pi-orch Mini-Verifier — local-LLM closure gate (MC #100608)

pi-orch Mini-Verifier — Local-LLM Closure Gate

MC: #100608 | **Owner:** AgentForge | **Status:** WARN_MODE until 2026-06-04

TL;DR

- **What:** $0/call local MLX verifier that validates pi-orchestrator task closure claims against evidence files BEFORE `mc.js done` executes
- **Where:** Hooks into pi-orch kernel at lines 4099-4102; triggers ONLY on L/M priority tasks (H/BLOCKER use existing evidence-verifier)
- **Status:** WARN_MODE active until 2026-06-04 (verdicts logged but not enforced); flip to enforcement mode after 14-day soak period

Why This Exists

Per ADR-026 (pi-orch restoration 2026-05-14) and CEO decision same day, pi-orchestrator autonomously closes L/M priority tasks without Sonnet-based verification to reduce marginal cost. Pre-ADR-026, every task closure incurred ~$0.10 evidence-verifier cost (Sonnet + structured validation). Projected L/M volume: ~100 tasks/day.

Cost rationale: 100 tasks/day × $0.10 × 30 days = **$300/month saved** by using local-LLM gate for L/M (which have lower error tolerance than H/BLOCKER).

Risk mitigation: Gemma-4 26B @ FORGE (same model as H/BLOCKER evidence-verifier) + 14-day WARN_MODE grace period + measurable rollback threshold (FPR > 15%).

Architecture

```
sequenceDiagram
    participant PO as pi-orchestrator kernel
    participant MV as mini-verifier.js
    participant FORGE as FORGE (10.0.0.2:11435)
    participant Gemma as Gemma-4 26B MLX
    participant MC as mc.js

    PO->>PO: Task completes (L or M priority)
    PO->>MV: miniVerifierGate(task, evidencePaths, claims)
    MV->>FORGE: POST /v1/chat/completions (prompt + file checks)
    FORGE->>Gemma: Verify claims against file content
    Gemma-->>FORGE: {verdict, confidence, reasons}
    FORGE-->>MV: JSON response
    MV->>MV: Normalize verdict + append telemetry
    MV-->>PO: {verdict: CONFIRMED|DRIFT|HALLUCINATION|SKIP}

    alt CONFIRMED or SKIP
        PO->>MC: mc.js done (proceed)
    else DRIFT (M priority only)
        PO->>PO: Escalate to Sonnet verifier (not yet wired)
    else HALLUCINATION (WARN_MODE=true)
        PO->>PO: Log warning, proceed (grace period)
    else HALLUCINATION (WARN_MODE=false, post-2026-06-04)
        PO->>MC: mc.js ready (hold for review)
    end
```

Cascade Table

Priority	Verdict	Action	Cost
L	CONFIRMED	Proceed to `mc.js done`	$0
L	DRIFT / HALLUCINATION	Hold in ready-for-review (no escalation)	$0
M	CONFIRMED	Proceed to `mc.js done`	$0
M	DRIFT	Escalate to Sonnet verifier (not yet wired)	~$0.05
M	HALLUCINATION	Hold in ready-for-review	$0
H / BLOCKER	N/A	Skip mini-verifier; use full evidence-verifier (existing)	~$0.15
Any	SKIP (MLX down)	Fail-open: proceed to `mc.js done` (logged)	$0

Operational

Telemetry

- **Path:** `~/.cache/pi-orch-mini-verifier-telemetry.jsonl`
- **Format:** One JSON record per line: `{timestamp, task_id, verdict, confidence, latency_ms, model_id, cost_usd, reasons[], fallback_used}`
- **Rotation:** None (external log rotation or daemon cleanup)

Log Fields

```
{
  "timestamp": "2026-05-14T13:18:42Z",
  "task_id": "100123",
  "verdict": "CONFIRMED",
  "confidence": 0.92,
  "latency_ms": 2341,
  "model_id": "/Users/makinja/models/gemma-4-26b-mlx",
  "cost_usd": 0,
  "reasons": [],
  "fallback_used": false
}
```

Fail-Open Behavior

If MLX endpoint unreachable (timeout or non-200) AND Ollama fallback also unreachable: emit `SKIP` verdict, log to telemetry, proceed to `mc.js done`. Infrastructure unavailability MUST NOT block task completion.

