

AI Research and Systems Portfolio

Cisco Caceres · AI systems architecture and independent AI research. Evidence snapshot: 6 October 2026, UTC.

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.

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-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.

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

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.

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

VERIFIER RELIABILITY

Status: Research proposal · Engineering feasibility measured. **Current artifacts:** Updated proposal, related-work matrix, frozen HumanEval splits, guarded local collector and networkless VM development qualification.

Research question: How do correlated verifier errors and calibration drift affect budgeted code-generation decisions?

Evidence scope: V2 retains fourteen response-contract failures; V3 has twelve public/hidden passes on six development tasks. This saturated set establishes no verifier-selection benefit, specificity, study hypothesis result or novelty.

Next milestone: Run the separately frozen 120-task pilot with matched-token baselines and retain all contract failures and costs.

AMBIE

Status: Model research · Live commercial-engine service. **Current artifacts:** Dated third-party speech benchmark, tested architectural assumptions and a live commercial-engine passthrough service. A frozen speaker-separated local third-party ASR diagnostic and insertion-heavy failure analysis.

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

Evidence scope: The published benchmark evaluates third-party models. It establishes no AMBIE-trained model accuracy. Four of five initial architectural assumptions were falsified.

Next milestone: Extend the frozen acoustic protocol to licensed domain-relevant channels and critical fields before assessing a proprietary checkpoint.

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 · Internal evaluation. **Current artifacts:** Browser demo, feature analysis, internal test methodology and documented failure cases. A frozen human-speech babble diagnostic with retained false alarms.

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

Evidence scope: Scores are uncalibrated. Internal convenience samples include substantial In-the-Wild failures; verdicts do not authenticate speakers. Streaming integrations remain proposed designs. The human-only diagnostic estimates no synthetic miss rate or binary accuracy.

Next milestone: Audit labels and licensing, freeze splits and thresholds, then compare against learned and simple acoustic baselines.

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; a controlled coding-task comparison is the next task-outcome study. [Architecture](#) and [measurement record](#).

Bootscri. 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.** Complete baseline feasibility and the task manifest for the public inference-time proposal, then run a controlled pilot with execution-based hidden checks. [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.

These are planned reports and experiments. 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.