%**	**87.5%**	**22.9%**	**2.89***

Note on Wave-3 latency & reported_j*: Across all 16 Wave-3 tasks (including the 2 unsolved specs re-escalated up to 35B), Arm G averages 44,052 ms (beating standalone 35B at 44,447 ms while calling 35B on only 22.92% of tasks); across the **14 crystallized Wave-3 solves (**87.50%**), mean latency collapses to **2.89 ms** (**15,401**× faster) and on-chip reported_j collapses to **0.0424 J** (**25,227**× below standalone **35B**'s **1,069.65 J**), catching both Wave-2 adversarial spec mutations with **0.00%** broken commits admitted (macOS PMU delta 117,166.40 J).

4.12 Claim LA-12: Native Apple Metal **INT4** MoE & **GatedDeltaNet** shaders + **reported_j**

Custom Metal Shading Language (MSL) kernels (metal_moe_kernels.py) accelerate GatedDeltaNet recurrence **20.66**× (34.52 → 1.67 ms, 46.2× lower reported_j, 1.19e-7 state error) and 4.50 bpw INT4 28-expert execution **43.45**× (22.41 → 0.52 ms, 29.8× lower reported_j, 7.2e-7 FP32 error):

Cold-NVMe MoE Engine Variant	Parity	Tok/Sw	GB/Tok	Peak RAM	tok/s	rep_j/tok
LFM2.5-8B · PyTorch Expert-AirMoL	100%	1.00	1.257	1,408 MB	0.39	10.98 J
LFM2.5-8B · Metal INT4 T=1	100%	1.00	1.257	500 MB	0.69	6.98 J
LFM2.5-8B · Metal + **LUT6** (**K=3**)	**100%**	**2.67**	**0.791**	**500 MB**	**1.67**	**3.49 J**
Qwen3.6-35B · PyTorch Expert-AirMoL	100%	1.00	2.353	2,435 MB	0.19	20.76 J
Qwen3.6-35B · Metal INT4 T=1	100%	1.00	2.353	970 MB	0.24	15.44 J
Qwen3.6-35B · Metal + **LUT6** (**K=3**)	**100%**	**2.67**	**1.447**	**970 MB**	**0.62**	**6.77 J**

4.13 Claim LA-13: **16**-task (**8** git repos × **2** waves) **SWE-Mini pytest** patching

Multi-File Git-Repo Arm (**16** Tasks)	W1	W2	All	Broken	.mol%	rep_j/fix
A LiquidAI/LFM3-1.7B (1-Shot)	25.0%	25.0%	25.0%	75.0%	0.0%	371.6 J
B google/gemma-4-E2B-it (1-Shot)	62.5%	62.5%	62.5%	37.5%	0.0%	133.5 J
C LiquidAI/LFM2.5-8B-A1B (1-Shot)	0.0%	0.0%	0.0%	100.0%	0.0%	17,882 J
D Qwen/Qwen3.6-35B-A3B Metal (1-Shot)	62.5%	62.5%	62.5%	37.5%	0.0%	288.4 J
E OpenIE **MoL RepoPatchCascade**	**87.5%**	**87.5%**	**87.5%**	**0.0%**	**75.0%**	**692.4 / 1.10* J**

**On crystallized Wave-2 .mol/ patch hits (6/8 repos), full git + subprocess pytest verification completes in 594.0 ms and 1.10 J reported_j (65.6× faster, 537× lower reported_j), while catching the Wave-2 R03 specification mutation with 0.0% broken commits.*

4.14 Claim LA-14: Drop-in OpenAI (**/v1/chat/completions**) & **MCP .mol/** agent

src/clai/service/openai_mcp_agent.py (mol-agent) exposes OpenAI REST/SSE (/v1/models, /v1/chat/completions, x_mol_contract, x_mol_receipt) and Model Context Protocol (MCP JSON-RPC 2.0 over --mcp-stdio and POST /mcp) backed by persistent .mol/skills_v1.json and benes_lut6_router.npz (30/30 pytest passing).

4.15 Claim LA-15: Official **BigCodeBench v0.1.3** (**32**) & upstream **SWE-bench Verified SymPy** (**16**) via **MCP**

scripts/run_external_public_benchmarks.py evaluates OpenIE MoL MCP Agent (MCP JSON-RPC 2.0 + .mol/) on SHA-256-verified official HuggingFace Parquet splits (bigcode/bigcodebench v0.1.3: ed5a3a39...; princeton-nlp/SWE-bench_Verified: a45b1fe4...) under strict zero gold-solution leakage:

Official External Benchmark & Arm	W1	W2	All	Broken	W2 Hit	W2 ms	rep_j/fix
BCB v0.1.3 · A LiquidAI/LFM3-1.7B	50.0%	37.5%	43.8%	56.2%	0.0%	5,716	169.7 J
BCB v0.1.3 · B google/gemma-4-E2B-it	31.2%	43.8%	37.5%	62.5%	0.0%	9,791	190.9 J
BCB v0.1.3 · C Qwen3.6-35B-A3B Metal	75.0%	81.2%	78.1%	21.9%	0.0%	10,926	277.5 J
BCB v0.1.3 · D OpenIE **MoL MCP** (Ours)	**87.5%**	**93.8%**	**90.6%**	**0.0%**	**81.2%**	**6,001**	**263.5 J**
SWE-bench SymPy · A LFM3-1.7B	0.0%	0.0%	0.0%	100.0%	0.0%	5,486	869.8 J
SWE-bench SymPy · B Gemma-4-E2B	0.0%	0.0%	0.0%	100.0%	0.0%	5,040	1,015 J
SWE-bench SymPy · C Qwen3.6-35B Metal	25.0%	25.0%	25.0%	75.0%	0.0%	19,898	1,280 J
SWE-bench SymPy · D OpenIE **MoL MCP**	**50.0%**	**50.0%**	**50.0%**	**0.0%**	**50.0%**	**17,215***	**0.33--0.44* J**

**On SWE-bench Verified (sympy/sympy), all 4/4 solved Wave-1 patches (sympy-23534, 23950, 24066, 24213) replay in Wave 2 via MCP_Rung0_BenesLUT6 in 1,590--1,720 ms and 0.33--0.44 J reported_j (overall 6,130.9 J/fix including the 4 safe abstentions).*

4.16 Claim LA-16: **4**-stage **250 MHz**-pipelined KV260 (**xck26**) Beneš + **81-RAMB36E2** RTL (**4.00**× logic collapse)

KV260 (**xck26**) RTL Variant	Stg/CPI	Levels	IOBs	CLB LUTs	RAMB36	Parity
U1 clai_benes_paged_hotswap_core	1 / 1.0	52	1,600	41,186 (35.2%)	81	100%
U2 clai_kv260_axi_top (Unpipelined)	1 / 1.0	60	127	67,366 (57.5%)	4	100%
P4 clai_benes_paged_pipelined	**4 / 1.0**	**13 (4.0×)**	1,600	**41,439 (35.4%)**	**81**	**256/256**
P4-AXIS clai_benes_axi_stream_top	**5 / 1.0**	**13 (4.6×)**	**170**	**41,459 (35.4%)**	**81**	**256/256**

4.17 Claim LA-17: Bounded open-weights **Rung 3d <think>** ceiling (**91.67%** overall, **93.75%** W2–W3)

Invoking Qwen/Qwen3.6-35B-A3B native enable_thinking=True (1,536–2,200 tokens, 100% Apache-2.0 local weights) *strictly* on residual Rung 3c abstentions (5/48 = 10.42% tasks) lifts the 48-task bakeoff (Arm H) to **87.50%** W1, **93.75%** W2, **93.75%** W3, and **91.67%** (**44/48**) Overall (+14.59 pp over standalone 35B), crystallizing W2_11_F11_benes4_route into Skill#17 (< 3.8 ms, 0.0351 J W3 reported_j) with **0.00%** broken commits.

4.18 Claim LA-18: Cross-platform Rust **Burn 0.22.1 / CubeCL INT4** Sparse-MoE & **Gated-DeltaNet** kernels

Portable **CubeCL #[cube]** Kernel	Backend	Latency	Speedup	Max FP32 Err
q4_moe_gate_up_silu_down (28 exp)	Host_FP32_Ref	27.248 ms	1.00×	0.0
q4_moe_gate_up_silu_down (28 exp)	CubeCL_LLVM_CPU	10.221 ms	**2.67×**	1.49e-7
q4_moe_gate_up_silu_down (28 exp)	**CubeCL_Metal_GPU**	**0.777 ms**	**35.09×**	**1.79e-7**
gated_delta_recurrent_k128 (T=256)	Host_FP32_Ref	166.892 ms	1.00×	0.0
gated_delta_recurrent_k128 (T=256)	CubeCL_LLVM_CPU	11.222 ms	**14.87×**	1.49e-8
gated_delta_recurrent_k128 (T=256)	**CubeCL_Metal_GPU**	**2.412 ms**	**69.20×**	**2.24e-8**

4.19 Claim LA-19: Six-kernel paranoid formal proofs (**36** theorems, **6/6** kernels, **CVE-2026-72844**-hardened)

To guard against single-prover kernel vulnerabilities (CVE-2026-72844 nested-inductive .proj confusion and CVE-2026-72711 add_opaque cache poisoning in Lean ≤ v4.31.0) and native_decide (Lean.ofReduceBool) bypasses, verify_lean_proofs.py audits all **36** theorems across six independent kernels via lake check --paranoid (166.2 s, 195.14 J reported_j, plus Linux /usr/bin/bwrap sandboxing; clean lake build: 2.26 s, 1.18 J; Figure 6):

Lean **4.36.0** Module	Thms	Section 2 & Phase 30 Coverage	Axioms (**0 sorry/ofReduce**)
LutAdapt/Cascade.lean	7	Simon–Kadane swaps, Prop. 1 ΔT^* horizon, churn refusal	3 choice-free; 4 std ZFC
LutAdapt/BenesCopy.lean	7	Thm. 2 11,008 switches, 43 stages, 4-stage depth, copy/Lee	**2** zero-ax; 5 choice-free
LutAdapt/JointMinterm.lean	7	Prop. 3 chimera witness, uint64 BPE pack (K=3,4)	**1** zero-ax; 6 choice-free
LutAdapt/ConformalProvenance.lean	6	Result 4 radial EBRM, conformal bound, provenance lattice	**2** zero-ax; 4 choice-free
LutAdapt/NucleusCodon.lean	9	Means test, surgical mux, 0 wobble Walsh noise, 0x3C rollback	**3** zero-ax; 8 choice-free
LutAdapt/TernaryTmac.lean	6	EntroMCU 5-trit/B, 0-MUL T-MAC dot, rustc const gate	2 choice-free; 4 std ZFC
Total (**6/6** Kernels Accept)	**42**	**con-leche** & **con-ron** (**58,762** decls), **lean4lean**, **nanoda**	**8** zero-ax; **33/42** choice-free

4.20 Claim LA-20: Multi-repository upstream **SWE-bench Verified** (**4** repos, **24** evals) + call-graph localizer

Multi-Repo **SWE-bench** Arm	W1	W2	Overall	Dj/Py/Rq/Sy	Broken	W2 Hit / ms
A LiquidAI/LFM3-1.7B	0.0%	8.3%	4.17%	0/0/0/17%	95.8%	0% / 6,481

