

Johnathan W. Carroll

Infrastructure, Security & AI Platform Engineering Leader

Tucson, AZ | [\(520\) 302-4921](tel:(520)302-4921) | j.carroll@techsoftsys.com

github.com/brandxx | www.techsoftsys.com

[Online resume](#) / [Cover letter](#) / [Research](#)

Applied AI Research

Independent research through TechSoft Systems in AI infrastructure, LLMOps, retrieval evaluation, stateful platform reliability, agent systems, and evidence integrity. Project leadership, hands-on integration, and AI-assisted development; implemented lab systems, operational personal deployments, and design-only proposals retain separate status.

[Dynamic GPU Residency & LLMOps](#)

[Stateful AI Platform Operations & Reliability](#)

[PIPER Intent Gateway](#)

[Embedding & Retrieval Evaluation](#)

[Open-WebUI Model Preloader](#)

[Inference & Platform Engineering](#)

[NVIDIA AI-Q Research Assistant](#)

[Production Agent Audio and Voice Integration](#)

[Ogden / Memory Workbench](#)

[AI Lab NVMe Storage Engineering](#)

[Reasoning Reliability Research Proposals](#)

Dynamic GPU Residency & LLMOps

Inference architecture and operational validation

Public project

Historical lab tests; residency components validated

Designed a private multi-model, multi-agent inference architecture and experimentally validated dynamic GPU residency components using vLLM and NVFP4-quantized MoE models. Evaluated retained-weight transitions under embedding load and controlled host-memory bandwidth contention.

Historical August/September 2026 independent lab results, not whole-architecture production qualification or confirmation of restored dual-GPU availability after RMA. Full-switch, target-wake, and paired-wake timings have different boundaries. Long-duration near-saturation contention testing remained outstanding; the Responses API was excluded from the Lightning acceptance campaign because of a separately tracked streaming defect. Upstream fix authorship is not claimed. Architecture name and internal design remain private.

- Measured strictly sequential source sleep, retained target wake, and streamed Chat Completions through the first nonempty visible token. Both short, nonthinking bidirectional probes completed their required response marker.
- Validated a pinned vLLM runtime incorporating an upstream sleep/wake fix. Distinguished cold host-mapping setup from warmed transitions, checked arithmetic/tool output after wake, and retained the previous runtime for rollback.
- Tested synchronized wakes across two GPUs while an embedder was active, then ran a separate five-level STREAM memory-bandwidth contention matrix with paired-wake timing and post-wake arithmetic gates.
- Preserved a failed initial Nano arithmetic gate and its zero-pressure reproduction before completing the replacement stable-gate matrix. Later numerical failures remain qualifications; successful residency does not establish general model correctness.
- Separated model-local KV capacity from per-request context and scheduler concurrency. The four deployments in the capacity snapshot were single-GPU TP1 engines on a two-GPU host, not a successful TP2 configuration or a cross-model shared KV pool.

Bidirectional model switching

0.960760 / 0.973000 seconds

September 1 selected Omni-to-Lightning and reverse tests. Wall time from source Level-1 sleep start through the target first nonempty visible content delta, including state verification; short nonthinking streamed marker requests.

Pinned-runtime sleep/wake acceptance

61/61 cycles; 122/122 checks; approx. 473 ms mean wake

Separate September 1 Lightning acceptance campaign with vLLM 0.28.1rc1.dev130+g44fe2a392. The dedicated 50-cycle soak is included in the 61-cycle total. First sleep established retained host mappings; Responses API testing was excluded.

Synchronized dual-GPU wakes

1.630-1.634 seconds

Median pair completion for four three-cycle conditions (twelve pair cycles overall) with an active embedder. Concurrent target-wake timing, not source-sleep-to-first-token or general agent-task latency.

Host-memory bandwidth contention

274.7 GB/s STREAM; 3.5-3.9% higher median pair wake

Separate near-saturation matrix: paired-wake medians rose from 1.625/1.622 to 1.682/1.685 seconds. Individual smaller-model wakes rose approximately 10%. Bandwidth stress, not RAM-capacity exhaustion or proof of PCIe saturation.

