Anthropic Outage Resilience — 529 Auto-Fallback Runbook

Anthropic Outage Resilience — 529 Auto-Fallback Runbook

MC: #104217 T5
Owner: Skillforge
Date: 2026-06-22
Status: Production (Active)
BookStack: System Architecture

Executive Summary

What It Does:
When Anthropic API returns HTTP 529 (overloaded) on ALAI agent/tool paths, the system auto-enables offline-mode and routes LLM work to local Ollama (FORGE or ANVIL) within 30 seconds. Auto-recovery occurs when Anthropic becomes healthy again (5-minute health check cycle).

What It Protects:

- Agent LLM calls via `adapters/claude-api.js` (line 194, 231)
- Company Mesh comms-responder legacy path
- Company worker CLI stderr path
- Tool execution requiring LLM reasoning

What It Does NOT Protect (Honest Limits):

- John's own Claude Code CLI session 529s (not interceptable — hooks run after Claude's internal API call)
- During full Anthropic outage, John-the-orchestrator degrades to `john-lite` for bounded triage only, NOT full orchestration
- H/BLOCKER/deploy/security tasks are rejected in offline mode (quality gates require full reasoning)

Cost:

- Development: $1,800 one-time (MC #104217 T1+T2+T4)
- Operational: $0/month (local Ollama)
- Avoided productivity loss: $1,200-$2,400/month (2-4 stalls/week × 2h × $150/h CEO time)

Key Dependency:
FORGE Ollama (10.0.0.2:11434) must be reachable. Falls back to ANVIL (localhost:11434) if FORGE down.

---

1. System Architecture

1.1 Auto-Detection Layer (T1)

File: `/Users/makinja/system/tools/anthropic-529-detector.js`
Owner: FlowForge
Evidence: `/tmp/evidence-104217/t1-hook/`

How It Works:

1. Wraps all Anthropic API calls with `wrapAnthropicCall()` middleware
2. Catches errors and applies `is529Error()` detector:
   - HTTP status code 529
   - Error message contains "overload" (case-insensitive)
   - Word-boundary regex `/\b(status|code|http|error)\s*529\b/i` (avoids false positives on "529ms", "in 529 milliseconds")
   - Anthropic SDK `error.type === 'overloaded_error'`
3. On 529 match:
   - Writes `/tmp/john-offline-mode` flag with metadata (timestamp, reason)
   - Spawns background recovery daemon (`node anthropic-529-detector.js recovery-daemon`)
   - Re-throws original error (caller decides how to handle)

Wired Call Sites (verified 2026-06-22):

```
// adapters/claude-api.js line 194 (initial message)
const detector = require('../anthropic-529-detector');
let response = await detector.wrapAnthropicCall(async () => {
  return await client.messages.create(apiParams, { signal: controller.signal });
});
```

```
// adapters/claude-api.js line 231 (tool-use round)
response = await detector.wrapAnthropicCall(async () => {
  return await client.messages.create(apiParams, { signal: roundCtl.signal });
});
```

Additional wired sites (per T2 job1-detector-wiring.md):

- `comms-responder.js` (Company Mesh legacy)
- `company-worker.js` (CLI stderr path)

State Files:

- `/tmp/john-offline-mode` — Offline-mode flag (checked by tier-router.js)
- `/tmp/anthropic-529-detector.json` — Detector state (trigger time, health check history)

Recovery Behavior:

- Auto-health-check every 5 minutes when offline-mode active
- If Anthropic responds with status != 529, auto-disables offline-mode
- TTL: Max 2 hours offline before forcing re-check
- Health check: `https OPTIONS api.anthropic.com/v1/messages` (any response except 529 = healthy)

1.2 Degraded Orchestration Layer (T2)

File: `/Users/makinja/system/tools/john-lite.js`
Owner: AgentForge
Evidence: `/tmp/evidence-104217/t2/`

Purpose:
Bounded orchestration continuity when `/tmp/john-offline-mode` flag is active.

Modes:

```
node john-lite.js loop         # REPL-like degraded orchestration loop
node john-lite.js once "<task>" # One-shot task dispatch
node john-lite.js triage       # MC triage (what needs attention)
node john-lite.js status       # Show capabilities + offline status
```

Capabilities (CAN DO):

- MC triage (list open tasks, show task details via `mc.js`)
- Task classification (priority, agent selection)
- Simple dispatch to Ollama-tier agents (research, analysis, draft)

- Read-only tool execution (git status, mc.js list, file reads)
- Bounded research/brainstorm/summarize tasks
- Status checks (daemon health, service status)

Capabilities (CANNOT DO — save for full John):