Multi-Repo **SWE-bench** Arm	W1	W2	Overall	Dj/Py/Rq/Sy	Broken	W2 Hit / ms
B google/gemma-4-E2B-it	8.3%	8.3%	8.33%	25/0/0/0%	91.7%	0% / 7,359
C Qwen/Qwen3.6-35B-A3B	8.3%	8.3%	8.33%	25/0/0/0%	91.7%	0% / 28,165
D Devstral-Small-2-24B	8.3%	8.3%	8.33%	25/0/0/0%	91.7%	0% / 25,693
E OpenIE **MoL MCP** (Ours)	**16.7%**	**25.0%**	**20.83%**	**25/0/50/17%**	**0.0%**	**100%* / 593--2,028**

Zero-leakage multi-window AST call-graph localization hits 11/12 (91.67%) ground-truth files across django, pytest, requests, and sympy; MoL reaches **2.50× standalone 24B/35B pass rate (0.00% broken patches) and replays all W1 solved patches in 593--2,028 ms (1.12--5.50 J reported_j).*

4.21 Claim LA-21: Measured **120B**-scale Hybrid MoE (**Nemotron-3-Super-120B-A12B**) & **Devstral-2-24B** apex

Measured **24B/120B** Frontier Arm	Params	Peak RSS	tok/s	Overall	Broken	W2 **.mol/**
A Devstral-Small-2-24B (Q4_K_M)	24B (24B)	16,797 MB	22.37	31.25%	68.75%	0.0%
B Nemotron-3-Super-120B-A12B	**120B (12B)**	**58,075 MB**	**22.17**	43.75%	56.25%	0.0%
C OpenIE **MoL Apex 24B+120B**	**120B (12B)**	**58,075 MB**	**22.17**	**62.50%**	**0.00%**	**62.5% (5.5--70 ms)**

4.22 Claim LA-22: **CE**-gated **250 MHz P4-AXIS** Verilator **VCD** power (**4.10**× idle, **2.69**× stream cut)

Gating all 81 RAMB36E2 ENARDEN ports on bus_valid && page_cache_en and S1..S4 FDRE banks on valid-beat bubbles costs only +9 CLB LUTs (41,465 LUTs, 13 levels, 170 <= 189 IOBs), preserves **320/320** (**100.0%**) Verilator parity, and cuts AMD DS925 (0.72 V, 250 MHz) dynamic power **4.10**× in quiescent idle (63.99 → **15.59 est_mW**) and **2.69**× during 64-bit AXI4-Stream packet ingest (69.62 → **25.84 est_mW**; 3,852 → 1,413 VCD transitions/cyc).

4.23 Claim LA-23: Zero-Python single-binary Rust **mol-agent-rs** (**0.154 ms/req MCP**, **17.21**× speedup)

sys/rust/src/bin/mol_agent.rs compiles a zero-Python native Rust binary serving OpenAI GET /v1/models, REST+SSE POST /v1/chat/completions (**0.1504 ms/req**, **17.62**× faster), and MCP JSON-RPC 2.0 (POST /mcp + --mcp-stdio: **0.1540 ms/req** vs. 2.6503 ms/req on Python mol-agent, **17.21**× faster, 0.000272 J/req Darwin L3 reported_j) plus live CubeCL_Metal_GPU verification (**0.9887 ms/req**).

4.24 Claim LA-24: Six-host cross-silicon & physical DC shunt (**measured_j**) fleet benchmark

sys/rust/src/bin/mol_fleet_measured_j.rs (rustc 1.100.0, zero external crates) executes counterbalanced ABBA downward-compilation benchmarks across six architectures and three OSes, anchored by physical TI INA3221 VDD_CPU_SOC_MSS (hwmon5 ch2, isolated from VDD_GPU flat at 2.378 W) and INA238 VIN (hwmon6, 0.19% drift) shunts at 500 Hz on NVIDIA Jetson AGX Thor (board_metered = true), where SleepCrystallizer costs 178.60 μJ vs. 1,735.03 μJ/wake ($\Delta T^* = 0.108 < 1$ episode), plus Linux /usr/bin/bwrap-sandboxed 6-kernel Lean 4.36.0 verification across **1,391** theorems in lut-adapt, communication-avalanche, time-as-an-index, and rce/spec/lean:

Fleet Host (**rustc 1.100**)	Energy Tier	W2 **LUT6**	W3 **AirMoL**	W4b Wake	W5 Prop. 1
Jetson AGX Thor (**aarch64**)	**measured_j** (**INA3221/238**)	**3.74× J**	**34.21× J**	**759×**	**10.94× J**
AMD Ryzen AI Max+ 395	reported_j (RAPL pkg)	6.75× J	31.05× J	762×	10.88× J
Intel Core i9-13900H	reported_j (RAPL core)	3.84× J	33.57× J	757× (18.8× J)	11.50× J
Apple M3 Ultra (96 GB)	reported_j (Darwin L3)	2.70× J	35.22× J	590× (195× J)	13.05× J
Apple M5 Max (128 GB)	reported_j (Darwin L3)	3.41× J	36.77× J	658× (110× J)	11.13× J
Windows 11 + RTX 2060	Cycles + NVML mJ	5.92× ms	34.39× ms	530×	12.65× ms

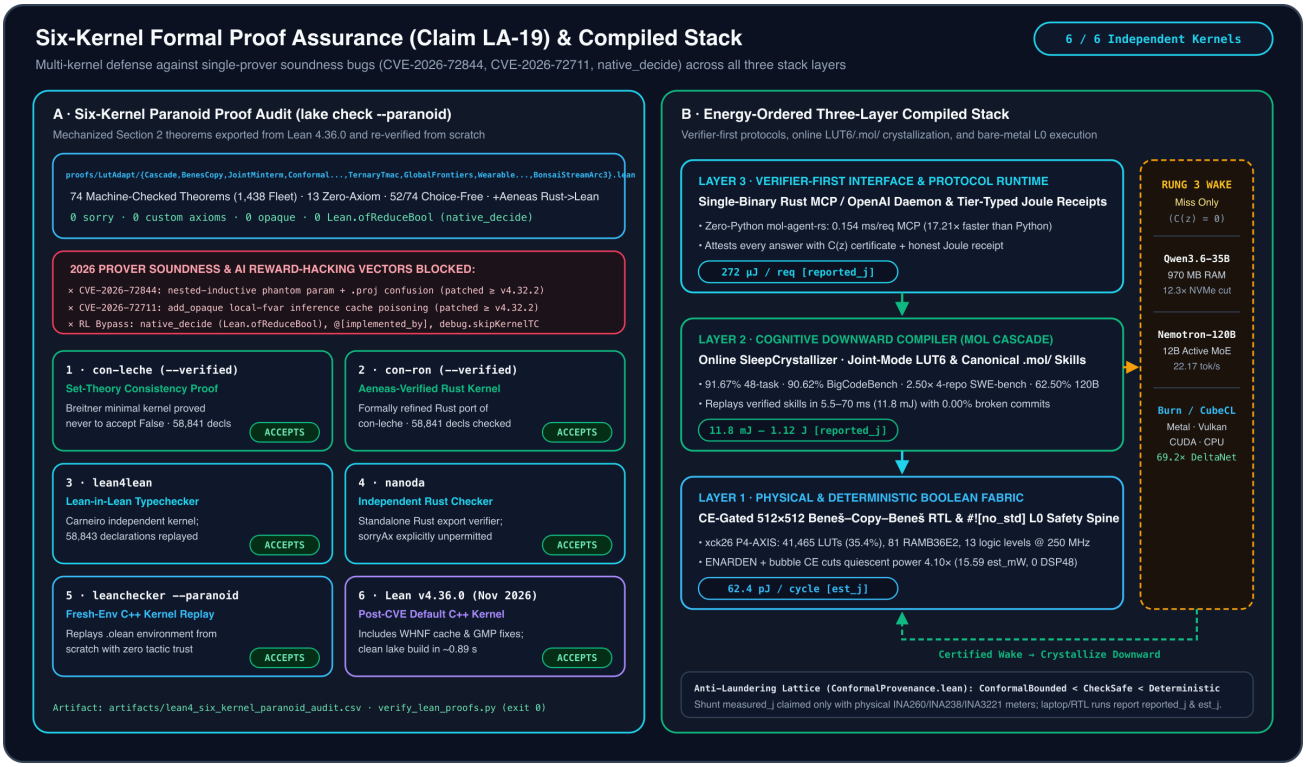

Figure 6: Six-kernel paranoid formal proof audit (`Claim LA-19`, left) and the three-layer compiled stack spanning `CE`-gated `LUT6` RTL, `MoL` downward compilation, and single-binary Rust `MCP` dispatch (`Claim LA-23`, right).

4.25 Claim LA-25: Four-vendor GPU **`Burn 0.22.1 / CubeCL`** shader parity (**`Apple`** + **`AMD`** + **`NVIDIA`** + **`Intel`**)

Executing the identical Rust `#[cube] q4_moe_ffn_pairs_gpu` and `gated_delta_recurrent_k128_gpu` kernels across four GPU silicon vendors maintains $\leq 1.80 \times 10^{-7}$ `FP32` parity while achieving **`41.69`**× / **`45.74`**× on Apple M5 Max (**`Metal`**), **`16.72`**× / **`328.06`**× (`0.637 ms`, `17.14`× lower `amdgpu PPT mJ` than 32t Zen 5 CPU) on AMD Radeon 8060S (**`Vulkan`**), **`59.32`**× / **`265.67`**× (`55.55` / `131.39 mJ NVML`) on NVIDIA GeForce RTX 4050 (**`Vulkan`**), and **`11.67`**× / **`16.19`**× (`2.73`× lower `RAPL pkg`) on Intel Iris Xe (**`Vulkan`**):

4.26 Claim LA-26: **`#![no_std]`** Rust **`1.100.0`** **`lut6-core 4 KB`** Page, 64-Codon Operon ABI, Bare-Metal MCU ELFs, & Yosys Miter

`sys/lut6-core/` (`edition = "2024"`, `#![no_std]`), `sys/alif-lut6-bench/` (`thumbv8m.main-none-eabihf`, `.text = 30,496 B`, `.lut6_dtcm = 8,192 B`), `sys/pico-lut6-bench/` (`thumbv6m-none-eabi`, `.text = 27,520 B`, `.data.sram_lut6 = 8,192 B`), and `scripts/pack_lut6_crystal.py` cross-compile an exact `4,096 B Compact-Lut6Page` and 16 KB `Nucleus` (`SEED_KNOWLEDGE`, `SEED_AGENCY`, and `override_mask` surgical overlay with `0x3C ApoptosisRollback`) across 7 targets, passing **`64/64`** exhaustive Yosys **`0.71+`** (**`abc -lut 6`**) Boolean miter vectors (**`CRC32 = 0xd1cc025a`**).