Final stable-gate contention matrix

60/60 arithmetic checks

Final replacement-gate matrix with no observed restarts, matching GPU/kernel faults, or swap use. Earlier failed Nano gates and later numerical failures remain part of the record; this is not universal model correctness or a long-duration soak.

NVIDIA AI-Q Research Assistant

Agentic research and evidence integrity

Public project

Implemented private research platform; evaluation ongoing

Built a self-hosted AI-Q/NeMo research platform with persistent jobs, cited reports, and reproducible evidence evaluation. Diagnosed and corrected a specific citation-accounting defect, validated with 196 focused passing tests.

Active independent AI-assisted research, not commercial deployment. Token/call counts cover selected retained work, not lifetime totals. Software tests and execution journals do not establish factual correctness; bounded comparisons do not establish overall model superiority or reliable general research.

- Integrated NVIDIA AI-Q, NeMo Agent Toolkit, FastAPI, Dask, PostgreSQL, Next.js, vLLM, and Ollama for persistent research jobs, cited reports, and streaming progress.
- Built bounded cross-model evaluation with frozen evidence, positive/negative controls, deterministic source-passage identifiers, immutable inputs, and hash-chained execution journals.
- Designed typed request, budget, and evidence contracts to make research jobs traceable and replayable. Separated successful software execution from accepted research quality.
- Investigated source-binding, structured-output, and agent-handoff failures; tested source-first comparisons in independent contexts and identified limits in domain-specific controllers and generalization.
- Diagnosed a citation-accounting defect and restored 30 cited sources from a 77-source report in an exact-image replay; validated the repair with 196 focused passing tests and regression comparisons.
- Analyzed 4.68M+ reported model tokens in selected retained studies and 1,448 calls across 45 diagnostic runs. These supplied study scopes are not lifetime totals or a newly reconciled aggregate.

Stateful AI Platform Operations & Reliability

Inference migration, recovery, and integration engineering

[Public project](#)

Implemented independent lab operations; historical functional checks

Migrated a stateful Hermes/Honcho assistant platform from Ollama GGUF to vLLM-hosted NVFP4 inference while preserving 3,045 active memory documents and the existing embedding service. Developed recoverable deployment procedures and tested the actual agent paths for browser, vision/video, speech, and other native tools.

Independent AI-assisted lab engineering, not a commercial or Tribal deployment. The 3,045-document count describes the recorded migration, not a current inventory or the separate memory-review cohort. Functional checks do not establish visual accuracy or a security certification; LLM/TTS/STT inference was not repeated in the October 1 update. No measured uptime, SLA, recovery-time objective, or general fault-elimination claim. Private configuration, credentials, and architecture remain undisclosed.

- Coordinated service cutover with PostgreSQL, Redis, assistant, and Memory Workbench backups, preserving memory and review progress. Used stopped-state snapshots, retained container images, verified Git bundles, SHA-256 manifests, and database integrity checks.
- Separated image-only rollback from full-state restoration so application recovery would not unnecessarily discard newer conversations or review work. Diagnosed persistent Python native-extension/ABI and missing declared dependency failures that an image revert alone did not resolve.
- Repaired service-user ownership, protected configuration access, and stale MCP registry problems through scoped recovery actions rather than broad state replacement.
- Corrected a false-positive browser smoke check by following the agent's actual managed-browser resolver and readiness contract, rather than treating a cached Chrome launch as proof of usable agent tooling.
- Repaired unavailable auxiliary-model dependencies and verified native image/video tooling without frame-extraction workarounds. A September 30 functional audit covered terminal, file, browser, search/extraction, image/video, and local speech paths; the October 1 update used narrower health, catalog, browser, and deterministic audio checks.

Production Agent Audio and Voice Integration

Model-directed audio and operational engineering

[Public project](#)

Operational personal deployment; owner-confirmed October 2026

Designed and deployed a self-hosted audio stack for a personal AI assistant using Hermes, Whisper, Qwen3 TTS, and Kokoro. Built model-callable media tools and evidence-aware guidance that separate speech transcripts, deterministic acoustic measurements, and visual inference.

