ANVIL Memory Troubleshooting — Mac Studio

ANVIL Memory Troubleshooting — Mac Studio (M2 Ultra 192GB)

Incident Summary

Date: 2026-04-20
Symptom: System freezes, Chrome/Claude unresponsive, OOM kernel panics
Root Cause: Zombie Ollama runner processes + duplicate launchd agents + runaway grep processes
Resolution: Ollama config tuning, duplicate agent removal, zombie cleanup daemon, Ollama 0.21.0 upgrade

Root Causes

1. **Ollama zombie runners:** `ollama ps` reports 0 models loaded, but `pgrep -fl ollama_llama_server` shows 4-6 GB processes still resident
2. **Duplicate launchd agents:** Both `com.alai.ollama-serve.plist` and `com.alai.ollama-serve-v2.plist` running simultaneously → 2x Ollama daemons
3. **grep memory leak:** `grep -rn` commands on large codebases hang and consume 8+ GB RAM each
4. **Preload warmup bloat:** `com.john.ollama-warmup.plist` loading 3 models on boot → 48 GB baseline before any work

Permanent Fix — Ollama Config

File: `~/Library/LaunchAgents/com.alai.ollama-serve-v2.plist`

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE plist PUBLIC "-//Apple//DTD PLIST 1.0//EN" "http://www.apple.com/DTDs/PropertyList-1.0.dtd">
<plist version="1.0">
<dict>
  <key>Label</key>
  <string>com.alai.ollama-serve-v2</string>
  <key>ProgramArguments</key>
  <array>
    <string>/usr/local/bin/ollama</string>
    <string>serve</string>
  </array>
  <key>EnvironmentVariables</key>
  <dict>
    <key>OLLAMA_HOST</key>
    <string>0.0.0.0:11434</string>
    <key>OLLAMA_KEEP_ALIVE</key>
    <string>60s</string>
    <key>OLLAMA_MAX_LOADED_MODELS</key>
    <string>1</string>
    <key>OLLAMA_NUM_PARALLEL</key>
    <string>1</string>
  </dict>
  <key>RunAtLoad</key>
  <true/>
  <key>KeepAlive</key>
  <true/>
  <key>StandardOutPath</key>
  <string>/tmp/ollama-serve.log</string>
  <key>StandardErrorPath</key>
  <string>/tmp/ollama-serve-error.log</string>
</dict>
</plist>
```

Key parameters:

- `OLLAMA_KEEP_ALIVE=60s` — unload model after 60s idle (default 5m causes bloat)
- `OLLAMA_MAX_LOADED_MODELS=1` — only one model resident at a time
- `OLLAMA_NUM_PARALLEL=1` — no parallel inference (reduces contention)

Zombie Cleanup Daemon

File: `~/Library/LaunchAgents/com.alai.zombie-cleanup.plist`

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE plist PUBLIC "-//Apple//DTD PLIST 1.0//EN" "http://www.apple.com/DTDs/PropertyList-1.0.dtd">
<plist version="1.0">
<dict>
  <key>Label</key>
  <string>com.alai.zombie-cleanup</string>
  <key>ProgramArguments</key>
  <array>
    <string>/bin/bash</string>
    <string>/Users/makinja/system/tools/zombie-proc-cleanup.sh</string>
  </array>
  <key>StartInterval</key>
  <integer>3600</integer>
  <key>StandardOutPath</key>
  <string>/tmp/zombie-cleanup.log</string>
</dict>
</plist>
```

Script: `~/system/tools/zombie-proc-cleanup.sh`

```
#!/bin/bash
# Kill zombie Ollama runners (no parent process or disconnected from ollama serve)
pgrep -fl ollama_llama_server | while read -r pid rest; do
  parent=$(ps -o ppid= -p "$pid" | xargs)
  if [[ -z "$parent" ]] || ! ps -p "$parent" | grep -q ollama; then
    echo "$(date): Killing zombie Ollama runner $pid"
    kill -9 "$pid"
  fi
done

# Kill grep processes older than 5 minutes (likely hung)
ps -eo pid,etime,command | grep 'grep -rn' | while read -r pid etime rest; do
  minutes=$(echo "$etime" | awk -F: '{print ($1*60)+$2}')
  if [[ "$minutes" -gt 5 ]]; then
```

```
    echo "$(date): Killing hung grep process $pid (runtime: $etime)"
    kill -9 "$pid"
  fi
done
```

Disabled Agents

```
launchctl unload ~/Library/LaunchAgents/com.alai.ollama-serve.plist
launchctl unload ~/Library/LaunchAgents/com.john.ollama-warmup.plist
rm ~/Library/LaunchAgents/com.alai.ollama-serve.plist
rm ~/Library/LaunchAgents/com.john.ollama-warmup.plist
```

Ollama Upgrade

```
brew upgrade ollama  # 0.19.0 → 0.21.0
# Changelog: Fixed memory leak in runner cleanup (issue #4821)
```

OOM Symptom Recognition

Command:

```
vm_stat | awk '/Pages free/ {printf "%.1f GB\n", $3*16384/1024/1024/1024}'
```

Thresholds:

- **< 5 GB free:** Alert — investigate top memory consumers
- **< 2 GB free:** Critical — kill non-essential processes immediately
- **< 500 MB free:** Imminent OOM — force quit Claude/Chrome, restart Ollama

Quick triage:

```
ps aux | sort -nrk 4 | head -10  # Top 10 memory hogs
pgrep -fl ollama_llama_server    # Zombie Ollama runners
pgrep -fl grep                    # Hung grep processes
```

Prevention Checklist

1. Monitor free RAM hourly: `vm_stat` check in cron
2. Zombie cleanup daemon running: `launchctl list | grep zombie-cleanup`
3. Only one Ollama launchd agent: `launchctl list | grep ollama` → expect 1 line
4. No warmup preload agents: `launchctl list | grep warmup` → empty
5. Grep with timeout: `timeout 60 grep -rn ...` instead of bare `grep -rn`

Revision #2
Created 2026-04-20 14:40:46 UTC by John
Updated 2026-05-31 20:06:12 UTC by John