- H/BLOCKER priority orchestration (quality gates demand full reasoning)
- Mehanik/prompt-forge workflows (multi-turn agentic planning)
- Company Mesh P2P verifier orchestration
- AI Factory workflow dispatch
- Production deploys, security decisions, architecture changes
- Evidence ledger writes (L2+ verification)
- Complex multi-agent coordination
- Anything requiring Opus/Sonnet-level reasoning

Rejection Logic:
Tasks matching these patterns exit with code 3:

```
const COMPLEX_PATTERNS = [
  /\b(deploy|production|staging|release)\b/i,
  /\b(security|auth|encrypt|vulnerability)\b/i,
  /\b(architecture|refactor|migrate)\b/i,
  /\b(H|BLOCKER|P0|P1)\b/i,
  /\b(mehanik|prompt-forge|company-mesh|ai-factory)\b/i,
  /\b(evidence|verification|validator|proveo)\b/i,
  /\b(multi-file|cross-service|integration)\b/i,
];
```

Exit Codes:

- 0 = success
- 1 = Anthropic healthy (john-lite not needed)
- 2 = No reachable Ollama host (FORGE + ANVIL both down)
- 3 = Task too complex for john-lite (needs full John)

Output Storage:
All john-lite output saved to `~/system/offline-queue/<timestamp>_john-lite_<type>.md` with `NEEDS_REVIEW` flag for post-outage review.

Log File:
`/tmp/john-lite-log.jsonl` (append-only JSONL)

1.3 Local Ollama Fleet

Primary: FORGE (10.0.0.2:11434)
Fallback: ANVIL (localhost:11434)

FORGE Models (verified 2026-06-22)

```
$ curl -s http://10.0.0.2:11434/api/tags | jq -r '.models[].name'
qwen2.5:7b-instruct-q8_0
qwen3-coder:30b          # Code primary
qwen3.5:27b
deepseek-r1:70b          # Deep reasoning (42GB)
qwen2.5-coder:32b-instruct-q8_0
qwen3:32b                # Reasoning primary
qwen3:8b-q8_0
bge-m3:latest            # Embedding
```

Status: UP (2026-06-22)
Network: Listens on `*:11434` (all interfaces)
Fix History: MC #104217 T2 Job 3 — OLLAMA_HOST=0.0.0.0:11434 added to launchd plist to enable remote access

ANVIL Models (verified 2026-06-22)

```
$ curl -s http://localhost:11434/api/tags | jq -r '.models[].name'
bge-m3:latest
llama3.1:8b              # Reasoning fallback
nomic-embed-text:latest
llama-guard3:8b
llama-guard3:8b-q8_0
```

Status: UP (2026-06-22)
Network: Localhost only (127.0.0.1:11434)

---

2. Operator Procedures

2.1 Check Offline Mode Status

```
# Quick status
node ~/system/tools/anthropic-529-detector.js status
```

```
# Example output:
=== Anthropic 529 Detector Status ===

Offline Mode: ACTIVE
Trigger Reason: Anthropic API 529 overload detected: status 529
Offline Since: 2026-06-22T14:23:15.123Z (12 minutes ago)
Last Health Check: 2026-06-22T14:28:00.456Z
  Result: unhealthy
  Status Code: 529
Auto-Recovery: enabled
```

2.2 Check john-lite Status

```
node ~/system/tools/john-lite.js status

# Example output:
=== JOHN-LITE STATUS ===

Offline Mode: □□ACTIVE
Reason: Anthropic API 529 overload detected

Ollama Hosts:

  □ FORGE (http://10.0.0.2:11434)
     Models: qwen3-coder:30b, qwen3:32b, deepseek-r1:70b, qwen2.5-coder:32b, ...
  □ ANVIL (http://localhost:11434)
     Models: llama3.1:8b, nomic-embed-text:latest, ...
```

2.3 Manual Enable/Disable Offline Mode

Enable (test mode):

```
node ~/system/tools/anthropic-529-detector.js test-529
# Simulates 529 trigger, enables offline-mode
```

Disable (manual clear):

```
node ~/system/tools/anthropic-529-detector.js clear
# Removes /tmp/john-offline-mode flag
```

Force Health Check:

```
node ~/system/tools/anthropic-529-detector.js recovery-check
# Runs one health check cycle immediately
```

2.4 Monitor Logs

Detector State:

```
cat /tmp/anthropic-529-detector.json | jq .
```

john-lite Activity:

```
tail -f /tmp/john-lite-log.jsonl | jq .
```

Offline Queue (output awaiting review):

```
ls -lt ~/system/offline-queue/*.md | head -5
```

2.5 Check FORGE/ANVIL Reachability

FORGE (from ANVIL):

```
curl -s --max-time 3 http://10.0.0.2:11434/api/tags | jq -r '.models[].name' | head -5
```

