Warden-S1: A Sub-1 GiB, 128K-Context Local System-One Verifier for Long-Horizon Coding Agents

Hybrid Gated DeltaNet + GQA Multi-Head Verification, Counterfactual Calibration, and 10-Session Laptop Multiplexing

Ryan Joshua Z. Gapac
ORCID: 0009-0004-2591-0219 · DOI: 10.5281/zenodo.24494405
mail@ryanjoshuagapac.com · https://ryanjoshuagapac.com · November 24, 2026

Abstract. Autonomous coding agents routinely accumulate 64K to 128K tokens per session while running concurrently across multiple repositories on developer laptops. Verifying their tool calls, secret-handling boundaries, repository rules, and completion claims with cloud judges hits three structural walls: payload ceilings around 33K tokens, recurring API latency and cost, and data-residency risk. Meanwhile, dense small transformers exceed 7 to 14 GiB of KV state at 64K to 128K context, exhausting 16 GB unified-memory laptops on a single session. We introduce **Warden-S1**, a 752M-parameter local System-One verification model built by surgically stripping the vision tower and multi-token prediction heads from `Qwen3.5-0.8B` (exact `0.000000` logit parity), yielding an 18 Gated DeltaNet + 6 GQA hybrid backbone paired with five zero-decode heads (`noul` , `choice` , `score` , `locus` , and `salience`). Trained on **WardenS1-Bench** (4,200 trajectories, 225,650,644 tokens, 67,200 probes, 2,100 counterfactual twin pairs) and polished via a Stage 3.5 zero-leakage `<4K` + `4K-128K` replay curriculum, `Warden-S1-752M-MLX-Q4` (415.5 MiB on disk) achieves `0.9957` Noul AUROC (`0.9964` BF16), `0.8726` Choice F1, `0.0175` ECE, `0.9115` counterfactual net sensitivity, `0.9297` Locus Top-1 recall, and `0.9998` Salience AUROC across 4K to 128K tokens. On Apple M5 Metal, a singleton weight-pinned daemon multiplexes 10 concurrent 64K to 98K sessions inside `858.3 MiB` peak active residency (`< 1 GiB`) with `20/20` BPE-boundary-healed `O(Delta L)` delta hits (`18.76x` warm turn speedup on M5, `18.92x` on A100) and `104.4 ms` p50 L2 host-RAM swaps (`22.5x` faster than naive copies). In a 100-request stratified head-to-head against commercial Cloud `Jev` , local `Warden-S1` wins matched Noul AUROC (`0.9933` vs. `0.9449`), accuracy (`96.1%` vs. `84.9%`), Brier (`0.0270` vs. `0.0998`), and abstention overconfidence (`0.0%` vs. `100.0%`) while accepting `100/100` requests up to 117,899 tokens where Cloud `Jev` rejects `34%` past 33K. Stage 3.5 also lifts out-of-distribution `pi-warden` static `judge-bench` stuck accuracy to `85.0%-87.5%` and `rules-bench` recall at 0.5 to `41.9%` (`7.75x` Stage 2). A zero-PyTorch `461.3 MiB` sharded `ONNX-Int4` build brings the same five-head verifier to Linux, Windows, Node, and browser WebGPU (`0.9923` Noul AUROC, `max |Delta p| = 6.85e-7` CPU delta equivalence).

1 Introduction

Long-horizon coding agents fail in ways that static linters cannot see: a model reads an untrusted web page 40,000 tokens earlier and later interpolates a local secret into an outbound shell command, weakens an assertion to manufacture a passing test run, oscillates between failing edits, or silently ignores a repository rule stated in `AGENTS.md` [1], [2], [3], [4], [5]. Runtime guardrails such as `pi-warden` pair fast deterministic checks (`System-0`) with semantic verification (`System-1`) so unsafe actions are steered inline rather than blocking the human operator on every step.

Cloud System-One judges [6] work well on short snippets, but production multi-session coding exposes three operational bottlenecks: (1) hard context ceilings near `33K` tokens (`~100 KiB`), causing silent truncation or `HTTP 400` rejections on mature sessions; (2) recurring per-turn upload latency and billing across 10+ concurrent agent workspaces; and (3) privacy exposure when proprietary diffs and credentials must leave the developer workstation.

Running a dense sub-1B verifier locally does not solve the problem either. A 28-layer full-attention transformer such as `Qwen3-0.6B` [7] allocates `14.3 GiB` of BF16 KV cache at `128K` tokens (`4.4 GiB` even with 4-bit KV quantization), violating a strict `< 1 GiB` background residency budget on a 16 GB laptop.