Owner-operated personal production confirmed October 10, 2026; exact cutover date, commercial scale, and Tribal deployment are not established. Functional results are dated historical checks, not new benchmark runs. Processing limits are not full-length success measurements; timestamps are segment-level. No standardized speech-accuracy, latency/SLA, native hearing, DTMF decoding, cloning, or diarization claim. Specific security checks do not establish general prompt-injection immunity. Internal deployment details and raw evidence remain private.

- Directed architecture, integration, operations, and acceptance for a three-tool native audio plugin and companion skill. Implementation and testing used AI coding assistance; the speech models and Hermes framework remain upstream components.
- Added audio-track selection, source-relative ranges, sequential five-minute 16 kHz mono PCM chunks, approximate speech-segment timestamps, and explicit continuation on transcript output-budget exhaustion.
- Separated Whisper speech recognition from FFT-based tone and timing analysis. Recorded controlled rising/falling/constant-tone, silence, speech, track-selection, and fresh-session native-tool checks using fixture manifests with expected results and tolerances.

- Restricted media decoder protocols and formats, rejected playlists, retained protected-file and approved-output controls, and documented specific unsafe-input/output rejection tests. Agent guidance treats recorded speech as evidence, not instruction authority.
- Integrated profile-scoped Qwen3 TTS through native speech providers while preserving existing Kokoro output; verified MP3 generation and 24 kHz PCM through the actual Hermes read-aloud path.
- Maintained custom capability through backend upgrades with registration and deterministic DSP checks, configuration fingerprints, stopped-state backups, retained images, and documented narrow rollback procedures.

PIPER Intent Gateway

Intent classification and memory routing

[Public project](#)

Implemented advisory lab integration; historical regression coverage

Architected a Hermes advisory intent/memory integration with deterministic contracts, monitoring, human review, and failure analysis. Preserved a historical 114-test intake milestone and a separate selected 106-decision audit.

Independent R&D, documented May 2026 - Present. The native plugin is advisory, not autonomous memory enforcement. Historical packages do not establish feature parity across generations. The selected audit is failure analysis, not a general accuracy score; historical tests are not a full current-plugin passing run.

- Led AI-assisted integration research across early intake tooling, Open-WebUI filters, and a later native Hermes advisory plugin. Historical work used structured output, contract validation, failure tracing, and rollback.
- Separated advisory classification from agent execution, using schema validation and output checks while retaining evidence and explicit uncertainty in handoffs.
- Evaluated retrieval and memory-selection behavior without treating model output or remembered content as independent evidence of correctness. Detailed routing and memory architecture remain private.
- Built gateway monitoring and review tooling with FastAPI/React, authenticated source-turn review, append-only corrections, revision-checked configuration, atomic replacement, backups, and rollback.
- Expanded the early gateway acceptance suite to 114 reported passing tests in May 2026. Re-reviewed 106 previously recovery-labeled decisions in September to separate memory-routing errors, classifier misses, internal turns, correct routes, and downstream recovery.
- Retained historical regression source files and a test log supporting the early 114-test milestone. Investigated dependent-hop latency, oversized handoffs, nested timeouts, output-contract failures, and test traffic entering durable memory; those artifacts do not establish a current native-plugin passing suite.

Ogden / Memory Workbench

Proposal-only memory evaluation

[Public project](#)

Implemented review tooling; model proposals do not mutate live memory

Evaluated 280 conclusions across 72 cases, completed after retries and serving recovery; produced 19 proposals for human review without mutating live memory.

Independent AI-assisted research. The 19 proposals are not 19 proven errors or applied fixes. Semantic clustering, a full memory compiler, and compaction concepts remain research/planned work, not completed features.

- Designed and evaluated a proposal-only semantic memory reviewer with complete-artifact validation, idempotent imports, immutable model output, separate human edits, and append-only review events.
- Revalidated source records at review time and blocked approval on source drift. Kept live-memory application disabled, with auditable exports and human review separate from model-generated proposals.

