AI Research and Systems Portfolio

Cisco Caceres · Independent AI systems researcher focused on reliable machine agency; systems architect. Evidence and method summary updated 7 October 2026, America/Los_Angeles.

Grounded in decades of hands-on systems engineering. How can models become systems that perceive, reason, act, coordinate and recover under real constraints, with verification and measured cost?

I design, build and evaluate AI systems across speech, language models, agents, distributed infrastructure and physical hardware. My research asks how those systems behave under noise, limited compute, unreliable infrastructure and imperfect verification. My work combines model experiments with the engineering needed to execute, measure and operate them.

This portfolio links inspectable artifacts and distinguishes completed experiments, operating implementations and proposed studies. Most recent measurements are self-published and internally evaluated. Independent replication and peer review are separate milestones.

How I test a research claim

I start with a testable hypothesis, compare it with simpler alternatives, and keep the evidence that changes my conclusion.

1. **Define the comparison.** State the hypothesis, task and failure criterion. Compare simpler baselines under matched information and budgets; distinguish my implementation work from upstream models, data and methods.
2. **Freeze conditions before outcomes.** Record exact source, model and data identities, splits, scoring rules, stopping decisions and resource limits. Keep development and calibration separate from held-out evaluation.
3. **Retain and expose the evidence.** Keep failures, abstentions, unrun cases and unknown costs in the registered accounting. Link versioned artifacts and descriptive uncertainty, and state which records an authorized reviewer can actually rerun.

AMBIE: revise the assumptions: The acoustic evidence changed planning priorities; third-party benchmark results do not establish AMBIE-trained accuracy.

Verifier selection: inspect the headroom: A completed engineering pilot exposed a limited candidate pool. It did not test the separate joint-verifier hypothesis.

Echo: separate coordination mechanisms from outcomes: The reused-fixture coding diagnostic remains negative; separate bounded path-lease and crash-recovery checks now test coordination primitives, not coding-task benefit.

Information limits: state the assumptions: A short fixed-pool note proves a ceiling on what any selector can gain from verifier signals and lists seven ways it fails to transfer to real policies. It reports no model result.

A source-only guard, a load-qualified runtime and a completed task experiment support different claims. My self-published reports and same-workflow reviews remain separate from outside peer review and independent reproduction.

Selected engineering experience

Speech and radio systems. Speech-recognition products since approximately 2012; CTO work across hearing technology, firmware, clients and APIs in 2017–2018; co-founder/CTO of a voice-AI company from 2018–2024. The work included U.S. Coast Guard and DHS radio deployments and a hardware interface that allowed software to key a radio transmitter. These deployment descriptions are my reported professional history. Chronology and current speech evaluation.

Distributed systems. Architecture and implementation work across publishing infrastructure, identity and large-scale messaging. The historical ZettaZing platform was load-tested to 30 million sustained concurrent connections across approximately 3,000 AWS instances; that is a load-test claim, distinct from production users or a current AI benchmark. Selected work and contribution.

Product and technical leadership. Founder/CTO work includes Abundera's application federation and identity/integration architecture, alongside technical diligence through CodeProvenance. Cloud services, financial-data APIs, cryptographic primitives and model providers are dependencies; the engineering contribution is system design, implementation, evaluation and operational ownership. Professional record and architecture and diligence practice.

Model training, adaptation and evaluation

TuneHarness includes substantial completed experiments from its former VastOps period. Work spans operational triage and log specialists, structured tool calls, document field extraction, text-based speech-turn completion, speech benchmarks and small transformers. Completed method comparisons include prompted baselines, fine-tuning, LoRA, quantization, vocabulary trimming, and prompt/schema ablations. The matrix also documents an implemented structural-pruning/recovery route refused by the planner before training.

I built experiment infrastructure, run accounting, evaluation gates and specialist workflows around third-party base models and learning libraries. Documented two-GPU runs establish single-node multi-GPU execution; each comparison retains its dataset, simulator, budget and hardware limits. The next publication milestone is reproducibility and validation on independent real data, rather than a symbolic first training run. Experiment history and scope.