Contributions. We present a complete, end-to-end local verification system:

1. **Backbone arena and exact weight surgery:** Benchmarking four sub-1B hybrids and dense transformers under sandwiched timing and true GPU allocator residency, stripping vision and MTP modules from `Qwen3.5-0.8B` [8] (`873.4M -> 752.4M` , `0.000000` logit parity) to hold 128K live Metal residency at `855.3 MiB` base (`867.4 MiB` with heads).
2. **Zero-decode five-head System-One architecture:** Jointly producing calibrated `noul` , `choice` , `score` , `locus` span localization, and `salience` compaction

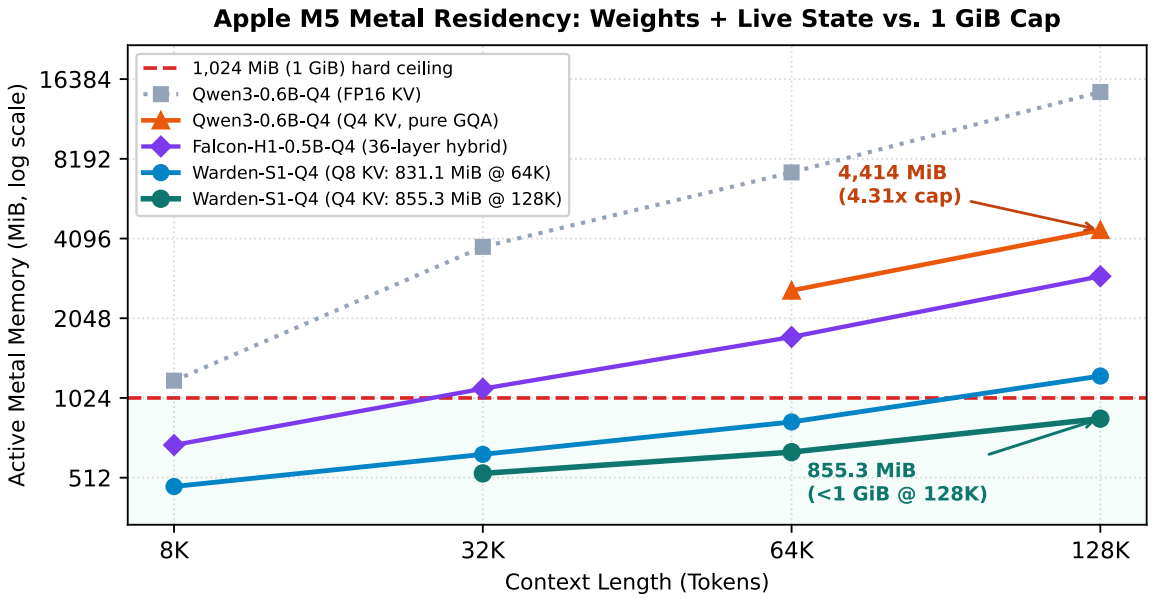


Figure 1: Measured live Apple M5 Metal residency (4-bit weights plus KV and recurrent state) from 8K to 128K tokens. `Warden-S1-752M` (`18` Gated DeltaNet + `6` GQA) stays strictly under the `1,024 MiB` ceiling at both 64K (`831.1 MiB` Q8 KV) and 128K (`855.3 MiB` Q4 KV), whereas dense `Qwen3-0.6B` hits `4,414 MiB` even with 4-bit KV.

scores in a single forward pass with `O(Delta L)` incremental prefix reuse and `0.0` cross-probe state mutation.

3. **WardenS1-Bench and Stage 3.5 replay polish:** A `225.7M`-token long-horizon corpus (`10,368` held-out test probes, `2,100` counterfactual twin pairs, zero repository overlap) combined with a zero-leakage `<4K` replay mix that preserves `0.9957` 4K-128K AUROC while multiplying out-of-distribution short-schema `rules-bench` recall by `7.75x`.
4. **Multi-session `< 1 GiB` runtime and portable `ONNX-Int4`:** Singleton Apple M5 Metal multiplexing across 10 concurrent 64K-98K sessions in `858.3 MiB` with 2-token BPE boundary healing, plus a `461.3 MiB` zero-PyTorch sharded `ONNX-Int4` package for Linux, Windows, Node, and browser WebGPU.

2 Backbone Arena & Weight Surgery

We evaluated four Apache-2.0 or permissive sub-1B architectures on NVIDIA A100-SXM4-40GB (`1,352 GB/s`, `227.7 TFLOPS` BF16 measured) and Apple M5 16 GB Metal (`78.0 GB/s`, `9.77 TFLOPS` FP16 measured, `mlx 0.32.4`) using sandwiched `control -> candidate -> control` runs (`>= 6` samples per arm, control drift `<= 0.16%`) and true allocator residency (`mx.get_active_memory()` and `torch.cuda.memory_allocated()`).