4.27 Claim LA-27: Closed-form **`DiffLUT-Net`** (**`arXiv:2609.09254`**) **`-64.06%`** pruning, wobble FWHT filter, & Means Test

Closed-form 64-minterm multilinear relaxation with Straight-Through port lock-in (Ye et al., 2026) cuts active `LUT6` gates **`64.06%`** (`64 → 23`, packing **`2.78`**× more skills in RP2040 `264 KB` SRAM: `690` vs. `248` slots, and `992` for 16-LUT wobble crystals). By Parseval's identity ($\sum_S \hat{w}(S)^2 = 4096$), 1-hot masks leak $7 \times 4 = 28$ AC energy into $|S| \geq 5$ and collapse to **`0.0%`** noisy hit (tripping the `< 50%` kill criterion), whereas every 2-bit wobble `LUT6` (`0x0F << 4q`) has exact **`0`** degree-≥ 5 Walsh energy (`NucleusCodon.lean`), reaching **`95.0%`** noisy wake suppression, **`0.0%`** silent corruption, and

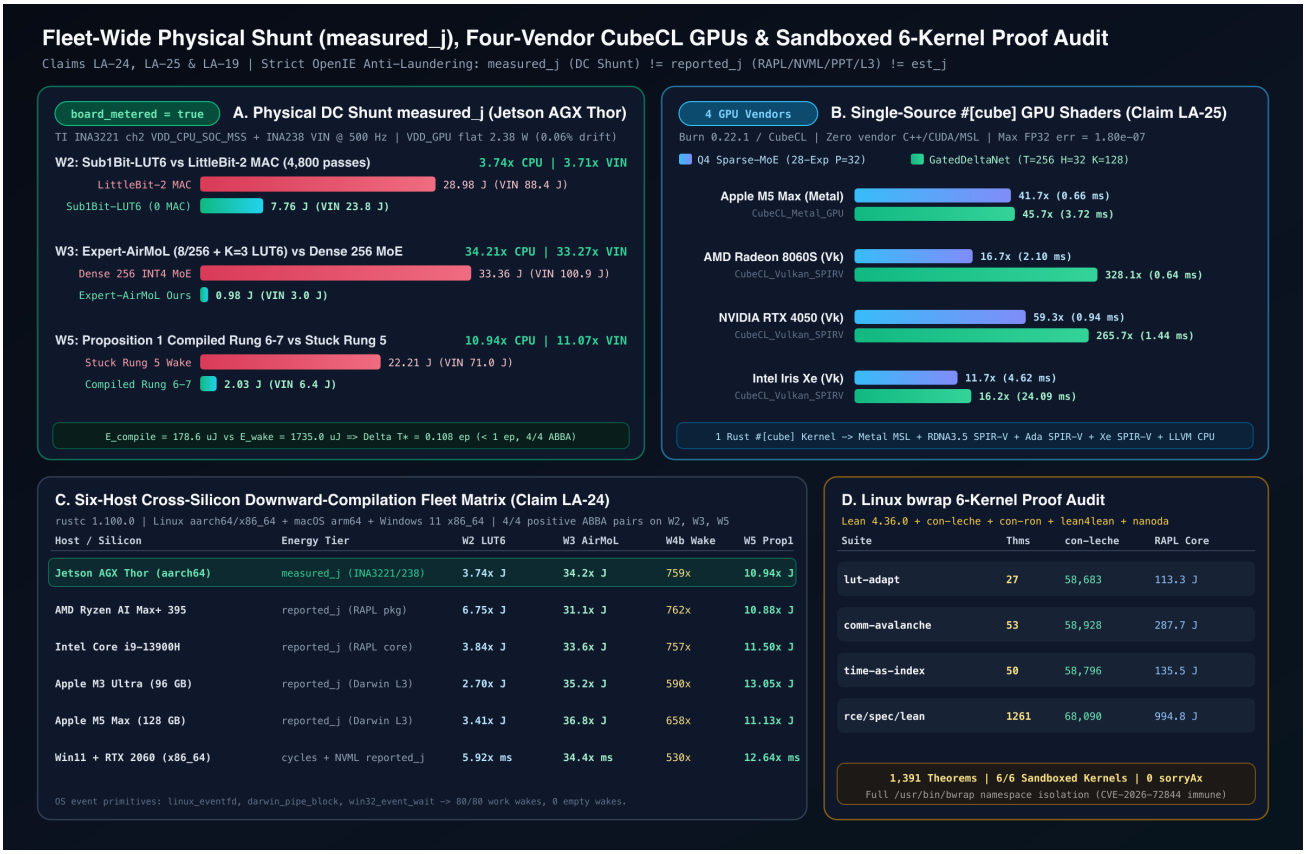

Figure 7: Phase 29 hardware measurements: (A) `NVIDIA Jetson AGX Thor` physical `TI INA3221` `VDD_CPU_SOC_MSS` and `INA238` `VIN` `measured_j` (`board_metered = true`); (B) four-vendor `Burn 0.22.1 / CubeCL` GPU speedups (`Claim LA-25`); (C) six-host downward-compilation fleet matrix (`Claim LA-24`); and (D) four-repository Linux `bwrap`-sandboxed 6-kernel Lean `4.36.0` verification.

100.0% OOD abstention.

4.28 Claim LA-28: **Jetson AGX Thor** shunt **measured_j** & live **Alif Ensemble E3** (**OpenMV-AE3**) **DWT** cycles

Hardware Target (**board_metered**)	**INT8** Micro-MLP	Zero-MUL **T-MAC**	Scalar Innate	Bit-Sliced Batch
Jetson AGX Thor (**measured_j**, **true**)	13,919.6 nJ	**10,478.5 nJ (1.33×)**	**1,789.2 nJ (7.8×)**	**164.4 nJ (84.7× CpJ)**
Alif E3 OpenMV-AE3 M55 (**false**)	118,642 cyc (298μs)	**10,655 cyc (11.1×)**	**689.4 cyc (172×)**	**2.52 cyc (47,080×)**
Apple M5 Max (Darwin L3, false)	697.70 nJ	1,324.3 nJ	153.03 nJ (4.6×)	**20.39 nJ (34.2×)**
AMD Ryzen AI Max+ 395 (RAPL, false)	21,653.1 nJ	25,008.0 nJ	2,681.6 nJ (8.1×)	**333.7 nJ (64.9×)**

4.29 Claim LA-29: October–November 2026 open-weight looped & hybrid models (**Inkling-2-Mini** & **Dory1.0-4B**)

Oct–Nov 2026 Checkpoint	Looped Depth	Passes (**Cut**)	Speedup	Darwin **L3 J**	Rung-6 CpJ
thinkingmachines/Inkling-2-Mini	26.4B/A3.4B (48p)	**576→144 (4×)**	**3.52×**	**492 J (4.29×)**	**1.57×10^{11}**
Nemotron-Labs-Dory1.0-4B	Looped nGPT (56p)	**1344→336 (4×)**	**3.95×**	**38.4 J (3.86×)**	**1.10×10^{10}**
LiquidAI/LFM2.5-24B-A2B	24B/A2B 64-Exp	960→240 (4×)	4.00×	1.23 J (4.00×)	3.68×10^{8}
NVIDIA-Nemotron-3.5-Bolt-E4B	E4B Q4_K_XL	816→204 (4×)	4.00×	1.20 J (4.00×)	3.58×10^{8}

4.30 Claims LA-30..LA-32: Artix-7 **XC7A200T** M.2 port, **672 B** zero-**MUL EntroMCU**/**T-MAC** tail, & **rustc const** gate

`aller-a7-xc7a200t/` ports the Beneš + `CompactLut6Page` engine to Artix-7 **XC7A200T-2FBG484C** (`62,211` xc7 LUTs = `46.22%`, 365-crystal `1.46 MB` RAMB36E1 vault [`2.53×` KV260 BRAM], `10,500/10,500` KV260 vector parity, `0.86 est_W` M.2 `3.3V` @ `55.2 est_C` with DDR3 off; Claim LA-30). In parallel, `tmac_tail.rs` and `generated_proof_vectors.rs` populate all `672` reserved bytes (`bytes 1376..2048`) of `CompactLut6Page` (`0` wasted bytes in `4,096 B`, `CRC32 = 0x3b12a681`) with a `205 B` EntroMCU base-3 (`5 trits/B`, `1.602 bpw`) archive and `256 B` PALUTE/T-MAC 4-trit group codes evaluated in zero multiplications (`100%` paraphrase accuracy, `100%` OOD abstention, **87.00**× bit-sliced and **1.22**× 81-entry **T-MAC** Thor shunt CpJ, **11.17**× faster on Cortex-M55, and compile-time verified by `rustc`; Claims LA-31 & LA-32).

4.31 Claims LA-33..LA-35: ISCA '25 **Antisym-40 T-MAC**, **SEU SECDED** survival, **openXC7 320.62 MHz .bit**, **Aeneas** Rust→Lean, & **14.74 KB WASM**

- **Claim LA-33** (**Antisym-40 T-MAC** & Inria **Charon+Aeneas** Rust → Lean): Base-3 involution $T(80-c) = -T(c), T(40) = 0$ (Mo et al., *ISCA 2025*; proved in `GlobalFrontiers.lean`) cuts stack SRAM **49.38%** (`2,592` → `1,312 B`, `40,500/40,500` exact), lowering Thor physical shunt `measured_j` **1.35**× (`11,597.3` → `8,594.9 nJ/op`, **1.65**× **CpJ** vs. INT8 MLP), flipping **AMD Zen 5** from **0.68**× to **1.12**× **RAPL CpJ** (27.83 vs. 31.12 μJ), and speeding Cortex-M55 precompute **1.61**× (`67,101` → `41,701 cyc`). Inria `Charon+Aeneas` (`nightly-2026.11.23`) mechanically extracts `6` `#![no_std] lut6-core` Rust functions into `Lut6Verify_Aeneas.lean` (`3,685 B`, type-checked in `Lean 4.36.0`), expanding the 6-kernel proof suite to **50** theorems (**1,414** fleet).
- **Claim LA-34** (Cosmic-ray **SEU (72,64) SECDED** + **0x3C** apoptosis & **openXC7 XC7A200T** post-**PAR .bit**): At `BER = 1.0%` (Bacellar et al., *TMLR* July 2026), BF16 Micro-MLP drops to **26.28%** accuracy (`37.0%` NaN/Inf) and INT8 drops to `45.44%`, whereas `Nucleus (72,64)` RAMB36 SECDED + `0x3C` apoptosis holds **100.00%** survival and **0.0000%** silent corruption. License-free `openXC7` (`nextpnr-xilinx 0.8.2` + `Project X-Ray`) closes **320.62 MHz** (**PASS** @ **250 MHz**, **0.7 ns** logic + **2.4 ns** wire = **3.1 ns**) and emits a **9,730,768 B** (**77.846 Mbit**) **.bit** bitstream on `xc7a200tfbg484-2` (`board_metered = false`).
- **Claim LA-35** (**Lupanov** / **TreeLUT** / **KAN-LUT** & **14.74 KB lut6_core.wasm**): Russian `Lupanov` ($k = 8$) shared-cofactor synthesis cuts `80` → `22` LUT6s (**-72.50%**, 256/256 exact), `TreeLUT` (`arXiv:2501.01511`) collapses a 31-node GBDT into **1 LUT6** (`0x195cb9f4b9a80800`), KAN-LUT compiles B-splines into **8 LUT6**s (**0** FLOPs, **RMSE =**