Research projects and evidence

TUNEHARNESS

Status: Operating research infrastructure · Documented ML experiments. **Current artifacts:** Training and provisioning tooling, a dated cost ledger, model experiment matrices, and per-arm ablation reports, including the former VastOps work. Completed work includes saved-output tool-calling comparisons, a 120-task local verifier pilot, native FunctionGemma checkpoint diagnostics, a 160-case authored full-schema comparison, and bounded native environment export and restoration checks. These are distinct experiments and prerequisite stages; their full numerical record is retained below.

Research question: How can specialist-model training, adaptation and evaluation improve task outcomes within measured compute budgets?

Evidence scope: The comparisons show task-specific differences and failures, including FunctionGemma invocation collapse in a lossy native condition and a separate JSON endpoint audit. Training-content overlap, earlier probe exposure, omitted schemas, authored fixtures, and unidentified original SFT lineage prevent

causal training-benefit or broad generalization claims. Single-node multi-GPU work does not establish multi-node or arbitrary-scale training or marketplace-wide rates. Environment export and finite restoration checks qualify prerequisites, not serving identity, exhaustive ELF closure, collector admission, or a verifier-study result. The complete dated record and unresolved dispositions remain below.

Next milestone: Validate specialist findings on independent real data and held-out domain shifts with a frozen protocol.

Evidence and method

Full experiment and prerequisite record

VERIFIER RELIABILITY

Status: Research proposal · 120-task engineering pilot complete; H1 study unrun. **Current artifacts:** Updated proposal, related-work matrix, task-label manifest, guarded local collector, networkless VM outcomes and public aggregate results for the separate 120-task HumanEval engineering pilot (480 candidates and 480 verifier calls). The earlier 504 task-metadata rows and 4,032 candidate records are retained as a seeded synthetic simulation: candidate correctness, public-test outcomes, verifier scores, costs and latencies were generated, not measured. No models or benchmark tests ran. The three named splits share the same 14 domain labels and are not a held-out-family evaluation. The stated 0.85 early-stop threshold is unreachable under the implemented score formula. The prospective 576-slot H1 collection remains unrun.

Research question: Under a fixed total inference budget and task-family shift, does modeling cross-verifier error dependence improve hidden-test success when selecting candidates and deciding when to stop, beyond strong single-judge, fixed-consensus and matched-retention filter baselines?

Evidence scope: The separate 6 October 2026 HumanEval pilot passes 109/120 tasks with the first candidate; higher-cap SoftBoN reaches 91.39%, a descriptive +0.56-point difference with interval [0, +1.67]. Candidate-pool headroom is only one task in that pilot. The earlier 504-task Pillar 1 figures (including covariance, pass rates and cost outputs) are withdrawn as empirical evidence: source audit shows they are deterministic outputs of a seeded synthetic simulator using benchmark task metadata. They do not measure model/verifier behavior, code correctness on benchmark tests, held-out-family generalization or a VEC-SCR gain. The simulator's three splits reuse the same 14 domains; its 0.85 stopping threshold cannot be reached (the implemented public-pass score is at most 0.75), so no early stopping or token conservation occurred. The joint-verifier and natural-shift H1 study remains unrun.

Next milestone: Reconcile serving model/tokenizer/template identity, qualify and execute the frozen 576-slot development collection locally, then fit source-only controllers and matched baselines, measure candidate-pool headroom and cost, and freeze the preregistration before any confirmatory run.

Evidence and method

Separate 120-task HumanEval pilot: methods, results and limits

AMBIE

Status: Model research · Bounded synthetic-input speech diagnostics recorded. **Current artifacts:** Dated third-party speech benchmark, reviewed planning assumptions, commercial-engine passthrough service, E1 and E2 per-model/condition output ledgers and a separately released 1,050-file TRCS fixture package. The release and the smaller evaluation input sets have distinct denominators.