Table 1 summarizes the hardware arena. Stripping the unused vision encoder (`100.6M` params) and multi-token prediction head (`20.5M` params) from `Qwen3.5-0.8B` removes `121.0M` dead parameters (`242 MiB` BF16) while verifying `max |Delta logit| = 0.000000` across text inputs. Because 18 of 24 layers use Gated DeltaNet linear recurrence [9] (`18.6 MiB` fixed recurrent + convolutional state independent of context length) and only 6 layers hold GQA KV caches, live state grows `9.2x` slower than a dense 28-layer model. Both `Falcon-H1-0.5B` [10] and `LFM2-700M` [11] failed the 64K/128K needle-in-a-trajectory gate.

Backbone (Q4)	Params	M5 64K	M5 128K	Gate
Warden-S1 (18 DN+6 GQA)	752M	**831 MiB**	**855 MiB**	**PASS**
`Qwen3-0.6B` (28 GQA)	599M	2,606 MiB	4,414 MiB	FAIL (>1G)
`Falcon-H1` (36 M2+GQA)	522M	945 MiB	981 MiB	FAIL (0/4)
`LFM2-700M` (10 C+6 GQA)	743M	854 MiB	878 MiB	FAIL (neg)

Table 1: Apple M5 Metal live residency (4-bit weights + Q8 KV at 64K / Q4 KV at 128K) and long-horizon needle gate.

3 Zero-Decode Multi-Head Architecture & `O(Delta L)` Probing

Given a coding-agent trajectory serialized into chunked turns (`C` chunks, `L <= 131,072` tokens), `Warden-S1` runs a single prefix forward pass and caches the 18-layer Gated DeltaNet state (`18.6 MiB`) alongside the 6-layer GQA KV cache (`216 MiB` Q4 KV at 128K; `408 MiB` Q8 KV at 64K). On subsequent turns, only the appended delta `Delta L` (`~1K-2K` tokens) is processed.

Five lightweight probe heads (`11.1 MiB` FP32 total) operate over chunk representations and short query encodings without autoregressive token decoding:

Model / Checkpoint	Weights	Noul AUC	ECE	Choice F1	Score rho	CF Net	Locus R@1	Sal AUC
`Warden-S1-752M` (S3.5 BF16 Cal)	1,458 MiB	**0.9964**	**0.0126**	**0.8726**	0.8784	**0.9174**	0.9219	**0.9998**
`Warden-S1-752M-MLX-Q4` (S3.5)	**415.5 MiB**	**0.9957**	**0.0175**	**0.8726**	**0.8814**	**0.9115**	**0.9297**	**0.9998**
`Warden-S1-752M-ONNX-Int4`	**461.3 MiB**	**0.9923**	0.0324	0.7846	**0.8927**	**0.9259**	0.9167	**1.0000**
`Warden-S1-752M` (No-CF Ablation)	1,458 MiB	0.9801	0.0096	0.6174	0.7302	0.7507	0.6265	0.9574
Zero-shot `Qwen3-1.7B`	3,440 MiB	0.6492	0.0754	0.3827	0.0000	0.2191	0.0733	0.7796
Zero-shot `Qwen3.5-0.8B`	1,665 MiB	0.5663	0.3451	0.1465	0.0000	0.0059	0.0455	0.7797

Table 2: Held-out `WardenS1-Bench` evaluation across all five heads. Stage 3.5 `MLX-Q4` (`415.5 MiB` on disk) matches BF16 within `0.0007` AUROC (`0.00271` overall mean probability delta on 32-state validation parity; `0.9844` Locus Top-1 agreement).

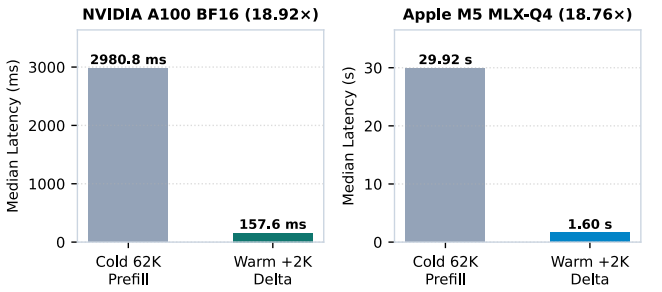


Figure 2: Sandwiched `O(Delta L)` latency (`60K` cached + `2K` delta vs. cold `62K` prefill): `18.92x` speedup on A100 (`157.6 ms` vs. `2,981 ms`) and `18.76x` on Apple M5 (`1.59 s` vs. `30.4 s`).

- **`noul` (Epistemic Binary):** Outputs calibrated probability `p in [0, 1]` with explicit abstention supervision (`p = 0.50`) when evidence is insufficient.
- **`choice` (Multi-Option):** Scores arbitrary named criteria strings against the cached trajectory representation.
- **`score` (Ordered Rubric):** Predicts a 5-bin calibrated severity/quality distribution mapped to `[0, 1]`.
- **`locus` (Evidence Pointer):** Attends over all trajectory chunks to identify the exact turn containing a rule violation, tainted input, or shortcut edit.
- **`salience` (Compaction Scorer):** Assigns per-chunk retention probabilities so upstream harnesses can mask stale tool outputs without dropping security-relevant context.