ANVIL (local):

```
curl -s --max-time 3 http://localhost:11434/api/tags | jq -r '.models[].name' | head -5
```

If FORGE down:

1. SSH to FORGE: `ssh makinja@10.0.0.2`
2. Check Ollama service:

   ```
   lsof -nP -iTCP -sTCP:LISTEN | grep ollama
   launchctl list | grep ollama
   ```

3. Verify `OLLAMA_HOST=0.0.0.0:11434` in `~/Library/LaunchAgents/homebrew.mxcl.ollama.plist`
4. Reload if needed:

   ```
   launchctl unload ~/Library/LaunchAgents/homebrew.mxcl.ollama.plist
   launchctl load ~/Library/LaunchAgents/homebrew.mxcl.ollama.plist
   ```

5. If unrecoverable, system auto-falls back to ANVIL localhost:11434

---

3. Recovery Behavior (Auto)

3.1 Normal Recovery Cycle

1. 529 detected → offline-mode ENABLED → recovery daemon spawned
2. Every 5 minutes: health check `https OPTIONS api.anthropic.com/v1/messages`
3. If response status != 529 → offline-mode DISABLED → daemon exits
4. Next agent/tool call routes to Anthropic normally

Timeline:

- Detection to offline-mode: <30 seconds
- Recovery check interval: 5 minutes
- Max offline duration (TTL): 2 hours (forces health check)

3.2 Manual Recovery (if auto-recovery stuck)

```
# Check if Anthropic is healthy
node ~/system/tools/anthropic-529-detector.js health

# If healthy, manually clear offline mode
node ~/system/tools/anthropic-529-detector.js clear
```

---

4. What Is NOT Protected (Honest Limits)

4.1 Claude Code CLI Session 529s

Problem:
When you (John) interact with CEO via Claude Code CLI and Claude's backend returns 529, the CLI's internal error handling kicks in BEFORE the `anthropic-529-detector.js` hook can intercept it.

Why:
The detector wraps `adapters/claude-api.js` (ALAI's own agent tool calls), not the Claude Code

executable's internal network stack.

Workaround:
Use `john-lite.js loop` for bounded orchestration during outages. Accept degraded quality for the duration.

Evidence:
MC #104217 T1 IMPLEMENTATION.md line 35-40:

```
CONSTRAINTS (HONEST):
  - CANNOT intercept Claude Code CLI's own 529s (those are CLI-internal)
  - CAN detect 529s from ALAI agent/tool calls (company-worker, tier-router path)
  - Focus: agent workflow continuity, not CLI session continuity
```

4.2 High-Priority/Complex Work

Rejected in offline mode:

- H/BLOCKER priority tasks
- Deploy/production/security decisions
- Architecture changes
- Multi-agent orchestration (Company Mesh, AI Factory)
- Evidence synthesis (L2+ verification)

Rationale:
Local Ollama 32B models lack the reasoning depth for quality gates. These tasks wait for Anthropic recovery.

How to check:
`john-lite.js` exits with code 3 and logs rejection reason.

---

5. Cost Analysis (Why Not API Priority Tier?)

Full Analysis: `/Users/makinja/system/specs/anthropic-priority-tier-analysis.md`
Conclusion: NO-GO on Priority Tier / Provisioned Throughput API migration

Rationale:

1. **Anthropic does NOT offer a "Priority Tier" that prevents 529 errors.**
   Their tier system (Tier 1-5) controls rate limits (RPM/TPD/TPM), NOT capacity guarantees.

A Tier 4 user can still hit 529 if Anthropic's backend is overloaded.

2. **No API migration path for Claude Code subscription.**
   ALAI's orchestration runs on Claude Code CLI (subscription-based, no `ANTHROPIC_API_KEY`). Cannot "upgrade to priority tier" — different product line.
3. **API migration cost vastly exceeds productivity loss:**
   - Current subscription: ~$500-2,000/month (embedded in Claude Code Enterprise license)
   - Hypothetical API (Tier 4): $13,400-$18,367/month (2-2.5x increase due to loss of free caching)
   - Hypothetical Provisioned Throughput: $15,000-$30,000/month (estimated, unverified)
   - Productivity loss from 529 stalls: $1,200-$2,400/month (2-4 stalls × 2h × $150/h CEO time)
   - **ROI: NEGATIVE. Cost increase >> productivity loss.**
4. **Auto-fallback to local Ollama delivers 529 resilience at $0 marginal cost.**
   - Development: $1,800 one-time (MC #104217 T1+T2+T4)
   - Operational: $0/month (FORGE/ANVIL already owned, Ollama free)
   - **ROI: POSITIVE. Payback in <1 month.**

Recommendation:
Maintain hybrid model (Claude subscription + auto-fallback). Defer API migration unless Anthropic provides SLA-backed capacity guarantee + cost < $5K/month.