Research question: Which acoustic and deployment choices make noisy-speech recognition more reliable under radio and environmental constraints?

Evidence scope: E1 retains 27 authored synthetic input clips across 6 models and two decoding regimes (324 output evaluations). Overall per-model ungated rates are 24–55 words per minute; 160 is a condition-cell peak. The gate reduces overall WPM by 19.5–25.0%, leaving hallucinations on 108/144 non-speech evaluations. E2 retains 5 authored synthesized utterances across 30 declared conditions and 6 models (900 model-by-condition records). Mean WER across recorded models and paths is 129.4% at 0 dB. These are repeated views of a small synthetic input set, not independent samples or authentic Coast Guard recordings. No population confidence interval, genuine-voice preservation, codec-compliance guarantee, operational-radio accuracy or proprietary AMBIE model accuracy is established.

Next milestone: Retain model, launch and codec-path receipts; freeze licensed, speaker-diverse real-domain inputs and critical-field scoring before assessing proprietary or operational performance.

Evidence and method

THERMOCOG

Status: Whitepaper and design · Implementation pending. **Current artifacts:** Written whitepaper and design materials; public project description.

Research question: What does machine intelligence cost in energy, and which proposed computational mechanisms can be tested against physical constraints?

Evidence scope: No software or hardware implementation and no measured results. Proposed energy figures are not demonstrated performance.

Next milestone: Define a minimal falsifiable experiment, implement it and measure it against a conventional baseline.

NOISE OR VOICE

Status: Experimental acoustic research · DSP and SSL diagnostics completed. **Current artifacts:** Browser demo, feature analysis, internal test methodology and documented failures. Completed 64-record ODSS and original 192-record ASVspoof 5 diagnostics; full acquisition, PCM/WAV replay and blind scoring deposits precede label analysis. Aggregate reports retain false flags and misses. A separate developer-selected, known-label 168-trial paired diagnostic compares the frozen JavaScript DSP rule with an author-released SSL checkpoint; the committed per-trial ledger reconciles all overall and dataset-level counts.

Research question: Can interpretable acoustic features help distinguish synthetic speech across generators, speakers and recording conditions?

Evidence scope: Scores are uncalibrated and verdicts do not authenticate speakers. The ODSS diagnostic evaluates a pinned historical JavaScript research snapshot, not automatically the current deployed detector. Its four test-speaker clusters and descriptive interval are fragile. Earlier human-only failures remain separate; two-class results do not validate the detector, streaming, radio use or general-population performance. ASVspoof selected labels cover two base TTS families, a related adversarial derivative and shared vocoder. In that separate internal paired diagnostic, DSP classified 123/168 correctly versus SSL 140/168. DSP falsely flagged 30/65 human recordings and missed 15/103 synthetic recordings; SSL had 26/65 false flags and 2/103 misses. DSP analyzed complete WAVs while SSL used the first 64,600 resampled samples or repetition/padding. Nonrandom selection, previously exposed labels, related sources, unresolved upstream label custody and training exposure prevent blind-validation, population-performance or general-superiority claims. This is separate from the original 192-record ASVspoof diagnostic.

Next milestone: Independently reproduce the historical-detector diagnostics, qualify the current deployed implementation separately, and extend acoustic coverage under a new prospective protocol. Resolve synthetic-content reuse scope, source linkage and training exposure before broader claims.