Evaluating `K = 32` independent probes against a frozen `64K` trajectory achieves `15.32 ms/probe` (`5.41x` batch speedup over sequential queries) with strictly `0.0` cross-probe state mutation (non-mutating cache fork).

4 WardenS1-Bench & Stage 3.5 Curriculum

Existing agent benchmarks [12], [13] rarely combine 50K-128K token depth, calibrated abstention, and minimal counterfactual pairs. We constructed **WardenS1-Bench** from permissively licensed open-source repositories using Union-Find fork clustering to guarantee `0` repository overlap across `train` (`3,120` states), `val` (`432` states), and held-out `bench` (`648` states, `10,368` probes):

- **Scale and depth:** `4,200` states totaling `225,650,644` exact `Qwen3.5` tokens (`5,592` min, `45,186` median, `53,726.3` mean, `127,892` max; `47.14%` of all states and `55.56%` of held-out `bench` states exceed `50K` tokens).
- **Six balanced domains:** `action_guard`, `taint_guard`, `repo_rules`, `done_verify`, `trajectory_health`, and `s1_general` (`700` states each).
- **Counterfactual twin pairs & abstention:** `2,100` verified single-chunk counterfactual pairs (`8,700` target probes flipped, `24,900` control probes invariant) plus `8,484` unanswerable abstention probes supervised at `0.50`.

Stage 2 + Stage 3.5 Training. Following `4K -> 48K -> 128K` curriculum training with counterfactual contrast loss (`lambda_cf = 2.2`), Stage 3.5 addresses short-schema transfer (`<4K` single-edit JSON payloads used by `pi-warden` static guards) via a 32-step L4 polish (`1,942 s`, `6.47 GiB` peak VRAM) mixing `50%` zero-leakage synthetic `<4K` single-edit/oscillation/pre-stop pairs (`512` pairs; verified `0` SHA-256 and `0` 13-gram overlap against `1,319,494` held-out benchmark 13-grams) with `50%` `4K-128K` `WardenS1` replay pairs (`rule_violation` choice weight `3.0`), followed by per-head post-hoc temperature scaling [14] (`T_noul = 1.039`, `T_choice = 1.139`, `T_score = 1.220`, `T_locus = 2.119`, `T_sal = 0.979`).

5 Experimental Results

5.1 Held-Out WardenS1-Bench (4K to 128K)

Table 2 reports held-out `WardenS1-Bench` results (Stage 3.5 BF16 and MLX-Q4 evaluated on the stratified 64-state / 1,032-probe slice; baselines and No-CF ablation on the full 648-state / 10,368-probe test split; `ONNX-Int4` on the 36-state parity slice).

Across length buckets, Stage 3.5 `Warden-S1-752M-MLX-Q4` shows zero long-context collapse: `0.9957` (`4k-16k`), `0.9969` (`16k-48k`), `0.9957` (`50k-80k`), and `0.9947` (`80k-128k`) Noul AUROC, with `0.9500` Locus Top-1 recall in `80k-128k`. Counterfactual twin training is critical: removing counterfactual contrast drops Choice F1 from `0.8726` to `0.6174` and Locus Top-1 from `0.9297` to `0.6265`.