---

6. Evidence & Sources

Implementation Evidence

MC #104217 T1 (FlowForge):
`/tmp/evidence-104217/t1-hook/`

- FLOWFORGE-REPORT.md
- IMPLEMENTATION.md (detector design)
- verification-output.txt (test results)

MC #104217 T2 (AgentForge):
`/tmp/evidence-104217/t2/`

- job1-detector-wiring.md (wired call sites)
- job2-john-lite.md (degraded orchestration)
- job3-forge-ollama-fix.md (network binding fix)
- SUMMARY.md

MC #104217 T4 (Proveo):
`/tmp/evidence-104217/t4-proveo/`

- test-results.txt (simulation + validation)

MC #104217 T3 (AgentForge):
`/Users/makinja/system/specs/anthropic-priority-tier-analysis.md`
(Tier analysis, cost/benefit, NO-GO recommendation)

Source Files (canonical)

- `/Users/makinja/system/tools/anthropic-529-detector.js` (T1 detector + recovery daemon)
- `/Users/makinja/system/tools/john-lite.js` (T2 degraded orchestration)
- `/Users/makinja/system/tools/adapters/claude-api.js` (wired call site lines 194, 231)
- `/Users/makinja/system/tools/comms-responder.js` (legacy Company Mesh path)
- `/Users/makinja/system/tools/company-worker.js` (CLI stderr path)

Web Sources (Tier Analysis)

- Claude Subscription Plans (Google Vertex AI Search grounding-api-redirect, 2026-06-22)
- Anthropic API Rate Limit Tiers (Google Vertex AI Search grounding-api-redirect, 2026-06-22)
- Claude Opus 4 / Sonnet 4.6 API Pricing (Google Vertex AI Search grounding-api-redirect, 2026-06-22)
- Prompt Caching & Batch API (Google Vertex AI Search grounding-api-redirect, 2026-06-22)

---

7. Frequently Asked Questions

Q: Why not just buy API priority tier?

A: Anthropic does not offer a "priority tier" that prevents 529 overload errors. Their tier system (Tier 1-5) only controls rate limits (requests per minute/day, tokens per minute), not capacity guarantees. Even Tier 4 users can hit 529 during backend overload.

Provisioned Throughput (enterprise-only, pricing undisclosed) might reduce exposure, but estimated cost ($15K-$30K/month) vastly exceeds productivity loss from 529 stalls ($1.2K-$2.4K/month).

Q: How long does it take to switch to offline mode?

A: <30 seconds from 529 detection to `/tmp/john-offline-mode` flag active. Next agent/tool call routes to Ollama.

Q: How long does it take to recover when Anthropic is healthy again?

A: 5-minute health check cycle. Once Anthropic responds with status != 529, offline-mode is auto-disabled. Next call routes to Anthropic.

Q: What if FORGE Ollama is down?

A: System auto-falls back to ANVIL localhost:11434 (llama3.1:8b reasoning, nomic-embed-text embedding). If both FORGE + ANVIL down, `john-lite.js` exits with code 2 and logs "No reachable Ollama host."

Q: Can I manually trigger offline mode for testing?

A: Yes.

```
node ~/system/tools/anthropic-529-detector.js test-529
```

Clear with:

```
node ~/system/tools/anthropic-529-detector.js clear
```

Q: How do I review john-lite output after outage recovery?

A: Check `~/system/offline-queue/*.md` for all output generated during offline mode. Each file includes:

- Timestamp
- Task description
- Model used (qwen3:32b, llama3.1:8b, etc.)
- Output
- `NEEDS_REVIEW` flag

Review before using in production (local model accuracy < Claude Opus 4).

Q: Where are the logs?

A:

- Detector state: `/tmp/anthropic-529-detector.json`
- john-lite activity: `/tmp/john-lite-log.jsonl`
- Offline-mode flag: `/tmp/john-offline-mode`
- Offline output queue: `~/system/offline-queue/<timestamp>_john-lite_*.md`

---

8. Related Documentation

- **MC #104217:** [H] Anthropic-outage resilience: firma ne smije stati kad Claude API vrati 529/overloaded
- **Tier Analysis:** `/Users/makinja/system/specs/anthropic-priority-tier-analysis.md`
- **FORGE Ollama Fix:** `/tmp/evidence-104217/t2/job3-forge-ollama-fix.md`
- **Cost Tracking:** `node ~/system/tools/cost-tracker.js summary week`

Last Updated: 2026-06-22T21:29:00Z
Owner: Skillforge
Status: Production (Active)
Runbook Version: 1.0

END OF RUNBOOK

Revision #2
Created 2026-06-22 19:33:44 UTC by John
Updated 2026-06-22 19:34:05 UTC by John