TUNEHARNESS FULL EXPERIMENT AND PREREQUISITE RECORD

Artifacts: Training-run tooling, dated provisioning ledger, model experiment matrices, ablation reports and per-arm metrics, including the former VastOps work. Exact saved-output rescore, recovered ablation provenance, local verifier diagnostics and a checked mathematical reference. A completed 120-task local verifier pilot with public aggregate analysis and explicit evidence limits. A separately reconstructed names-ablation CPU contract probe completed eight cases: six correct, all eight parseable and strict-format compliant. This is not the missing original SFT checkpoint or a historical generation reproduction. A completed matched 400-record comparison retained 69.7222% class-balanced correctness for the reconstructed names-ablation versus 34.7222% for publisher Qwen3-1.7B; pooled correctness was 254/400 versus 74/400. A separate FunctionGemma base/checkpoint-600 diagnostic qualified finite tensors and repeated synthetic forward outputs, not task accuracy. A later tokenizer-only native fit retained all 400 historical FunctionGemma inputs without oversize; a separate incomplete schema inventory covered 381 records with 19 unsupported, not task accuracy. A completed same-lossy-native FunctionGemma comparison retained both 400-case arms: base 228/360 call and 0/40 no-call correct; checkpoint-600 0/360 call and 40/40 no-call. This was checkpoint invocation collapse, not improved tool calling: macro 31.6667% to 50%, pooled 57% to 10%, with eight checkpoint output truncations. A subsequently completed publisher full-schema + instruction control retained 235/360 call and 27/40 no-call correct, macro 66.3889% and pooled 262/400, with 138 errors. A separately frozen argument-sensitivity rescore reconciled all original labels and changed zero of 1,200 correctness labels across the three Qwen conditions. A post-hoc retained-output audit subsequently replayed all 800 native labels unchanged. Exact stored strict JSON scored 0/400 in each arm. Removing exactly one terminal <eos> only when the recorded stop ID is 1 gave checkpoint-600: 159/360 call and 36/40 no-call correct, pooled 195/400 (48.75%) and class macro 67.0833%; base remained 0. Typed/legacy equality agreed, and all eight call truncations stayed incorrect. This alternate JSON endpoint qualifies blanket incapacity interpretations, not the original native invocation-collapse table; it establishes no causal improvement, equivalent full-schema comparison or external peer review. A later matched full-schema and instruction diagnostic on 160 authored structural fixtures retained reconstructed names-ablation 160/160 versus publisher Qwen3-1.7B 132/160 correct (strict contract validity 160 versus 137). All 320 requests completed without truncation. These fixtures were authored after prior findings; unresolved training lineage and toy structure prevent causal training-benefit or frontier-generalization claims.

Evidence scope: Completed experiments retain task, dataset, simulator and hardware limits. The reported structural-pruning/recovery arm was planner-refused before training. Single-node multi-GPU execution does not establish multi-node or arbitrary-scale training; provisioning samples are not marketplace-wide rates. The eight-case probe remains separate. The completed names-only comparison retains five exact training-content overlaps and eight earlier probe exposures; omitted argument schemas may favor a catalogue-trained specialist. The original selected SFT remains unidentified. This establishes neither causal training benefit, general tool-calling superiority, frontier transfer nor production readiness. FunctionGemma evidence includes load/finite-forward qualification, native input fit and incomplete schema inventory; these earlier load/fit/audit stages did not establish task accuracy. The later completed base collection remains a separate exposed, lossy-native condition; the completed comparison showed checkpoint invocation collapse with zero correct calls and eight truncations. No-call measures invocation absence, not semantic refusal quality; the cause is unknown from aggregate review. Equivalent full-schema information is not established. Checkpoint-400 remains truncated. Full schemas and instruction both changed; the exposed historical prefix and different model catalogue priors remain. This follow-up is not causal training benefit or fresh generalization; strict argument sensitivity is distinct from strict-format validity. The original native invocation-collapse finding is endpoint-specific; the later one-terminal-EOS JSON audit does not replace it or establish