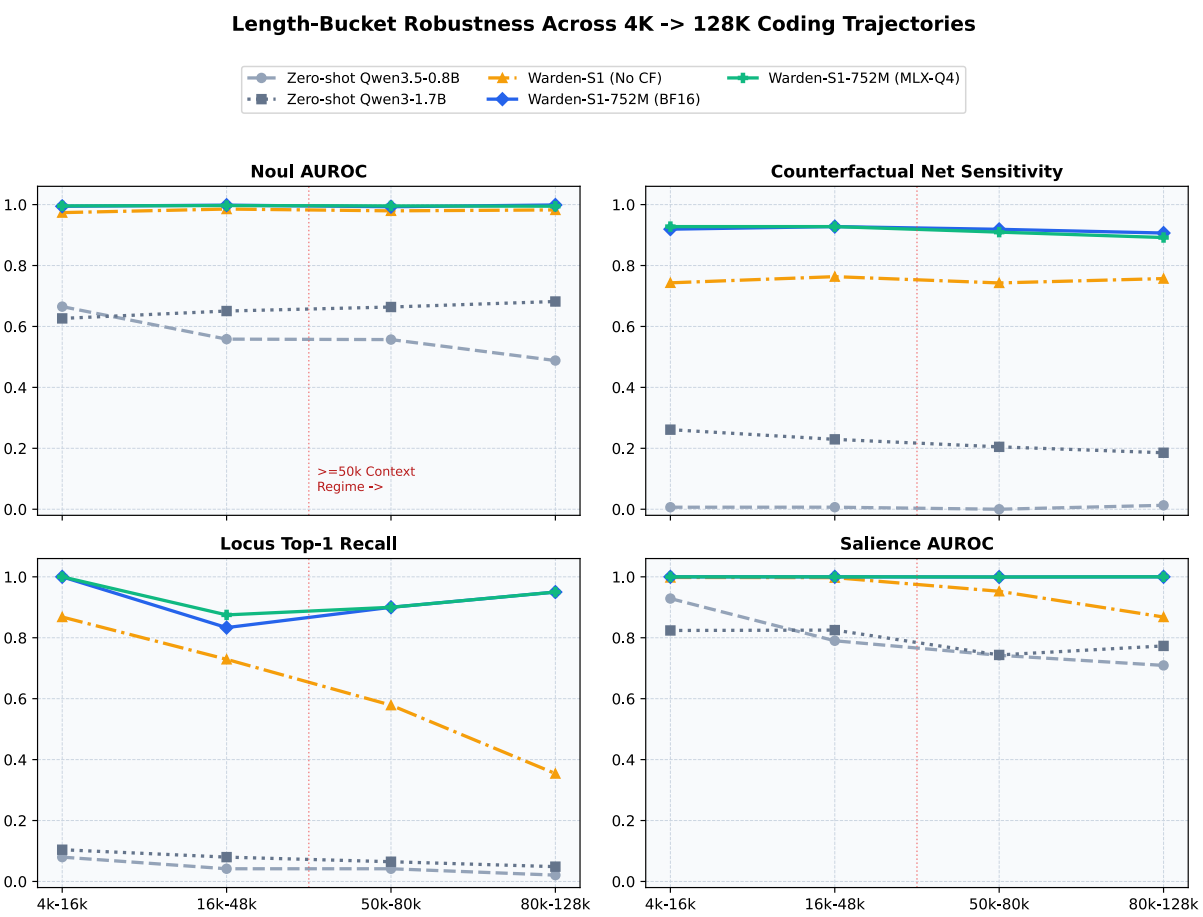


Figure 3: Length-bucket robustness from `4K` to `128K` tokens. Unlike zero-shot and No-CF ablations, Stage 3.5 `Warden-S1` holds flat Noul AUROC (`0.9957` at `4K-16K` to `0.9947` at `80K-128K`), Counterfactual Net Sensitivity (`0.8919-0.9278`), Locus Top-1 (`0.8750-1.0000`), and Salience AUROC (`0.9994-1.0000`).