- Established provenance-aware export boundaries: retained model conclusions are unreviewed, not human-approved training examples. Schema-valid output remains a proposal.
- Implemented exact deduplication that reduced a 200-conclusion sample to 111 review rows while retaining source identities and audit metadata.
- Researched governance distinctions between confidence, authority, and purpose, including privacy risks from cumulative profile inference; broader runtime enforcement remains planned.

Embedding & Retrieval Evaluation

Quantized serving, retrieval quality, and evidence-aware evaluation

[Public project](#)

Historical matched lab comparisons; workload-specific results

Evaluated four 1B/8B BF16 and NVFP4 embedding releases under matched serving conditions, combining throughput, latency, vector integrity, retrieval quality, and sampled GPU-memory checks. Kept serving performance separate from retrieval relevance and label quality.

Historical independent AI-assisted lab results, not newly rerun benchmarks. The 2x serving comparison applies only to the recorded 8B workload; zero-error soaks do not establish a production SLA. Serving, public-retrieval, and memory-corpus evaluations are distinct scopes, not a combined accuracy score or an overall model ranking. Unreviewed labels are not gold labels; no statistical equivalence or human-approved training dataset is claimed.

- The 8B NVFP4 candidate recorded 32.0 versus 16.0 inputs/sec in the matched mixed-length, eight-worker soak. Across four 60-second soaks, 20,936 embedding requests completed with zero recorded errors; the rate comparison and total request count describe different scopes within the same four-release study.
- At 2,048 dimensions, three-dataset retrieval means were 0.6183 for NVFP4 and 0.6185 for BF16 across SciFact, NFCorpus, and ArguAna. The retained summary does not identify the exact metric for those means; close values are not proof of statistical equivalence.
- Evaluated a separate memory corpus of 214 sessions and 1,061 candidates using source-linked, unreviewed labels and paired bootstrap comparisons. Treated model-derived memory and source-link labels as evaluation context, not human-adjudicated ground truth.
- Investigated GPU contention, role-aware query/document formatting, and consistent vector truncation with L2 normalization. A separate public-dataset and 48-query hard-negative technical comparison showed workload-dependent model tradeoffs; truncation reduced stored vector payload, not inference memory or network payload in that run.

AI Lab NVMe Storage Engineering

High-throughput storage and reproducible I/O measurement

[Public project](#)

Engineered lab storage; historical sequential direct-read tests

Engineered two independent four-drive PCIe 5.0 NVMe RAID0 arrays for the AI lab. Repeated fio sequential direct-read testing against prepared 100 GiB files recorded 53.21 GB/s on the 8 TB array and 55.24 GB/s on the 16 TB array, with each array tested independently.

Historical owner-reported read results, not a new benchmark or independently reproduced performance test. Rates are per array, not combined throughput or write performance; no PCIe-bus saturation or application-level speedup is claimed. Fault checks cover the recorded test window. RAID0 provides no redundancy and is not presented as a backup or resilient production-storage design. Current hardware availability is not established by these results.

- The 8 TB array (4 x 2 TB) recorded 53.21 GB/s / 49.56 GiB/s at approximately 50,748 IOPS; the 16 TB array (4 x 4 TB) recorded 55.24 GB/s / 51.44 GiB/s at approximately 52,678 IOPS. Kept decimal and binary units distinct.

- Both tests completed with zero fio errors. All eight NVMe devices reported zero SMART critical warnings or media errors, no matching kernel faults were observed during testing, and all 54 running container workloads remained intact.
- Retained the direct-I/O workload, prepared-file size, per-array scope, and approximately 90-97% system CPU utilization per job. The CPU observation suggests submission overhead may constrain the single-job workload; it does not establish a specific bottleneck or PCIe saturation.

Open-WebUI Model Preloader

Browser-side AI workflow integration

[Public project](#)

Built and used in the lab; performance gains not benchmarked

Built and used a JavaScript/Tampermonkey integration that initiates model warmup on Open-WebUI selection, aiming to overlap initialization with user interaction before the first prompt.