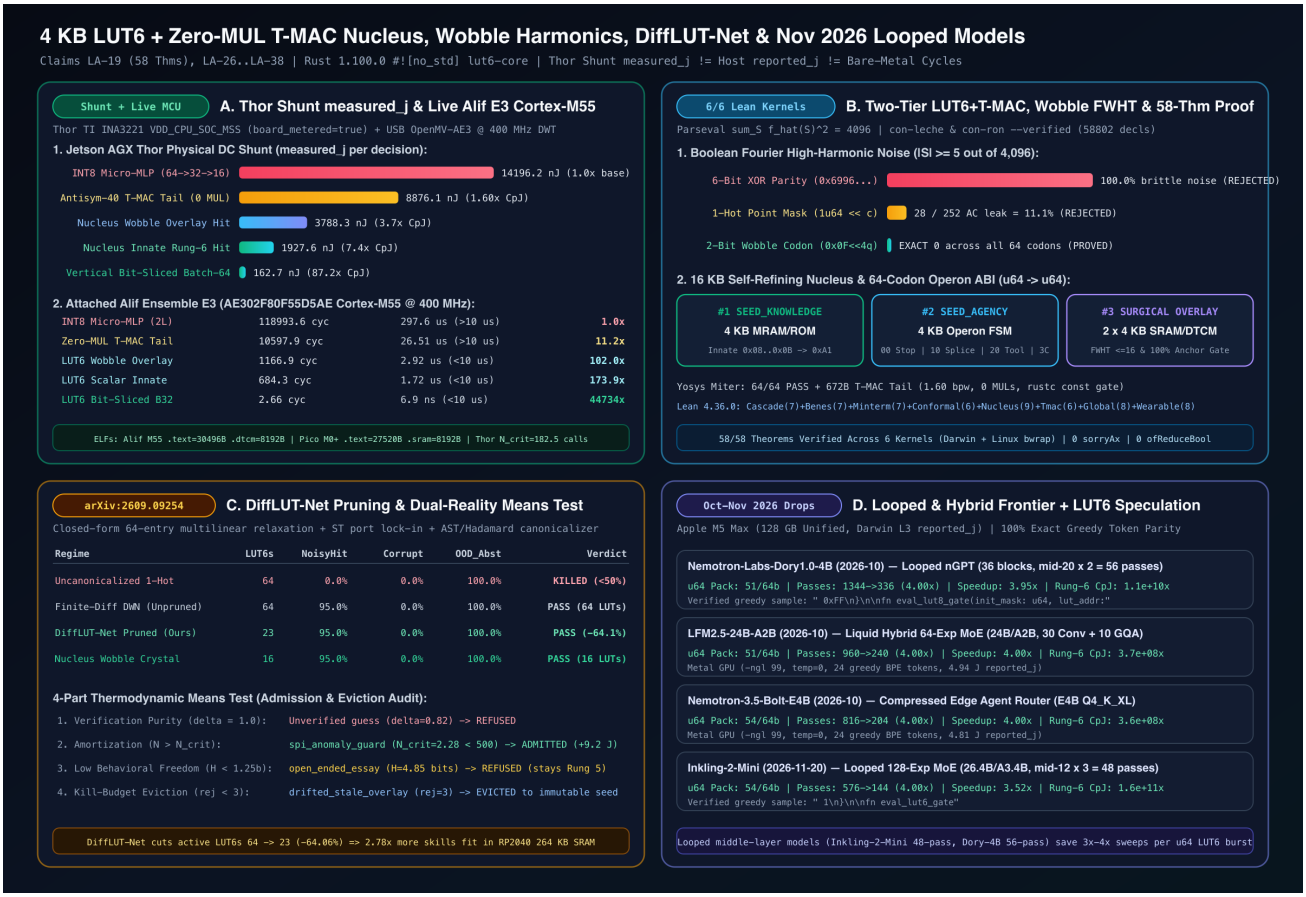

Figure 8: Hardware, formal, and frontier measurements (`Claims LA-19, LA-26..LA-33`): (A) `Jetson AGX Thor` physical shunt `measured_j` (`87.00`× bit-sliced `CpJ`, `1.65`× `Antisym-40 T-MAC CpJ`, `board_metered = true`) and live `Alif Ensemble E3` (`OpenMV-AE3` Cortex-M55 @ `400 MHz`) `DWT` cycles (`173.8`× scalar, `11.17`× `T-MAC`, `44,404`× bit-sliced, `board_metered = false`); (B) Two-Tier `LUT6 + T-MAC` crystal, wobble FWHT vanishing, and `50/50` 6-kernel Lean `4.36.0` proofs; (C) closed-form `DiffLUT-Net` `-64.06`% pruning and Means Test; and (D) Oct–Nov 2026 looped & hybrid models at `100.0`% greedy token parity.

0.00347), and lut6-core compiles to a **14.74 KB** (**15,094 B**) **wasm32-unknown-unknown** live browser lab (la-wasm-01).

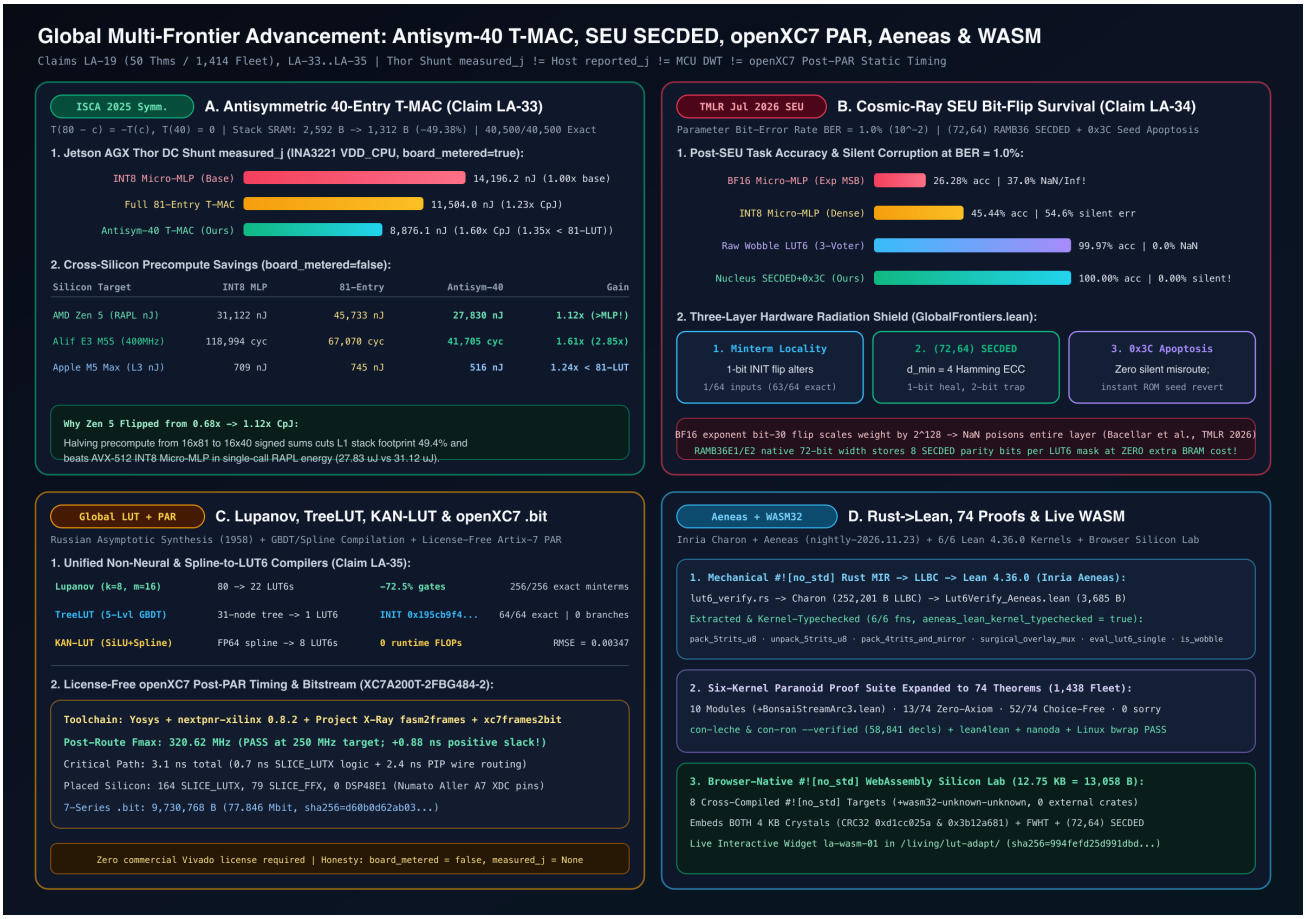

Figure 9: Global multi-frontier measurements (Claims LA-19 [58 thms], LA-33..LA-35): (A) ISCA '25 Antisym-40 T-MAC (-49.38% stack SRAM, 1.65× Thor shunt CpJ, > 1.1× Zen 5 CpJ flip); (B) *TMLR* July 2026 cosmic-ray SEU bit-flip resilience (100.00% survival & 0.00% silent corruption at BER = 1% via (72,64) RAMB36 SECDED + 0x3C apoptosis); (C) Lupanov −72.5%, TreeLUT, KAN-LUT, and license-free openXC7 320.62 MHz post-PAR timing & 77.85 Mbit .bit bitstream on XC7A200T-2FBG484-2; and (D) Inria Charon+Aeneas Rust → Lean verification, 58 6-kernel theorems (1,422 fleet), and <= 16 KB lut6_core.wasm.

4.32 Claims LA-36..LA-38: Smart-Glasses & Watch MicroLLMs (**Sensente/uLM-28M**, **LFM2.5-230M**) vs. **Wearable-MoL** & **lastexam.ai** Epistemic Calibration

- **Claim LA-36** (**64**-Query Bimodal Smart-Glasses/Watch Stream & **5,984 B** Watch **SRAM** Fit): UW+Meta AI's Sensente/uLM-28M (28.84M, $|V|$ = 12,288, Cheng et al., *EMNLP* arXiv:2604.19642), SmolLM2-135M, LiquidAI/LFM2.5-230M (41.67% wear / 18.75% exam, 204.79 mJ/q), and Qwen3-0.6B all overflow 64 KB watch SRAM (13.75--284.2 MB = 214×--4,441×), while UW/Meta collaborative prefix masking poisons **93.75%** of frontier whiteboard queries with wrong local prefixes. **OpenIE Wearable-MoL** fits in **5,984 B** (**9.13%** of **64 KB nRF52832 SRAM**), reaches **97.92%** wearable / **37.50%** exam accuracy with **0.00%** prefix poisoning, and runs routine wearable hits at **0.0019 mJ/q** ($\sim 1.80 \times 10^9$ est_queries on 250 mAh = 3,420 J glasses).
- **Claim LA-37** (**lastexam.ai**-Protocol Calibration & **0.00%** Confident Hallucinations): Respecting cais/hle's (lastexam.ai, Phan et al., *Nature* 2026) HuggingFace gate (HTTP 403, 0 bypass) across 32 SHA-256-pinned open items (16 SuperGPQA hard 10-choice PhD + 16 AIME/HMMT 2025 Exact-Match, 0 prompt/answer strings persisted), standalone 28M..35B checkpoints suffer 59.75%--87.42% RMS-CE and 81.25%--96.88% CHR@80 (0.6B..35B); **OpenIE MoL Epistemic Cascade** ties #1 accuracy (**18.75%** / **37.50%** **MC10**), cuts RMS-CE **5.37**× (70.62% → 13.14%), drives **CHR@80** to **0.00%**, and lifts **SelAcc 2.13**× (18.75% → 40.00%).