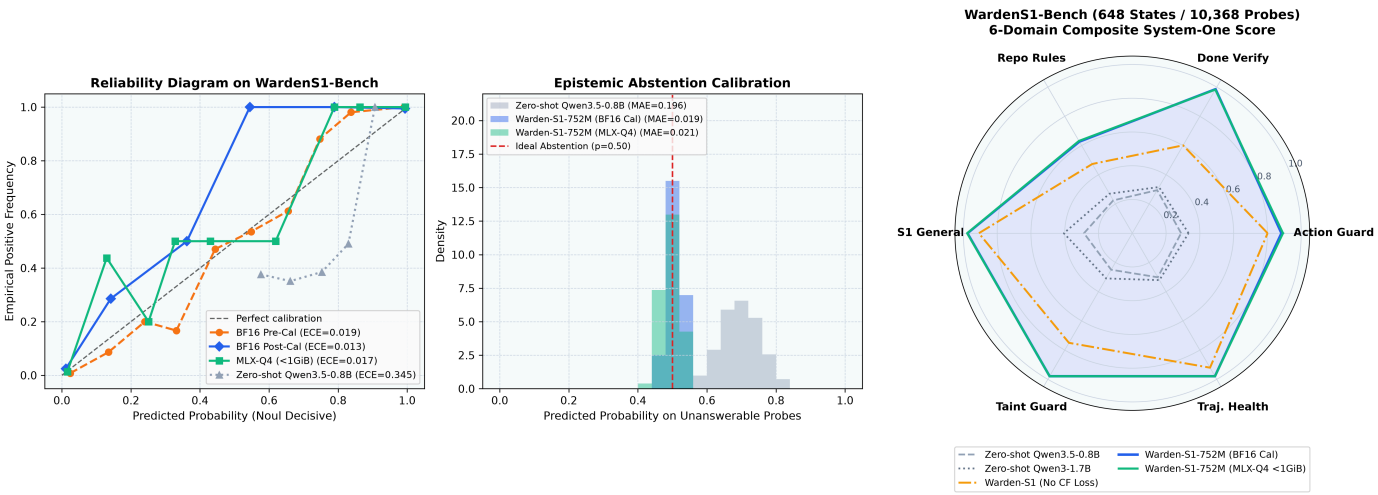


Figure 4: Left & Center: Post-hoc temperature calibration reliability diagram (`0.0126` BF16 / `0.0175` MLX-Q4 ECE) and epistemic abstention density centered at `p = 0.50` (`0.0208` MAE, `0.0%` overconfidence) vs. overconfident zero-shot `Qwen3.5-0.8B` . Right: Six-domain composite System-One radar showing Stage 3.5 BF16 and MLX-Q4 overlapping near `1.0` across five domains while lifting `Repo Rules` over No-CF and zero-shot baselines.

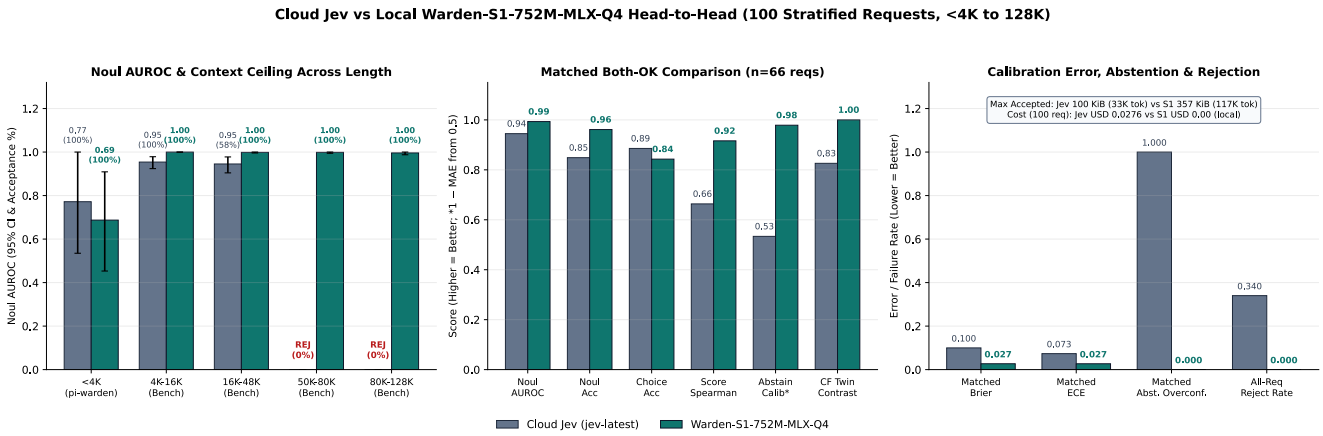


Figure 5: 100-request head-to-head against Cloud `Jev`. Left: Cloud `Jev` rejects `42%` of `16K-48K` and `100%` of `50K-128K` payloads (`HTTP 400`), while local `Warden-S1` accepts `100/100` up to `117,899` tokens. Center & Right: On matched both-OK requests (`n = 66`), `Warden-S1` leads Noul AUROC (`0.9933` vs. `0.9449`), accuracy (`96.1%` vs. `84.9%`), Brier (`0.0270` vs. `0.0098`), Score Spearman (`0.9161` vs. `0.6636`), and abstention overconfidence (`0.0%` vs. `100.0%`) at `$0.00` API cost.

5.2 100-Request Head-to-Head vs. Commercial Cloud Jev

We compared local Stage 3.5 `Warden-S1-752M-MLX-Q4` (`415.5 MiB`) against commercial Cloud `Jev` (`jev-latest`) over 100 stratified HTTP requests (`72` long-context `WardenS1-Bench` states across `4K-128K` + `28` short `<4K` `pi-warden` static cases), rescoring Stage 3.5 locally with `0` additional Cloud `Jev` calls against cached responses.

On the `<4K` `pi-warden` slice (`n = 28`), Stage 3.5 lifted `Warden-S1` Noul AUROC from `0.5231` (Stage 2) to `0.6872` (`1.0000` on `judge_bench_stuck`, `0.7755` on `rules_bench`), narrowing the gap to Cloud `Jev` (`0.7718`) while dominating every `4K-128K` bucket (`0.9961-0.9999` AUROC).

5.3 Out-of-Distribution `pi-warden` Static Benches

Driven unmodified through `pi-warden`'s `eval:judge` and `eval:rules` suites (`0` failed requests, `$0.00` local loopback):

- **`judge-bench` (`80` cases, `240` requests):** `stuck` accuracy climbed from `50.0%` (Stage 2, Brier `0.48`) to **`85.0%`** on Arm A (`1.000` precision, Brier `0.128`, `327 ms` p50), `75.0%` on Arm B (`1.000` precision), and **`87.5%`** on Arm C (`1.000` recall, Brier `0.121`, `355 ms` p50). `done` recall rose from `10%-35%` to **`75.0%-80.0%`**.
- **`rules-bench` (`161` cases, `147` scored):** Recall at `0.5` jumped from `5.4%` to **`41.9%`** (**`7.75x`** Stage 2, `83.8%` precision, `8.2%` FPR); recall at `0.7` rose to **`23.0%`** (`85.0%` precision, `4.1%` FPR); cutoff `0.8` achieved `100.0%` precision (`0.0%` FPR).