Implemented independent lab software, not an upstream Open-WebUI contribution or commercial deployment. No retained before/after latency or GPU-overhead benchmark establishes cold-start elimination. Client cancellation does not guarantee backend work stops. Public repository, license clearance, and current-version compatibility are not established by the supplied artifacts.

- Implemented debounced DOM-driven selection handling, deterministic model-ID mapping, client request cancellation, cooldown/force-warm controls, authenticated API fallback, and persisted state.
- Maintained versioned source, architecture documentation, change notes, troubleshooting guidance, and a structured manual-validation checklist.
- Preserved implementation caveats and separated confirmed lab use from static source checks, unperformed runtime tests, and roadmap features.

Reasoning Reliability Research Proposals

Research design and evaluation planning

[Public project](#)

Proposal and design only; no combined-system experiments

Authored technical proposals on LLM reasoning reliability, with planned evaluation of correctness and computational cost. The work remains proposal and design only, with no combined-system implementation, experiments, or measured gains.

Proposal/design only, as confirmed by the author. No combined-system implementation, experimental improvement, safety guarantee, or publication is claimed. Papers, diagrams, and internal architecture details remain private; this entry is a high-level abstract only.

- Defined evaluation questions, controls, failure modes, and comparison criteria before implementation, keeping intended benefits separate from measured results.
- Distinguished model confidence from correctness and proposed independent artifact validation for agent repairs.
- Credited existing research methods; the contribution is synthesis and integration design, not an established first-ever algorithm or published/peer-reviewed paper.

Inference & Platform Engineering

AI infrastructure, LLMOps, and reproducible operations

[Public project](#)

Implemented lab serving; historical workload-specific benchmarks

Convert base models to NVFP4 using NVIDIA Model Optimizer, validate and optimize the converted checkpoints, and deploy through vLLM. Work includes FP8 KV caching, MTP speculative decoding, prefix caching, chunked prefill, CUDA graphs, and model residency. Personal conversion experience is distinct from deploying publisher-supplied NVFP4 checkpoints.

Integrate embedding/reranking services, local speech, and multimodal workloads. LLMOps includes pinned runtime images, functional acceptance checks, model-local capacity planning, observability, verified backups, state-preserving container updates, and narrow rollback procedures. Ollama remains for basic testing and selected embedding workloads.

Historical local-lab results use different workloads, timing definitions, and recorded runtime configurations. They are not direct cross-model rankings, universal model-performance claims, or guarantees on the current runtime. Residency acceptance-phase measurements do not replace these throughput campaigns. Detailed source artifacts and internal configuration remain private.

- Designed reproducible inference comparisons using matched decoding settings, immutable model/runtime identities, bounded concurrency, and explicit timing definitions; retained failed attempts and workload-specific acceptance limits.
- Investigated unsuccessful TP2 attempts and tested a bounded single-GPU serving configuration for Nemotron Super. Workspace and retained-VRAM headroom remained narrow. Distinguished per-sequence context limits from each engine's own KV capacity; startup capacity did not establish simultaneous maximum-context workload support.
- Evaluated role-aware embedding formats against public retrieval datasets and a separate hard-negative technical suite, informing workload-specific model selection rather than an overall quality ranking.
- Used infrastructure and GPU telemetry with Prometheus, Grafana, Loki, and NVIDIA DCGM Exporter, alongside fault logs and post-update checks. Monitoring coverage and bounded fault-free tests are not a production SLA.

Nemotron Super 120B NVFP4

212.20 output tokens/sec

Warmed single-request technical workload; end-to-end rate across the configured generation, not decode-only throughput.

Nemotron Lightning 30B-A3B NVFP4

568.15 output tokens/sec

Mean of five warmed synthetic runs; 59.6% above the measured target-only control on the recorded runtime.

Qwen3.8-27B NVFP4

133.64-164.12 output tokens/sec

Median single-request decode rates across recorded prompt workloads; 914.20-1,130.73 aggregate output tokens/sec across eight concurrent requests.