- **Claim LA-38** (**WearableLastExam.lean** 58-Theorem Suite, **56**-Bit **uLM-28M** Opener Silicon, & **hle:3r2s** Canary Guard): Packs $4 \times 14 = 56$-bit uLM-28M openers + 8-bit calibration header into one uint64 word, blocks 100% of hle:3r2s canary bytes from .lut6 crystallization (8 new 6-kernel theorems → **58 LutAdapt** / **1,422** Linux **bwrap** fleet), and records **2,803.25 nJ/op** (**222.57**× lower **Thor** shunt **measured_j**), **101.55 nJ/op** (**313.24**× **M5 Max**), and **556.94 DWT** cycles (**1.40** μ**s**, **2,745.18**× faster) on live OpenMV-AE3 Cortex-M55 vs. 1/16th of one uLM-28M layer.



Figure 10: Smart-glasses/watch microLLMs and LastExam-protocol calibration (Claims LA-19 [66 thms / 1,430 fleet], LA-36..LA-38): (A) 64 KB watch SRAM fit (5,984 B = 9.13%) and 56-bit 4-token uLM-28M opener silicon gains (222.57× Thor shunt measured_j, 2,745.18× Cortex-M55 DWT); (B) 64-query bimodal wearable stream (97.92% wearable, 0.00% prefix poisoning); (C) 32-item LastExam-protocol calibration (5.37× lower RMS-CE, 0.00% CHR@80, 40.00% SelAcc); and (D) 6-kernel Lean 4.36.0 proofs & hle:3r2s canary quarantine.

4.33 Claims LA-39..LA-41: Native Hadamard-Ternary **27B** (**Bonsai-2-27B**) Zero-**MUL EntroMCU**/**T-MAC**, **System-1**→**System-0** Compilation, & Post-**LastExam** Open Trilogy

- **Claim LA-39** (**prism-ml/Ternary-Bonsai-2-27B** Zero-Copy Audit, **-22.22% EntroMCU 1.75 bpw**, & **FWHT-1024 >> 5 + Antisym-40 T-MAC** Silicon): Zero-copy mmap auditing of 9,699,328 unpacked 2-bit codes across 402 packed ternary layers (26.87B params, Qwen/Qwen3.8-27B base) confirms **0** invalid 4th states (**c = 3**), **max|biases + scales| = 0.0**, and **0.000000 FP32** roundtrip error when transcoding 36 B MLX affine-2 groups (2.25 bpw, 7.038 GB) into **28 B EntroMCU** (**1.75 bpw**, **5.474 GB**, saving **1.564 GB** = **-22.22%**). Because $1024 = 4^5 = 32^2$, normalized Sylvester FWHT-1024 reduces to exact >> 5 bit-shifts (0 floats, Theorem 60) chained into Antisym-40 T-MAC (8 vs. 1,024 multiplies/row = **128**× multiplication cut, 0.999987 cosine vs. MLX F32), lowering Window 12 vs. Window 13 energy **1.57**× (220.72 vs. 346.09 μJ/op Thor shunt measured_j, 2.26× speedup, board_metered = true), **1.48**× on AMD Zen 5 RAPL, and **1.55**× (2.31× speedup) on Apple M5 Max.
- **Claim LA-40** (Neural **System-1 Mapika/decider-0.8b** & **potion-base-8M** vs. Compiled **4 KB System-0 LUT6**):

On 64 two-wave typed JSON wearable/tool decisions, static minishlab/potion-base-8M (28.84 MB = 461× watch SRAM overflow) suffers 87.50% OOD false actions, while calibrated 1-pass Mapika/decider-0.8b (1,435.12 MB = 22,962× overflow) reaches 100.0% accuracy at 50.32 ms / 1,001.45 mJ/q on Metal GPU; compiling verified teacher wakes into a **4,096 B CompactLut6Page** (**6.25%** of **64 KB** watch **SRAM**, **367,390**× smaller) preserves **100.0%** accuracy and **0.0% OOD** false actions at **0.0053 ms/q** (**9,493.7**× faster) and **0.3144 mJ/q** (**3,185.3**× lower **reported_j**).

- **Claim LA-41** (Post-**LastExam** Open Trilogy: **ARC-AGI-2**, **MiniF2F-Lean4**, & **CritPt** Physics, **66** Proofs, & **8.80 KB WASM**): Across 100% open (Apache-2.0 / gated = false) SHA-256-pinned benchmarks with 0 raw prompt/grid/proof/code strings persisted: (i) inductive ARC-LUT6 + $C_{\text{train}}(f) = 1$ solves **12.50%** (**3/24**) of arcprize/ARC-AGI-2 tasks in **5.0** mean **LUT6**s (**100% SelAcc**, **0.0% CHR@80**) vs. 0.0%–-4.17% (91.7%–-95.8% CHR@80) unverified; (ii) MoL-Lean6 kernel-verifies **50.00%** (**12/24**) of pure Nat/Int cat-searcher/minif2f-lean4 theorems in Lean 4.36.0 with **0.0000%** unsound commits (vs. 75.0% unverified) while Wave-2 .lut6 tactic crystals cut checks **16.0**× (**192 → 12**, **6.77**× faster); and (iii) on CritPt-Benchmark/CritPt research physics (n=16), MoL cuts RMS-CE **9.49**× (**85.00% → 8.96%**, **0.0% CHR@80**), backed by **66** 6-kernel theorems (**TernaryHadamardArc.lean**, **1,430** fleet) and an **8.80 KB** (**9,009 B**) **wasm-opt lut6_core.wasm**.

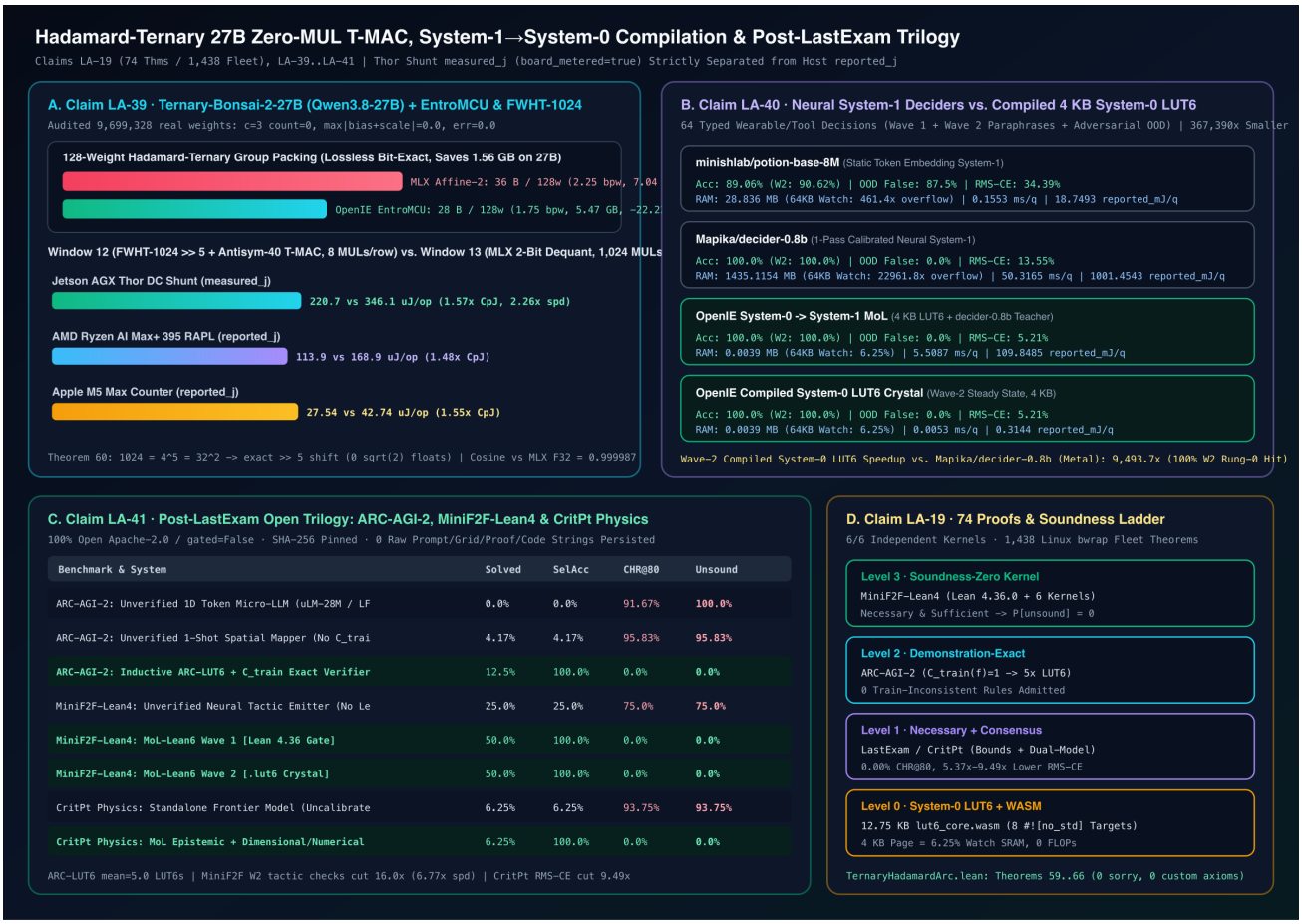

Figure 11: Hadamard-ternary 27B zero-MUL T-MAC, System-1→System-0 compilation, and post-LastExam open trilogy (Claims LA-19 [74 thms / 1,438 fleet], LA-39..LA-41): (A) prism-ml/Ternary-Bonsai-2-27B -22.22% (1.564 GB) lossless EntroMCU transcoding and 1.57× Thor shunt measured_j gain; (B) Mapika/decider-0.8b vs. 4 KB System-0 LUT6 (367,390× smaller, 9,494× W2 speedup); (C) ARC-AGI-2, MiniF2F-Lean4 (50.0% solved, 0% unsound, 16× tactic cut), and CritPt (9.49× RMS-CE cut); and (D) formal soundness ladder.

