Zhaohui Wang

ML Systems Researcher & Engineer

zhaohui.geoffrey.wang@gmail.com Los Angeles, CA LinkedIn

GitHub Portfolio

// Professional Summary

ML systems researcher and engineer working on **verifiable and efficient LLM infrastructure** — from CUDA/Triton sparse-attention kernels and GPU cryptography up to agent security, routing, and edge feasibility.
First author on **six peer-reviewed 2026 main-track conference papers** (ICML, ECML PKDD, CIKM, ICICS, NLPCC ×2) plus ten workshop papers at NeurIPS, ICML, ICLR, AAMAS, ProbML and C3NLP, with production backend and DevOps experience across AWS, Docker, and Kubernetes. Currently a Research Assistant at the CMU Infini-AI Lab.

// Core Competencies

Technical Skills

GPU & Inference — CUDA, Triton kernels, GPU MSM, CUDA Graph, sparse attention, KV cache, speculative decoding, quantization (INT4/INT8), memory pooling; PyTorch, SGLang, vLLM, FlashInfer, TensorRT-LLM, ONNX Runtime
ML & Agents — Transformers, DeepSpeed, LoRA/QLoRA/PEFT, scikit-learn, XGBoost, LangChain, multi-agent orchestration, large-scale red-teaming, W&B/MLflow
Methods — Sequential decision theory, optimal stopping, bandits, mechanism design, uncertainty quantification, statistical testing (ANOVA/FDR), zero-knowledge proofs, Merkle trees, sumcheck
Security & DevOps — OWASP Top-10, SAST (bandit, semgrep, gosec, Clippy, Cppcheck), pip-audit/trivy, secrets management, JWT / RBAC / OAuth2 (AWS Cognito); Docker, Kubernetes (HPA, health probes), GitHub Actions CI/CD, Infrastructure-as-Code, Prometheus, Grafana, OpenTelemetry
Cloud & Backend — AWS (EC2, S3, RDS, IAM, MediaConvert, CloudFront), Nginx, Spring Boot, Node.js/Express, FastAPI, Django/Wagtail, Kafka, Redis, PostgreSQL, MongoDB

Languages — Python, C/C++, CUDA, Rust (published crates `hanfei-shu`, `hanfei-fa`), Java, TypeScript/JavaScript, Go, Bash, SQL

Core Strengths

- Independent research direction-setting — six first-author main-track papers in a single year, from problem formulation through camera-ready
- Written communication for mixed audiences: peer-reviewed papers, public technical notes, and internal engineering documentation
- Cross-boundary systems work — comfortable moving between CUDA kernels, distributed backends, and production operations in the same project
- Reproducibility discipline: benchmark harnesses, idempotent deployment commands, and honest reporting of negative results

// Professional Experience

CMU Infini-AI Lab — Research Assistant
PI: Prof. Beidi Chen · Pittsburgh, PA (Remote) · Mar 2026 – Present

- Core contributor to **Vortex**, a sparse-attention LLM inference framework on **SGLang/FlashInfer**; designed custom sparse-attention **CUDA** kernels with **CUDA Graph** capture for **2.7× end-to-end speedup**.
- Built page-level cross-layer sparse-index sharing and a masked top-k kernel to cut redundant attention memory traffic across **1B–70B** models.
- Built large-scale reproducible **evaluation and benchmarking harnesses** (multi-model sweeps, 30K+ runs, Docker/AWS, W&B tracking) — the backbone of the ICML 2026 red-team study.

Meetfood — Software Engineer Intern
Los Angeles, CA · May 2025 – Aug 2025

- Designed and shipped RESTful APIs with **Node.js/Express/MongoDB** — **JWT** auth, **role-based access control**, **AWS Cognito**, media uploads, and cross-device sync endpoints.
- Cut p95 latency of hot search/feed endpoints by **~40%** (≈500 ms → 300 ms) via response caching and MongoDB query and index optimization.
- Built the video pipeline on **AWS**: S3 uploads, **MediaConvert** transcoding, **CloudFront** CDN delivery — cutting average video start time by **~50%**; deployed on EC2 with auto-scaling and RDS under least-privilege IAM.
- Set up **GitHub Actions CI/CD**, cutting release turnaround from ~1 hour of manual steps to **<10 minutes**.

Tencent — Java Backend Engineer Intern
Shenzhen, China · Jul 2019 – Aug 2019

Built backend modules for an internal messaging platform in **Spring Boot** with **MyBatis**; **Redis** caching of hot session data cut hot-path DB reads by **~60%**.

Refactored synchronous service calls into async **CompletableFuture** pipelines, reducing average response time by **~30%** (≈200 ms → 140 ms).

Deployed the team's microservices to **Kubernetes** (TKE) via declarative manifests with liveness/readiness probes and Prometheus/Grafana monitoring, cutting manual deploy time from ~30 min to **<5 min**.

// Key Projects & Achievements

Interruptr — Multi-Agent Vulnerability Scanner
Python, LangGraph, AutoGen · 2025–2026

Multi-agent static-analysis pipeline deployable as a **CI/CD security gate**, scanning C/C++ for memory-safety bugs, race conditions and other CWE-classified vulnerabilities: three cloud experts plus one *free* local adversarial verifier.

77.2% F1 at **$0.002/sample** on NIST Juliet with 100% recall; the local verifier alone adds **+10.3 pp precision** ($p<10^{-6}$, McNemar) and parallel execution gives a 3.0× speedup.

