CTC AI Operations — Research Program and Roadmap

Trustworthy, reproducible, low-cost evaluation of frontier and multi-agent AI on sovereign commodity hardware

● Program overview – research agenda and phased roadmap

Abstract

This document frames five individual project papers as one coherent research program and sets a phased roadmap for validating, publishing, and funding it. The unifying question is whether frontier and multi-agent AI systems can be evaluated trustworthily, reproducibly, and cheaply on sovereign commodity hardware — a single 32 GB GPU rather than a datacentre. Four invariants run through every workstream: the system under test is never compressed; references (rubrics, benchmarks) are frozen and content-hashed; untrusted model code executes only in a hardened sandbox; and everything must fit and be measured within one card. The program is organised as five workstreams, each stated as a falsifiable hypothesis with a defined metric, and sequenced into four phases aligned with preprint publication (ORCID / SSRN) and non-dilutive research funding.

§01

Research thesis

Evaluation is the bottleneck of trustworthy AI: capability advances faster than our ability to measure whether a system is correct, honest, and safe. The dominant assumption is that credible evaluation requires frontier-scale cloud infrastructure. This program tests the opposite thesis — that a disciplined, hypothesis-driven protocol on commodity hardware can produce evaluation evidence that is reproducible (frozen references, content hashes), grounded (deterministic execution facts, not opinion), and economical (local low-precision judging anchored by sampled cloud arbitration) — and that the same discipline extends from single-model correctness to multi-agent, fleet-level safety.

§02

The five workstreams

Each project is a workstream with a core hypothesis and a primary metric. Papers marked measured have established infrastructure results; papers marked agenda define an evaluation to be run.

#	Workstream	Core hypothesis (abbreviated)	Primary metric	Stage
W1	Hybrid Evaluation Pipeline	Execution-grounded, frozen-rubric, local FP4 judging is trustworthy and ~10× cheaper	Cohen's □; cost / 1k	Infra measured
W2	Contamination-Resistant Code Eval	Benchmarks regenerated from live repositories resist memorisation	Contamination gap (fresh – stale)	Pilot

#	Workstream	Core hypothesis (abbreviated)	Primary metric	Stage
W3	Three-Tier Workstation	Single-residency time-multiplexing serves three tiers in 32 GB without collision	Peak VRAM; swap cost	Infra measured
W4	Multi-Agent Safety Evaluation	Single-model safety evaluation misses fleet-level emergent failure modes	Emergent-risk detection rate	Agenda
W5	Sovereign Assistant	A local-plus-remote assistant preserves data sovereignty at usable latency	Latency; egress = 0	Design

W1–W3 share the same substrate (32 GB budget, hardened sandbox, frozen references); W4 reuses the W1 sandbox for its testbed; W5 demonstrates the sovereignty case that motivates local evaluation in the first place.

§03

Program-level hypotheses

Beyond the per-project hypotheses, three cross-cutting claims are what the program as a whole tests:

- **P-H1 — Sovereign sufficiency.** A single 32 GB GPU is sufficient for trustworthy dataset-scale evaluation of frontier agentic systems, at judgment quality non-inferior to an all-cloud baseline within a pre-set margin.
- **P-H2 — Reproducibility by construction.** Freezing references under content hashes yields verdicts reproducible across time and software updates, at test–retest agreement above a pre-registered threshold.
- **P-H3 — Emergence gap.** Safety properties that hold for isolated agents do not compose: a measurable fraction of failure modes appear only under multi-agent interaction and are invisible to single-model evaluation.

§04

Roadmap and milestones

Phase	Window	Focus	Publication / funding milestone
P0	done	Pilot infrastructure for W1–W3; hardened sandbox	Working papers v1 (this set)
P1	near term	Labelled-set validation of W1 (□ study); W2 contamination-gap measurement	SSRN preprints v2 with agreement statistics; ORCID linkage
P2	mid term	W4 taxonomy + instrumented testbed; W2 multi-repository generalisation	Multi-agent safety research-call submission; preprint

Phase	Window	Focus	Publication / funding milestone
P3	later	Released protocols (rubric-gate spec, reproducibility packages); W5 deployment study	Open tooling release; follow-on funding

The critical path is P1: it converts infrastructure papers into empirical studies with human-labelled evidence, which is what both peer venues and funders require. P2 is the funding-defining phase, aligning W4 with multi-agent-safety and trustworthy-evaluation programmes.

§05

Publication and dissemination plan

Working papers are versioned and posted as preprints on SSRN, linked to a single ORCID identifier for authorship continuity, using `research@arenskrieger.dev` as the permanent corresponding-author address and the University of Pittsburgh affiliation for the SSRN record. Each paper follows the same template: abstract, motivation, related work, explicit falsifiable hypotheses, methodology and metrics, results (clearly separating measured from predicted), roadmap, limitations and threats to validity, and references. Version 1 establishes the design and measurement plan; version 2 adds empirical agreement statistics from P1. This staged approach is deliberate — it lets the design be cited and time-stamped now while the empirical study runs, and it keeps every claim honest about whether it is measured or predicted.

§06

Funding strategy

The program is designed for non-dilutive research funding rather than equity dilution. Its natural fit is with programmes on AI safety, trustworthy evaluation, and multi-agent risk: W1 and W2 speak to trustworthy, low-cost evaluation; W4 speaks directly to multi-agent safety. The near-term target is a multi-agent-safety research call (to be confirmed against current deadlines); the preprint set plus the P1 empirical results form the evidence base for the application. A key differentiator for funders is the sovereign-hardware angle — credible evaluation that does not depend on frontier-scale compute lowers the barrier for independent, reproducible safety research.

§07

Risks and mitigations

The principal scientific risk is correlated error between the local judge and the cloud arbiter; it is mitigated by anchoring on independent human labels rather than treating the arbiter as ground truth. The principal external-validity risk is single-operator, single-hardware results; it is mitigated by the P2 generalisation work across repositories and task distributions. The principal programme risk is scope: five workstreams is ambitious for one operator, which is precisely why the roadmap sequences them and why funding is sought to resource the empirical phases. The multi-agent emergence claim (P-H3) is the highest-variance and highest-value bet; it is deliberately staged behind the more tractable W1–W2 validation so that the program accrues credibility before its most speculative claim.

References

[1] Zheng, L., Chiang, W.-L., Sheng, Y., et al. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. NeurIPS 2023 Datasets and Benchmarks. arXiv:2306.05685. [2] Jain, N., Han, K., Gu, A., et al. (2024). LiveCodeBench: Holistic and Contamination-Free Evaluation of Large Language Models for Code. arXiv:2403.07974. [3] Liang, P., Bommasani, R., Lee, T., et al. (2022). Holistic Evaluation of Language Models (HELM). arXiv:2211.09110. [4] Kwon, W., Li, Z., Zhuang, S., et al. (2023). Efficient Memory Management for Large Language Model Serving with PagedAttention. SOSP 2023. arXiv:2309.06180. [5] Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training language models to follow instructions with human feedback. NeurIPS 2022. arXiv:2203.02155.