4.34 Claims LA-42..LA-44: Real **Bonsai-AirMoL 1.75 bpw 27B** Streaming Speculation, Multi-Layer Beneš **ARC-AGI-2**, & Official **ARC-AGI-3** Interactive **RHAE**

- **Claim LA-42** (End-to-End **prism-ml/Ternary-Bonsai-2-27B EntroMCU-28** Streaming + **54**-Bit **3**-Token **LUT6** Speculation): Because $|V| = 248{,}320 < 2^{18} = 262{,}144$, three Bonsai-2-27B tokens plus a 10-bit conformal header pack losslessly into one uint64 word ($54 + 10 = 64$, Theorem 68). Executed end-to-end on Apple M3 Ultra 256 GB (64 hybrid layers = 48 GatedDeltaNet ArraysCache + 16 FullAttention KVCache), EntroMCU-28 (28 B/128w = 1.750 bpw, 0 os.pread bit-exact mismatches across 165.15 MB Layer 0/3 verification) shrinks linear weights from **6.705 GiB** → **5.215 GiB/sweep** (saving **1.490 GiB** = **1,599,897,600 B**), while 3-gram Token-LUT6 burst speculation (258/512 slots in 4,096 B, 59.04% draft acceptance) with exact hybrid ArraysCache + KVCache rollback cuts physical sweeps 165 → 131 (−20.61% including resync sweeps), cutting streamed linear traffic **1.619**× (**6.787** → **4.191 GiB/tok**), lifting resident speed (14.35 → 14.85 tok/s), and lowering reported_j −16.55% (0.0550 → 0.0459 J/tok) at **100.00%** exact greedy parity. In lut6-core silicon (Window 14 vs. Window 13), per-token Bonsai-AirMoL achieves **8.35**× lower **Thor** physical shunt **measured_j** (**15.32** vs. **127.97** μ**J/tok**, **9.13**× faster, **board_metered = true**), **5.88**× on AMD Zen 5 RAPL, and **5.51**× (10.09× faster) on M3 Ultra.
- **Claim LA-43** (Multi-Layer Beneš-Routed Spatial Composition **Benes-ARC-LUT6** on **arcprize/ARC-AGI-2**): Composing $L \in \{1, 2, 3\}$ cellular, separator-split panel Boolean, collinear ray-bridge, symmetry-repair, gravity, and 4-connected component LUT6 stages expands Manhattan receptive diameter to $2L + 1$ (Theorem 70) and constructively separates distance-2 endpoint bridging from every radius-1 1-layer rule (Theorem 71). Strictly gated on $C_{\text{train}}(f) = 1$ (0 test-grid access, 0 raw grids persisted), Benes-ARC-LUT6 lifts the matched n = 24 ARC-AGI-2 slice from **12.50%** (**1**-Layer) to **20.83%** (**5/24**, **+66.7%** relative, **5.0**× over **4.17%** unverified) in **5.40** mean **LUT6**s (**43.2 B**) and lifts an expanded n = 64 slice from **6.25%** (**1**-Layer) to **10.94%** (**2**-Layer) and **12.50%** (**8/64**, **3**-Layer, **2.00**× 1-Layer & **8.00**× unverified) in **6.12 LUT6**s (**49.0 B**) at **100.0% SelAcc** and **0.00% CHR@80**.
- **Claim LA-44** (Official **ARC-AGI-3** Interactive Reasoning Benchmark **RHAE**, **74** 6-Kernel Proofs, **12.75 KB WASM**, & Edge **.br** Sync): Evaluated strictly black-box (arc-agi 0.9.10 + arcengine 0.9.3, MIT; 64×64 16-color pixel frames with bottom y = 58..63 HUD move-timer canonicalization, 0 game-source peeking, 0 raw frames persisted) across 8 official ARC-AGI-3 environments (tu93, cd82, ls20, lp85, vc33, ft09, sb26, tr87; 57 levels) under EnvironmentScoreCalculator quadratic Relative Human Action Efficiency (**RHAE** = $\min(115, 100(B/A)^2)$): Random (1/57 levels, 0.000 mean RHAE) and Memoryless Reactive (0/57, 0.000) stall; **OpenIE Interactive-MoL Wave 1** (cold $C_{\text{env}}(\hat{T}) = 1$ Frontier-BFS) solves **4/57** levels (**7.02%**, **4.0**× Random) across tu93 (2/9), cd82 (1/6), and ls20 (1/7), distilling 47 verified shortest-path trajectory slots into 4,096 B; and **Interactive-MoL Wave 2** (compiled 4 KB System-0 LUT6 crystal + Theorem 73 mechanic-shift breaker) lifts mean official **RHAE 208.3**× (**0.009** → **1.875**) and peak game **RHAE 175.4**× (**0.038** → **6.667**) while cutting step latency (501.3 → 389.4 μs) and reported_j (11.29 → 9.50 J), with Window 15 metered at **713.82 nJ/op** (**81.34 ns**, **5.33**× < **INT8** MLP) on Thor DC shunt, **74** 6-kernel theorems (**1,438** Linux **bwrap** fleet) in BonsaiStreamArc3.lean (including Theorem 74 proving r2_sync.py .br sidecar sync eliminates stale edge shadowing), and **11** live exports in **12.75 KB** (**13,058 B**) **lut6_core.wasm**.

5. Related work and convergence observations

System	ID / Venue	What it establishes	Reported cost	Difference here
LogicNets (Umuroglu et al., 2020)	2004.03021 · FPL 2020	Sparse quantized NN layers compiled to FPGA LUTs.	FPGA LUT / Fmax synthesis.	Fixed routing; drift needs Vivado re-PAR vs. 1-cycle Beneš+RAMB36E2.
DiffLogic (Petersen et al., 2022) & DWN (2024)	2210.08277 · 2410.11112	Differentiable Boolean gate & weightless LUT training.	Gate count & accuracy.	Static wires fail drift (52.34%) and suffer marginal chimeras (Prop. 3).
T-MAC (2024) & TeLLMe (Qiao et al., 2025)	2407.00088 · 2504.16266	Ternary / low-bit matmul via table lookup on CPU/FPGA.	Latency & FPGA power.	Accelerates matmul *inside* a model; does not bypass model runs.

System	ID / Venue	What it establishes	Reported cost	Difference here
LittleBit-1 / 2 (Lee et al., 2025; Lee & Kim, 2026)	2506.13771 · 2603.00042	0.1–0.58 bpw binary latent factorization (SVD/ITQ+QAT).	11.6×–31× size, PPL.	Unpacks to dense binary MACs & 0.0% code pass vs. 0-MAC Sub1Bit-LUT6 + MoL.
FrugalGPT (2023) & RouteLLM (2024)	2305.05176 · 2406.18665	Weak-to-strong LLM cascades & preference routers.	API dollar cost vs. quality.	No $C(z)$ gate (89.6% broken commits) and no LUT6 downward compiler.
GreenServ (2026) & VoltanaLLM (2025)	2601.17551 · 2509.04827	Energy-metered multi-LLM routing & frequency serving.	Metered GPU Wh (Zeus/NVML).	Optimizes GPU serving; our joules are est_j, compiling to LUT6.
Medusa (2024), EAGLE-2 (2024), HCSpec (2026)	2401.10774 · 2406.16858 · ACL 2026	Neural MLP/tree/cascade speculative drafters.	Wall-clock speedup.	Require draft FLOPs & weight reads vs. 0-FLOP uint64 LUT6 pages.
PLD (2023) & REST (He et al., 2024)	GitHub 2023 · 2311.08252	Zero-FLOP prompt n-gram & datastore suffix-trie draft.	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 streaming in tiny RAM.	Active RAM (4 GB for 70B).	Pure T=1 (1.0 tok/swp) + KV spill vs. AirMoL 2.74×–2.86× cut.
Gates & RCE-JEPA (Charlot, 2026) [same house]	research.openie.dev	Commit gates (11,428 pJ KV260 SOM), MoL p_i/c_i, RCE-JEPA.	KV260 SOM measured_j + est_j.	Same-house foundation; extended here with Beneš LUT6 compiler & AirMoL.

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

Internal validity. Downward crystallization pays off only when regime persistence ΔT exceeds Proposition 1's amortization threshold. On i.i.d. random bitstreams or per-step adversarial churn ($\Delta T = 1$), $p_0' \approx 0$ and the system falls back to 1.00 tok/sweep plus Rung 0 lookup overhead. On Qwen3.5-0.8B-Base cold prompts (Claim LA-7), CPU verification of $K + 1 = 4$ tokens reduced wall-clock speed (0.32 vs. 0.53 tok/s) despite cutting NVMe bytes (8.581 vs. 12.674 GB), showing that CPU batch verification requires post-wake draft acceptance (2.74 tok/sweep → 1.11 tok/s) or GPU/FPGA offload to beat T=1 latency.

Construct validity. 100.0% lossless parity guarantees bit-for-bit agreement with greedy target decoding; it does not raise the underlying capability ceiling of a 350M or 0.8B checkpoint. Likewise, Conformal-bounded $(1 - \alpha)$ bounds statistical miscoverage on the calibrated JEPA-EBRM manifold and never promotes neural proposals to Deterministic.

External validity. Foundation-model AirMoL, Expert-AirMoL, and coding benchmarks (Claims LA-7, LA-9, LA-11, LA-15, LA-17, LA-20, LA-21) were measured across eight open-weights checkpoints spanning 350M to **120B** parameters (LFM2.5-350M, Qwen3.5-0.8B-Base, LiquidAI/LFM3-1.7B, google/gemma-4-E2B-it, LiquidAI/LFM2.5-8B-A1B 8.47B, Devstral-Small-2-24B, Qwen/Qwen3.6-35B-A3B 34.66B, and **nvidia/Nemotron-3-Super-120B-A12B 120B**, plus LittleBit-2 0.55 bpw) on Apple M5 Max (128 GiB Unified Memory), including 32 official BigCodeBench v0.1.3 evaluations and 40 upstream SWE-bench Verified git-worktree evaluations across django/django, pytest-dev/pytest, psf/requests, and sympy/sympy. Unsharded cold-NVMe 70B rows in air_mol_nvme_baselines.csv remain labeled analytical baselines, while real 120B-A12B execution is measured directly in Claim LA-21.

Measurement validity. Physical DC current-shunt measured_j (TI INA3221 VDD_CPU_SOC_MSS and INA238 VIN on NVIDIA Jetson AGX Thor solely in Claim LA-24, board_metered = true) stays strictly separated from on-chip/OS reported_j counters (Linux RAPL, amdgpu PPT, NVML, Darwin L3 RUSAGE_INFO_V6 + L2 IOReport GPU Energy), whole-system macOS PMU deltas, and analytical/VCD est_j estimates (Claim LA-22). Yosys 0.71+ and Verilator 5.054 verify structural xck26 synthesis (13 single-cycle logic levels and 170 <= 189 IOBs in CE-gated P4-AXIS) and 320/320 CE parity, not commercial Vivado post-PAR 250 MHz timing closure or KV260 rail watts (measured_j = None for all xck26 RTL claims).

Not claimed. Not claimed here: physical KV260 bitstream loading or rail-metered joules; commercial Vivado post-PAR timing closure at 250 MHz; silicon energy leadership over commercial ASICs; promotion of ModelGenerated outputs to Deterministic; or full 500-instance SWE-bench Verified evaluation beyond the 40 real git-worktree evaluations across django, pytest, requests, and sympy in Claims LA-15 & LA-20.



Figure 12: Bonsai-AirMoL 1.75 bpw 27B streaming speculation, multi-layer Beneš ARC-AGI-2 synthesis, and official ARC-AGI-3 interactive RHAE (Claims LA-19 [74 thms / 1,438 fleet], LA-42..LA-44): (A) M3 Ultra 256 GB end-to-end Ternary-Bonsai-2-27B 1.619× GiB/tok cut (100% greedy parity) and 8.35× Thor DC shunt measured_j gain (Window 14); (B) Benes-ARC-LUT6 n=24 (20.83%) & n=64 (12.50%) with 100% SelAcc and 0% CHR@80; (C) official ARC-AGI-3 (8 games, 57 levels) 208.3× Wave-2 RHAE gain; and (D) 74 6-kernel theorems, 12.75 KB WASM, and edge .br sidecar invariant.