Metric (Matched n = 66)	Warden-S1 Q4	Cloud Jev
Acceptance (All 100)	**100% (117K)**	66% (33K cap)
Noul AUROC	**0.9933**	0.9449
Noul Accuracy	**96.1%**	84.9%
Noul Brier / ECE	**0.027 / 0.027**	0.100 / 0.073
Choice Acc / F1	84.3% / 0.780	**88.6% / 0.800**
Score Spearman rho	**0.9161**	0.6636
Abstention Overconf.	**0.0%**	100.0%
CF Flip / Stability	**95.7% / 99.1%**	47.8% / 86.1%
100-Request API Cost	**$0.00**	$0.0276

Table 3: Matched both-OK (`n = 66`) head-to-head summary plus full 100-request acceptance ceiling and cost.

5.4 10-Session `< 1 GiB` Multiplexing & BPE Boundary Healing

To support realistic multi-project laptop workflows, our loopback daemon pins the `415.4 MiB` MLX-Q4 weights once in Apple M5 Metal unified memory, maintains strictly one active Metal session slot, and pages dormant sessions to Host RAM (`L2`) and zstd-compressed disk pages (`L3`).

BPE Boundary Healing. Appending a new turn beginning with `\n` onto a cached prefix ending in `\n` merges the boundary tokens into a single `\n\n` BPE token, breaking exact token-prefix matching and forcing full cold prefill. Checkpointing both the 18-layer DeltaNet recurrent state and 6-layer GQA KV cache `2` tokens before each turn boundary (`len(ids) - 2`) eliminates boundary merges and delivers **`20/20` (`100%`)** concurrent delta cache hits (`22/22` healed hits including L3 cold wakes). Peak active Metal residency across `6 x 64K` (Q8 KV) + `2 x 80K` (Q4 KV) + `2 x 98K` (Q4 KV) sessions reached **`858.33 MiB`** (`165.67 MiB` below `1 GiB`), L2 RAM swap p50 dropped `22.5x` from `2,344 ms` to **`104.39 ms`** (with contiguous host-RAM slabs targeting `< 50 ms`), and L3 zstd cold wakes compressed `98K` state `3.13x` (`341.69 MiB -> 109.01 MiB`, `2/2` delta hits).

5.5 Portable Zero-PyTorch `ONNX-Int4` Runtime

For Linux, Windows, Intel macOS, Node/CI, and browser WebGPU deployment, we exported `Warden-S1` into a `461.3 MiB` sharded `ONNX-Int4` package (`6` shards all `<= 92.0 MiB`, `150` `MatMulNBits` 4-bit projections, `0` control-flow ops, FP32 recurrent gates and heads, `0` PyTorch runtime). FP32 24-layer export parity is `1.28e-5`