general model incapacity. The 160-case follow-up is authored after earlier findings, not a blind broad benchmark or proof of causal training benefit. A V5.1 native environment diagnostic matched 168/168 pinned package versions but failed 1 of 12 ABI/import checks: Numba 0.55.0 could not import pkg_resources with setuptools 84.0.0 pinned. All 24 references and 9 controls were skipped; no H1 evidence resulted. A prospectively amended V7.2 run changed only runtime setuptools to 81.0.0: all 168 package identities, five source builds, twelve fixed ABI checks, 24 references and nine expected control outcomes qualified. No model calls occurred. V7.2 retained only the runtime inventory. A subsequent V5 run retained 2,646,208,612 regular-file bytes and 45,730 runtime entries in private host custody after repeated full prerequisite qualification and owned guest shutdown. A subsequent owned, networkless restoration replay reconciled eight registered native-file byte identities, all 168 package versions, 12 ABI checks, 24 references, nine registered controls and the unchanged 45,730-entry retained tree. The disposable VM was removed. Five historical source-build proofs were reused, not rebuilt. These finite checks supersede the earlier pending restoration descriptions for that replay; they do not establish exhaustive ELF dependency closure, serving-model/tokenizer/template identity or collector admission. The earlier 1,850 unresolved native metadata dispositions remain retained. Candidate supervision and real collection qualification remain pending. This is prerequisite qualification, not a verifier-study result.

Agent, infrastructure and learning implementations

Echo. I built coordination infrastructure for coding agents: durable messages, leased tasks, advisory file reservations, persisted results and recovery mechanisms. NATS and the model/agent providers supply underlying components. Dated transport and API-worker benchmarks measure their stated workloads. The first local-model diagnostic on eight original synthetic task pairs retained all 48 attempts; none passed pair acceptance. A separately qualified V2 contract completed another 48 attempts on the same exposed fixtures: single 8/16, ordinary two-worker messaging 0/16 and Echo 0/16. All 13 merge conflicts and complete usage are retained. These development results establish no coordination benefit or fresh benchmark claim; a fresh controlled task study remains planned. Architecture and measurement record.

Bootscry. I integrate physical observation and action through a Comet KVM, OCR, HID input and operational tooling. Published evidence includes ARM-board bring-up and older x86 installation/rescue, with OCR misses, hangs and manual power interventions retained. Libre Computer setup used a card prepared on another host. Physical read-tier evidence and mock-tested action controls have different scope. Hardware evidence and ARM bring-up.

LLM Lab. I implement interactive explanations and small-model learning mechanics: tokenization, backpropagation, attention, training, sampling and inference state. The code uses established architectures and methods. This makes implementation mechanics inspectable; it is an educational artifact, distinct from a novel model-method result. Open the lab.

Model access and retrieval. My internal gateway/router work combines provider interfaces, routing policies, budgets, retries and traces. SmartChannels explores persistent state and retrieval across communication interfaces. Held-out task studies are needed to quantify routing or retrieval benefits. Systems and evidence boundaries.

Publication program

1. **Reproduce existing specialist-model experiments.** Publish frozen data/model manifests, executable environments, baselines, multiple seeds, ablations, uncertainty and complete cost denominators. Add independent real-data and domain-shift validation before a generalization claim.

2. **Evaluate acoustic generalization.** Extend noisy-speech and synthetic-speech work with licensed, speaker- and generator-separated evaluation, fixed thresholds, meaningful error costs and independent reproduction.
3. **Test verifier reliability under budgets.** Baseline feasibility and a 120-task engineering pilot are completed. Next, develop diverse candidate pools and matched baselines for a separately frozen joint-verifier hypothesis study with new task families. Proposal and protocol.
4. **Measure coordinated agent outcomes.** Compare Echo with simpler isolated-worktree arrangements, retaining failures, merge work, operator intervention and cost per accepted task.

Completed diagnostics are identified above; the remaining reports and hypothesis studies are planned. Research relevant to frontier models can investigate inference, post-training and evaluation; this portfolio does not claim frontier-scale pretraining, a new foundation-model architecture or peer-reviewed acceptance.

How to assess the work

For each contribution, inspect the question, my design/implementation role, third-party methods, data provenance, exact executed conditions, failures and current limits. Ask for runnable permitted artifacts and an independent rerun. A test count establishes behavior covered by those tests; a polished report does not substitute for experimental evidence.

The current research agenda, AI systems page and linked project records provide the dated evidence trail. Contact Cisco for technical collaboration, artifact access or verification of professional history.