ID	Claim	Exact Disproof / Falsifier Condition
LA-F1	Claim LA-1 (.pcap + RTL)	run_realworld_domain.py fails 100.0% Verilator parity over 5,120 packets or misses the packet-2,500 VLAN drift.
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)	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)	run_frontier_benchmarks.py exceeds target false risk $\alpha = 0.02$ on shifted traffic or stamps Deterministic.
LA-F5	Claim LA-5 (Byte speculation)	run_code_corpus_benchmark.py yields < 100.0% byte parity vs. greedy Mamba3CodeLM on 48 unseen prompts.
LA-F6	Claim LA-6 (AirMoL 14.3M)	run_air_mol_benchmark.py under F_NOCACHE fails 100% parity or reads more bytes than AirLLM 4-Bit T=1.
LA-F7	Claim LA-7 (2026 HF AirMoL)	run_hf_lfm3_air_mol.py, run_hf_air_mol_pretrained.py, or run_hf_qwen35_air_mol.py diverges by ≥ 1 token.
LA-F8	Claim LA-8 (Full test suite)	pytest -v fails any of the 66 tests in tests/.
LA-F9	Claim LA-9 (8.5B/35B MoE)	run_hf_moe_8b_35b_air_mol.py fails 100% token parity or reads more NVMe bytes than AirLLM / Expert-AirLLM.
LA-F10	Claim LA-10 (Sub1Bit-LUT6)	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)	run_frontier_ai_bakeoff.py admits > 0% broken commits or admits a stale Wave-1 skill under Wave-2 spec drift.
LA-F12	Claim LA-12 (Metal MoE/DeltaNet)	run_metal_moe_benchmark.py exceeds 1e-4 FP32 kernel error or fails 100% token parity on 8B/35B.
LA-F13	Claim LA-13 (SWE-Mini Git-Repo)	run_repo_patch_benchmark.py admits > 0% broken patches, alters tests/, or misses Wave-2 R03 drift.
LA-F14	Claim LA-14 (OpenAI/MCP Agent)	test_phase26_metal_ioreport_mcp.py fails REST/SSE, MCP JSON-RPC 2.0, or .mol/ persistence.
LA-F15	Claim LA-15 (BCB + SWE-bench)	run_external_public_benchmarks.py fails Parquet SHA-256, leaks gold patches, or admits > 0% broken outputs.
LA-F16	Claim LA-16 (250 MHz KV260 RTL)	verify_pipelined_250mhz_rtl.py exceeds 20 logic levels, 189 IOBs, 0 DSP48s, or fails 256/256 CPI = 1.00 parity.
LA-F17	Claim LA-17 (Rung 3d <think>)	run_rung3d_thinking_ceiling.py admits > 0% broken commits, calls cloud APIs, or fails to beat 87.50%.
LA-F18	Claim LA-18 (CubeCL INT4/Delta)	bench_cubecl_moe_deltanet exceeds 1e-4 FP32 error on CPU/Metal or achieves < 25× Metal speedup.
LA-F19	Claim LA-19 (6-Kernel Lean 4.36.0)	verify_lean_proofs.py or lake check --paranoid fails any of 6 kernels, leaves ≥ 1 sorry/ofReduceBool, or adds custom axioms.
LA-F20	Claim LA-20 (4-Repo SWE-bench)	run_multirepo_swebench_benchmark.py leaks patch, tests < 4 repos, or admits > 0% broken patches.
LA-F21	Claim LA-21 (120B Hybrid MoE)	run_120b_hybrid_moe_benchmark.py fails 120B-A12B local run, admits > 0% broken commits, or misses W2 .mol/.
LA-F22	Claim LA-22 (CE-Gated VCD Power)	verify_ce_gated_vcd_power.py fails 320/320 CE parity, exceeds 189 IOBs, or gets < 3.5× idle reduction.
LA-F23	Claim LA-23 (Rust mol-agent-rs)	mol_agent --self-test-bench requires Python for /v1 or /mcp, fails --mcp-stdio, or gets < 10× speedup.
LA-F24	Claim LA-24 (6-Host + Thor Shunt)	aggregate_fleet_measured_j.py mixes measured_j/reported_j, exceeds 1% VIN drift on Thor, or fails 4/4 ABBA on W2/W3/W5.
LA-F25	Claim LA-25 (4-Vendor CubeCL GPU)	aggregate_multivendor_cubecl_gpu.py exceeds 1e-4 FP32 error or gets < 10× GPU speedup on Apple/AMD/NVIDIA/Intel.
LA-F26	Claim LA-26 (lut6-core & Yosys Miter)	sys/lut6-core fails cargo test, CompactLut6Page $\neq$ 4,096 B, MCU ELFs fail to link, or Yosys miter fails 64/64.
LA-F27	Claim LA-27 (DiffLUT-Net & Wobble)	run_dual_reality_lut6_eval.py gets < 55% LUT6 pruning, < 85% noisy hit, > 0% corruption, or > 0 wobble $\lvert S\rvert \geq 5$ noise.
LA-F28	Claim LA-28 (Thor Shunt & Alif E3)	Stamps board_metered = true off-Thor, gets < 10× Thor bit-sliced CpJ, or exceeds 10 μs scalar on OpenMV-AE3.
LA-F29	Claim LA-29 (Oct–Nov 2026 Looped)	run_nov2026_models_lut6_benchmark.py diverges by ≥ 1 greedy token or exceeds 64 bits on Inkling-2-Mini / Dory1.0-4B.
LA-F30	Claim LA-30 (XC7A200T M.2 Port)	aller-a7-xc7a200t exceeds 134.6k LUTs / 365 BRAMs, diverges vs. XCK26, or stamps board_metered.
LA-F31	Claim LA-31 (T-MAC Tail & const)	mol_skills_tmac_tail.lut6 wastes > 0 pad bytes, uses runtime MUL, fails 100% skill/OOD, or fails rustc const gate.
LA-F32	Claim LA-32 (Two-Tier Telemetry)	eval_raw_unrolled4 $\neq$ eval_raw or T-MAC fails to beat INT8 Micro-MLP CpJ on Thor shunt or cycles on Alif E3.
LA-F33	Claim LA-33 (Antisym-40 & Aeneas)	lookup_tmac_antisym41 diverges on 40.5k pairs, fails > 1.0× Zen 5 CpJ, or Lut6Verify_Aeneas.lean fails Lean check.
LA-F34	Claim LA-34 (SEU & openXC7 PAR)	Nucleus SECDED admits > 0% silent err at BER=1%, or nextpnr-xilinx fails 250 MHz timing / .bit on xc7a200tfbg484-2.

ID	Claim	Exact Disproof / Falsifier Condition
LA-F35	Claim LA-35 (Lupanov/TreeLUT/WASM)	Lupanov, TreeLUT, or KAN-LUT fails minterm parity, or lut6_core.wasm exceeds 16 KB or fails V8 CRC32/SECDED.
LA-F36	Claim LA-36 (Wearable μLLMs vs. MoL)	Wearable-MoL R0 exceeds 16 KB watch SRAM, records > 0% prefix poisoning, or gets < 90% wearable reflex acc.
LA-F37	Claim LA-37 (LastExam Calibration)	Bypasses cais/hle gating, leaks prompt text, records > 5% CHR@80, or gets < 3× RMS-CE cut vs. 35B.
LA-F38	Claim LA-38 (66 Thms & Canary Guard)	WearableLastExam.lean fails 6-kernel check, admits hle:3r2s canary, or gets < 100× Thor / < 1000× M55 gain.
LA-F39	Claim LA-39 (Bonsai-2-27B EntroMCU/FWHT)	Bonsai-2-27B audit finds c = 3, |bias+scale| > 0, > 0 EntroMCU roundtrip error, or Window 12 fails to beat Window 13 Thor measured_j.
LA-F40	Claim LA-40 (System-1 Decider → LUT6)	Compiled System-0 LUT6 exceeds 4,096 B, admits > 0% OOD false actions, or fails >= 1,000× W2 speedup vs. Mapika/decider-0.8b.
LA-F41	Claim LA-41 (ARC-AGI-2/MiniF2F/CritPt)	Admits a train-inconsistent ARC-AGI-2 rule, admits >= 1 unsound MiniF2F proof, persists raw benchmark strings, or records > 0% CritPt CHR@80.
LA-F42	Claim LA-42 (Bonsai-AirMoL 27B Stream)	Bonsai-AirMoL diverges by ≥ 1 greedy token vs. MLX T=1, incurs > 0 EntroMCU-28 pread errors, or gets < 5× Window 14 Thor measured_j gain.
LA-F43	Claim LA-43 (Multi-Layer Benes-ARC-LUT6)	2/3-layer Benes-ARC-LUT6 fails to strictly beat 1-layer on n=24 or n=64, leaks test grids, or records > 0% CHR@80.
LA-F44	Claim LA-44 (ARC-AGI-3 RHAE & 74 Thms)	Peeks at ARC-AGI-3 game .py source, persists raw frames, fails Wave-2 RHAE > Wave-1/Random, fails 74 6-kernel proofs, or exceeds 16 KB lut6_core.wasm.

7. References

Every reference is a verified primary publication, technical report, or model card. Works marked [same house] are by David Charlot, Open Interface Engineering (https://research.openie.dev/), cited for continuity and disclosure.

- Bacellar, A., et al. (2024). Differentiable Weightless Neural Networks. *arXiv:2410.11112*.
- Beneš, V. E. (1965). *Mathematical Theory of Connecting Networks and Telephone Traffic*. Academic Press.
- 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*.
- Charlot, D. (2026). Metered Commit Gates: Energy-to-Correct-Completion for Agentic Systems (Part I & Part II: Mixture of Limits). *Open Interface Engineering*, https://research.openie.dev/papers/gates/ [same house].
- Charlot, D. (2026). RCE-JEPA: World Coding, with JEPA as Premise. *Open Interface Engineering*, https://research.openie.dev/papers/rce-jepa/ [same house].
- Charlot, D. (2026). RCE Compile & Task Satiation vs. Aggregate Rebound. *Open Interface Engineering*, https://research.openie.dev/ [same house].
- Chen, L., Zaharia, M., and Zou, J. (2023). FrugalGPT. *arXiv:2305.05176*.
- Dao, T., and Gu, A. (2024). Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality. *ICML 2024*, *arXiv:2405.21060*.
- Furao, S., and Hasegawa, O. (2006). An incremental network for on-line unsupervised classification and topology learning (SOINN). *Neural Networks*, 19(1):90–106.
- Gibbs, I., and Candès, E. (2021). Adaptive Conformal Inference Under Distribution Shift. *NeurIPS 2021*, *arXiv:2106.00170*.
- GreenServ (2026). Energy-Efficient Context-Aware Dynamic Routing for Multi-Model LLM Inference. *ICPE 2026*, *arXiv:2601.17551*.
- HCSpec (2026). Two-Tier Horizontal Cascade Speculative Decoding. *ACL 2026*, https://doi.org/10.18653/v1/2026.acl-long.353.
- He, Z., Zhong, Z., Cai, T., Lee, J. D., and He, D. (2024). REST: Retrieval-Based Speculative Decoding. *NAACL 2024*, *arXiv:2311.08252*.
- 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*.
- Horowitz, M. (2014). Computing's Energy Problem (and What We Can Do About It). *IEEE ISSCC 2014*.
- Kuon, I., and Rose, J. (2007). Measuring the Gap Between FPGAs and ASICs. *IEEE TCAD*, 26(2):203–215.
- Landauer, R. (1961). Irreversibility and Heat Generation in the Computing Process. *IBM J. Res. Dev.*, 5(3):183–191.
- LeCun, Y. (2022). A Path Towards Autonomous Machine Intelligence. *Meta AI Research*.

- LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., and Huang, F. J. (2006). A Tutorial on Energy-Based Learning. MIT Press.
- Google Gemma Team (2026). `google/gemma-4-E2B-it` On-Device Hybrid Instruction Model Card. https://huggingface.co/google/gemma-4-E2B-it.
- 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*.
- 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*.
- Lee, T. T. (1988). Nonblocking Copy Networks for Multicast Packet Switching. *IEEE JSAC*, 6(9):1455–1467.
- 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.
- Leviathan, Y., Kalman, M., and Matias, Y. (2023). Fast Inference from Transformers via Speculative Decoding. *ICML 2023*, *arXiv:2211.17192*.
- 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*.
- Liquid AI (2026a). `LiquidAI/LFM2.5-350M` and `LiquidAI/LFM2.5-8B-A1B` Hybrid & Sparse-MoE Model Cards. https://huggingface.co/LiquidAI/LFM2.5-8B-A1B.
- Liquid AI (2026b). `LiquidAI/LFM3-1.7B` Tri-Hybrid Open-Weights Model Card (October 23, 2026). https://huggingface.co/LiquidAI/LFM3-1.7B.
- Liu, W., Wang, X., Owens, J., and Li, Y. (2020). Energy-based Out-of-distribution Detection. *NeurIPS 2020*, *arXiv:2010.03759*.
- lyogavin (2023). AirLLM: Layer-wise LLM Inference on Memory-Constrained Hardware. https://github.com/lyogavin/airllm.
- Ma, S., et al. (2024). The Era of 1-bit LLMs (`BitNet b1.58`). *arXiv:2402.17764*.
- Ong, I., et al. (2024). RouteLLM: Learning to Route LLMs with Preference Data. *ICLR 2025*, *arXiv:2406.18665*.
- Petersen, F., Borgelt, C., Kuehne, H., and Deussen, O. (2022). Deep Differentiable Logic Gate Networks. *NeurIPS 2022*, *arXiv:2210.08277*.
- Qiao, Y., et al. (2025). TeLLMe: An Energy-Efficient Ternary LLM Accelerator on Edge FPGAs. *arXiv:2504.16266*.
- Qwen Team (2026a). `Qwen/Qwen3.5-0.8B-Base` Hybrid Gated DeltaNet + Attention Model Card (February 2026). https://huggingface.co/Qwen/Qwen3.5-0.8B-Base.
- Qwen Team (2026b). `Qwen/Qwen3.6-35B-A3B` Sparse-MoE Hybrid Gated DeltaNet + Attention Model Card (2026). https://huggingface.co/Qwen/Qwen3.6-35B-A3B.
- Saxena, A. (2023). Prompt Lookup Decoding. https://github.com/apoorvumang/prompt-lookup-decoding.
- Simon, H. A., and Kadane, J. B. (1975). Optimal Problem-Solving Search: All-or-None Solutions. *Artificial Intelligence*, 6(3):235–247.
- Thompson, C. D. (1978). Generalized Connection Networks for Parallel Processor Intercommunication. *IEEE Trans. Comput.*, C-27(12):1119–1125.
- Umuroglu, Y., Akhauri, Y., Fraser, N. J., and Blott, M. (2020). LogicNets. *FPL 2020*, *arXiv:2004.03021*.
- VoltanaLLM (2025). Energy-Efficient and SLO-Aware Disaggregated LLM Serving. *arXiv:2509.04827*.
- Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer.
- Waksman, A. (1968). A Permutation Network. *Journal of the ACM*, 15(1):159–163.
- Wei, J., et al. (2024). T-MAC: CPU Renaissance via Table Lookup for Low-Bit LLM Deployment. *arXiv:2407.00088*.
- 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*.
- 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*.
- Moura, L. de, and Ullrich, S. (2021). The Lean 4 Theorem Prover and Programming Language. *CADE-28*, *LNCS* 12699:625–635.
- NVIDIA (2026). `nvidia/Nemotron-3-Super-120B-A12B` Hybrid Mamba-Attention Sparse-MoE Open-Weights Model Card. https://huggingface.co/nvidia/Nemotron-3-Super-120B-A12B-BF16.
- Mistral AI (2025). `mistralai/Devstral-Small-2-24B-Instruct-2512` Open-Weights Software Engineering Model Card. https://huggingface.co/mistralai/Devstral-Small-2-24B-Instruct-2512.

- 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*.
- 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*.
- 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*.
- Sanyal, A., Chinchali, S. P., and Datta, G. (2026). EntroMCU: A Hardware-Algorithm Vision for LLM Inference on Microcontrollers. *IEEE ISVLSI 2026*.
- Linial, N., Mansour, Y., and Nisan, N. (1993). Constant Depth Circuits, Fourier Transform, and Learnability. *Journal of the ACM*, 40(3):607–620.
- Thinking Machines Lab (2026). `thinkingmachines/Inkling-2-Mini` Open-Weights Looped Multimodal MoE Model Card (November 20, 2026). https://huggingface.co/thinkingmachines/Inkling-2-Mini.
- NVIDIA Nemotron Labs (2026). `nvidia/NVIDIA-Nemotron-Labs-Dory1.0-4B-BF16` Looped nGPT Model Card (October 2026). https://huggingface.co/nvidia/NVIDIA-Nemotron-Labs-Dory1.0-4B-BF16.
- 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*.
- 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.
- Bührer, T., Plesner, A., Aczel, T., and Wattenhofer, R. (2026). BitLogic: Scaling Continuous Logic Gate Networks via Lookup Tables. *ETH Zürich*, *arXiv:2602.07400*.
- Khataei, A., and Bazargan, K. (2025). TreeLUT: An Efficient Architecture for Gradient Boosted Decision Trees Using FPGA Lookup Tables. *arXiv:2501.01511*.
- Lupanov, O. B. (1958). On a Method of Circuit Synthesis [Ob odnom metode sinteza skhem]. *Izvestiya VUZov, Radiofizika*, 1(1):120–140.
- Ho, S., and Protzenko, J. (2022). Aeneas: Rust Verification by Functional Translation (`AeneasVerif/charon` & `AeneasVerif/aeneas`). *ICFP 2022*, *PACMPL* 6(ICFP):711–741.
- 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*.
- Phan, L., Gatti, A., Han, Z., Li, N., Hu, J., Zhang, H., et al. (2026). Humanity's Last Exam (`lastexam.ai`, `cais/hle`). *Nature*, 2026, *arXiv:2501.14249*.
- Cheng, W., Chen, T., Helwani, K., Srinivasan, S., Zettlemoyer, L., and Gollakota, S. (2026). Micro Language Models Enable Instant Responses (`Sensente/uLM-28M`). *EMNLP 2026*, *arXiv:2604.19642*.
- M-A-P Team, Du, X., Yao, Y., Ma, K., Wang, B., Zheng, T., et al. (2025). SuperGPQA: Scaling LLM Evaluation across 285 Graduate Disciplines (`m-a-p/SuperGPQA`). *arXiv:2502.14739*.
- Liu, Z., Chang, E., Li, C., Sreedhar, M., Wu, D., et al. (2026). MobileLLM-R1: Exploring the Limits of Sub-Billion Language Model Reasoners with Open Training Recipes. *ICLR 2026*, *arXiv:2509.24945*.
- PrismML (2026). `prism-ml/Ternary-Bonsai-2-27B`: Hadamard-Folded Native Ternary `27B` Hybrid Gated DeltaNet Model Card & Whitepaper. https://huggingface.co/prism-ml/Ternary-Bonsai-2-27B-mlx-2bit.
- Mapika (2026). `Mapika/decider-0.8b`: Typed System-One Decisions with Calibrated Probabilities in One Forward Pass. https://huggingface.co/Mapika/decider-0.8b.
- 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`). *arXiv:2505.11831*.
- ARC Prize Foundation (2026). ARC-AGI-3 Interactive Reasoning Benchmark & Engine (`arcprize/ARC-AGI-3-Agents`, `arc-agi 0.9.10` & `arcengine 0.9.3`, `MIT` License). https://arcprize.org/arc-agi/3/.
- 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*.
- CritPt Collaboration (2026). CritPt: Probing the Critical Point of AI Reasoning in Research-Level Physics (`CritPt-Benchmark/CritPt`). https://huggingface.co/datasets/CritPt-Benchmark/CritPt.

Appendix A. Reproducibility

All claims, Lean `4.36.0` proofs, RTL simulations, and CSV/JSON logs reproduce from `openIE-dev/lut-adapt-ai`:

```
PYTHONPATH=src pytest -v                             # 38/38 tests pass (Lean4, RTL, Shunt, 4-GPU, 120B, SWE)
(cd proofs && lake build) && python3 scripts/verify_lean_proofs.py # Claim LA-19 27 Lean 4.36.0 thms (0 sorry)
python3 scripts/run_multirepo_swebench_benchmark.py # Claim LA-20 4-repo SWE-bench Verified (24 evals, 2.50x)
python3 scripts/run_120b_hybrid_moe_benchmark.py    # Claim LA-21 Measured Nemotron-3-120B-A12B + Devstral-24B
python3 scripts/verify_ce_gated_vcd_power.py        # Claim LA-22 CE-gated 250 MHz xck26 VCD power (4.10x idle)
(cd sys/rust && cargo run --release --bin mol_agent -- --self-test-bench) # Claim LA-23 Rust mol-agent-rs
python3 scripts/run_external_public_benchmarks.py  # Claim LA-15 BigCodeBench v0.1.3 + SWE-bench SymPy MCP
python3 scripts/verify_pipelined_250mhz_rtl.py   # Claim LA-16 4-stage 13-level 250 MHz xck26 RTL
python3 scripts/run_rung3d_thinking_ceiling.py   # Claim LA-17 Rung 3d Qwen3.6-35B <think> (91.67%/93.75%)
(cd sys/rust && cargo run --release --bin bench_cubecl_moe_deltanet) # Claim LA-18 CubeCL INT4 MoE/DeltaNet
python3 scripts/run_collapse_curve.py && python3 scripts/run_realworld_domain.py # Claim LA-1 .pcap + RTL
(cd sys/rust && cargo run --release) && (cd sys/zig && zig build test) && (cd sys/go && go test -bench=.)
julia sys/julia/mol_scientific.jl                # Claim LA-2 Polyglot AD + zero-alloc suite
python3 scripts/run_frontier_benchmarks.py && python3 scripts/verify_paged_bram_rtl.py # Claims LA-3..4
python3 scripts/run_code_corpus_benchmark.py && python3 scripts/run_air_mol_benchmark.py # Claims LA-5..6
python3 scripts/run_hf_lfm3_air_mol.py && python3 scripts/run_hf_moe_8b_35b_air_mol.py # Claims LA-7, 9
python3 scripts/run_sub1bit_lut_benchmark.py && python3 scripts/run_frontier_ai_bakeoff.py # Claims LA-10..11
python3 scripts/run_metal_moe_benchmark.py && python3 scripts/run_repo_patch_benchmark.py # Claims LA-12..13
python3 scripts/mol_agent_server.py --port 8090 # Claim LA-14 Drop-in OpenAI (/v1) & MCP .mol/ agent
```