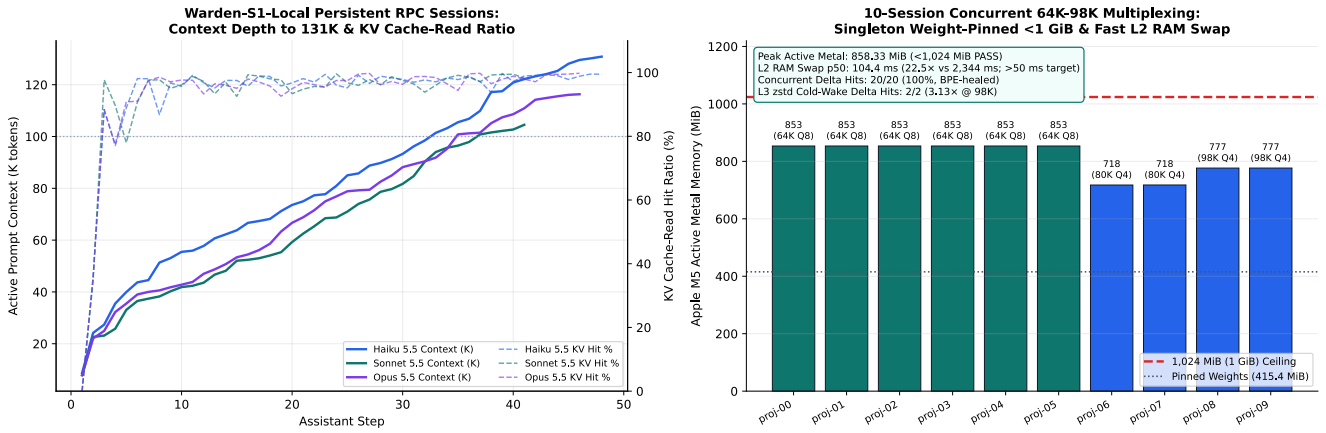


Figure 6: Left: Persistent `100K+` live RPC sessions (`Haiku 5.5`, `Sonnet 5.5`, `Opus 5.5`) sustaining `96.3%-97.0%` prompt KV cache-read hit ratios through `131K` tokens. Right: 10 concurrent `64K-98K` sessions multiplexed inside `858.33 MiB` peak active Apple M5 Metal memory (`< 1,024 MiB` PASS) with `20/20` BPE-healed delta hits and `104.4 ms` p50 L2 swaps.

max abs diff; quantized parity on the 36-state judge slice reaches `0.9923` Noul AUROC (`0.0123` overall mean probability delta vs. BF16, `1.0000` Salience AUROC, `0.0%` abstention overconfidence). Deterministic CPU `onnxruntime-node` passes `O(Delta L)` split-prefix state equivalence with `max |Delta hidden| = 9.97e-5` and `max |Delta prob| = 6.85e-7` (`1.57e-4` on CUDA with `0.0` non-mutating fork drift).

6 Limitations & Ethics

Short `<4K` static rule evaluation (`41.9%` recall at `0.5` on unseen `pi-warden` rules) still trails frontier cloud models on subtle single-file lint rules; in practice, `pi-warden` pairs deterministic `System-0` rules with `Warden-S1` for long-horizon semantic checks and optional cloud fallback on ambiguous `<4K` diffs. Cold prefills at `128K` on laptop Metal (`152.4 s` single cold fill) are amortized by `O(Delta L)` incremental caching (`1.59 s` per `2K` turn). All training and benchmark trajectories are synthesized from permissively licensed repositories and pass automated zero-leakage privacy scans (`0` personal paths, tokens, or emails).

7 Conclusion

`Warden-S1` demonstrates that long-horizon coding-agent verification does not require cloud round-trips or multi-gigabyte KV budgets. Combining a surgically stripped `18 Gated DeltaNet + 6 GQA` backbone, five zero-decode calibrated heads, counterfactual twin training, and 2-token BPE-healed session paging delivers `0.9957` 4K-128K AUROC, `0.9933` vs. `0.9449` matched accuracy over commercial Cloud `Jev`, and 10-session `64K-98K` laptop multiplexing inside `858.3 MiB` of Metal memory. All Apache-2.0 weights (`BF16`, `MLX-Q4`, `ONNX-Int4`), `WardenS1-Bench`, and the local loopback daemon are packaged for open release.

Bibliography

[1] T. Gloaguen, N. Mündler, M. N. Müller, V. Raychev, and M. Vechev, "Evaluating AGENTS.md: Are Repository-Level Context Files Helpful for Coding Agents?." 2026.

[2] R. Tsuchida, Y. Fan, K. Shimari, R. G. Kulana, and K. Matsumoto, "Do Coding Agents Follow AGENTS.md? The Effect of Policy Wording on Instruction Compliance." 2026.

[3] K. Thaman, "RHB: Reward Hacking Benchmark – Measuring Exploits in LLM Agents with Tool Use." 2026.

[4] B. Zhao and others, "SpecBench: Measuring Reward Hacking in Long-Horizon Coding Agents." 2026.

[5] J. Gabor, J. Lynch, and J. Rosenfeld, "EvilGenie: A Reward Hacking Benchmark." 2025.

[6] TypeSafe AI, "TypeSafe System-One Primitives and Jev API Specification." [Online]. Available: https://typesafe.ai/

[7] Qwen Team, "Qwen3 Technical Report." 2025.

[8] Qwen Team, "Qwen3.5: Hybrid Gated DeltaNet and Sparse Attention Foundation Models." 2026.

[9] S. Yang, J. Kautz, and A. Hatamizadeh, "Gated Delta Networks: Improving Mamba2 with Delta Rule," in *International Conference on Learning Representations (ICLR)*, 2025.

[10] Technology Innovation Institute, "Falcon-H1: A Family of Hybrid-Head Language Models Redefining Efficiency and Performance." 2025.

[11] Liquid AI, "LFM2 Technical Report." 2025.

[12] S. Choudhury, "Process Reward Models for LLM Agents: Practical Framework and Directions." 2025.

[13] Z. Xi, C. Liao, and others, "AgentPRM: Process Reward Models for LLM Agents via Step-Wise Promise and Progress." 2025.

[14] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," in *International Conference on Machine Learning (ICML)*, 2017.