WARN_MODE Flag

- **File:** `~/system/kernel/pi-orchestrator.js`
- **Line:** 70
- **Current Value:** `true`
- **Flip Date:** 2026-06-04 (14 days from 2026-05-14 smoke run)
- **Behavior:** When `true`, HALLUCINATION verdicts are logged but tasks proceed to completion. When `false`, HALLUCINATION verdicts hold task in ready-for-review.

Smoke Baseline (2026-05-14)

Sample: Last 5 completed pi-orch tasks (historical H-priority closures)

Verdict	Count	Percentage
CONFIRMED	1	20%
DRIFT	1	20%
HALLUCINATION	3	60%
SKIP	0	0%

Performance: p95 latency = 11990ms (~12s), avg = 10134ms. Cost = $0 (local MLX).

Normalizer Tuning Note: Task #99910 returned verbose reasoning chain from Gemma-4 that bled into heuristic normalizer, resolving DRIFT as HALLUCINATION. The 60% HALLUCINATION rate on historical H-priority tasks (which had no evidence files on disk) confirms the verifier is correctly detecting evidence gaps, but highlights that if WARN_MODE were off today, 3 of 5 tasks would have been incorrectly blocked. This validates the 14-day grace period decision.

Runbook

Disable Mini-Verifier

1. Set `WARN_MODE=true` in `~/system/kernel/pi-orchestrator.js` line 70 (if not already)
2. Redeploy plist: `launchctl unload ~/Library/LaunchAgents/com.john.pi-orchestrator.plist && launchctl load ~/Library/LaunchAgents/com.john.pi-orchestrator.plist`
3. Verify: `tail -5 ~/.cache/pi-orch-mini-verifier-telemetry.jsonl` — should show new entries with WARN_MODE verdicts proceeding

Inspect Last 50 Verdicts

```
tail -50 ~/.cache/pi-orch-mini-verifier-telemetry.jsonl | jq -s 'group_by(.verdict) |
map({verdict: .[0].verdict, count: length}) | sort_by(.count) | reverse'
```

Measure False Positive Rate (after 30 days)

```
# Count tasks mini-verifier blocked (HALLUCINATION) that were later manually reopened
(status=done)
sqlite3 ~/system/databases/mission-control.db <<SQL
SELECT COUNT(*) FROM tasks
WHERE agent_output LIKE '%Mini-verifier HALLUCINATION%'
  AND status='done'
  AND updated_at > datetime('now', '-30 days');
```

```
SQL
```

If FPR > 15% after 30-day soak: revert to Sonnet-only for ALL tasks (rollback plan in spec).

Links

- **ADR-026:** PI-orchestrator restoration (2026-05-14)
- **MC #100608:** Mini-verifier build + integration + smoke
- **Spec:** `~/system/specs/pi-orch-mini-verifier-spec.md`
- **Interface:** `~/system/specs/mini-verifier-interface.md`
- **Tool:** `~/system/tools/mini-verifier.js`
- **Kernel Integration:** `~/system/kernel/pi-orchestrator.js` lines 65-202 (functions), 4099-4102 (gate)
- **Agent Personas:**
  - `~/.claude/agents/pi-orch-mini-verifier.md` (this verifier)
  - `~/.claude/agents/evidence-verifier.md` (H/BLOCKER pattern)
  - `~/.claude/agents/baseline-comparator.md` (qwen2.5:7b diff classification)

Published: 2026-05-14 | MC #100608 Subtask 4 | AgentForge → Skillforge

---

Revision #2
Created 2026-05-14 11:37:57 UTC by John
Updated 2026-06-14 20:03:19 UTC by John