Timeline-Feed — Hardened High-Traffic Backend
TypeScript/Node.js, Kafka, Docker · 2025

Production-grade feed backend with **OWASP Top-10** hardening (XSS prevention, HMAC-signed cursors) and graceful-degradation circuit breakers.

Sub-15 ms P95 at **99.99%** availability; 30+ Prometheus metrics, Grafana, OpenTelemetry, and 80% test coverage across a 9-service Docker deployment.

geoffreywtech.me — Self-Hosted Research Portfolio
Django 5.2, Wagtail, Docker, Nginx, AWS EC2 · 2025–2026

Designed, deployed and operate this bilingual Wagtail CMS behind Nginx TLS termination on EC2, with content shipped as idempotent Django management commands rather than hand edits.

Ran a full **DevSecOps audit** of the live stack (bandit/semgrep/pip-audit/trivy plus runtime probing): fixed a production `DEBUG=True` exposure and a **CWE-601** open redirect, patched **132 CVEs** over two audit rounds, and reached zero Django `check --deploy` warnings.

// Education

University of Southern California — M.S. in Computer Science (STEM)
Los Angeles, CA · Jun 2024 – Dec 2025
University of Chinese Academy of Sciences — Ph.D. Candidate, Computer Science (discontinued)
Beijing, China · Aug 2020 – Mar 2023

Central South University — B.S. in Computer Science
Changsha, China · Sep 2016 – Jun 2020

// Publications & Open Source

First author on all. Full list with PDFs, figures and posters: geoffreywtech.me/publications

Main track

- **Persistent Semantic Entities in Tool-Augmented LLM Systems** — **ICML 2026**. Formalized a new class of LLM-agent vulnerability: implicit state in tool registries and event systems that persists across sessions and propagates across agents, invisible to standard prompt-injection monitoring. 34,000+ trials over 20 models from 9 vendor families (1.5B–1T); all susceptible (median 70% contamination), uncorrelated with scale (r^2=0.06); name-binding dominant (Cohen's d=3.26); cross-model validation defense eliminates the vector (median 100% reduction).
- **Uncertainty-Aware Sequential Decision Rules for Event-Triggered LLM Invocation in Streaming Systems** — **ECML PKDD 2026**. Cast *when* to invoke an expensive cloud LLM as risk-based sequential stopping; unified framework subsuming event-triggered / optimal-stopping / SPRT / CUSUM / bandit triggers (6 theorems, incl. $O(\sqrt{T \log T})$ regret). <5% miss at 6.2% invocation (~10× fewer calls); 63% lower total cost than CUSUM.
- **NanoZK: Layerwise Zero-Knowledge Proofs for Verifiable LLM Inference** — **ICICS 2026**. Made outsourced LLM inference cryptographically verifiable via layerwise ZK proofs (Halo2/IPA) with a custom GPU multi-scalar-multiplication CUDA kernel, released as the Rust crate `hanfei-shu`. 7.6× MSM speedup; 620 ms prove / 2.4 KB proof / 2 ms verify on LLaMA-3-1B at zero accuracy loss.
- **The Recall Ceiling: Why LLM Recommendation Reranking Fails in Practice** — **CIKM 2026**. Showed the oracle evaluation protocol used for LLM rerankers overstates NDCG@10 by 92–95%, and proved a deterministic ceiling: expected NDCG@k cannot exceed recall at the reranker's own candidate-window size under leave-one-out evaluation. Realistic retrieval covers only 2–19% of relevant items at K=100 across eight datasets in three domains.
- **IntegriRef: A Five-Layer Bayesian Framework for Cross-Domain Reference Integrity Verification** — **NLPCC 2026**. Five-layer pipeline over 62 registry adapters fusing 21 typed signals with Bayesian likelihood-ratio scoring: 99.9% recall on 1,215 hallucinated references and 98.3% on 5,488 retracted papers; L2 semantic verification reaches 93.5% precision against 89.1% human agreement.
- **When Retrieval Systems Encode Wrong Entity-Context Bindings** — **NLPCC 2026**. Index-level adversarial memorization: false entity–context associations need not be learned in pretraining, yet are preserved and amplified by the retrieval index. Multilingual case study over 1,535 articles; an evidence-grounded rewriting prototype removes 83.3% of pollution where the best filter reaches 20.8%.

Workshop

When Individually Calibrated Models Become Collectively Miscalibrated — ProbML 2026 (PMLR archival). Collective miscalibration under strategic prediction aggregation for edge–cloud inference, with an incentive-compatible (VCG/Brier) mechanism. 0/6 exploits; false-negative rate 0.027 vs 0.045–0.067 for neural baselines; 4× more data-efficient; sub-2 ms at 200 agents.

Interruptr — AAMAS 2026 SE Workshop · arXiv:2604.21282

AgentTrace — AIWILD @ ICLR 2026 · arXiv:2603.14688

Universe Routing — LLA @ ICLR 2026 · arXiv:2603.14799

NanoZK (workshop version) — VerifAI @ ICLR 2026

Block-Level Weight-Space Structure Persists Under Post-Training — Weight-Space Symmetries (WSS) @ ICML 2026. Post-training rewrites every tensor yet preserves block-level geometry (cosine > 0.99), enabling 18–48% GPU memory savings across co-resident model variants.

Learning from World Feedback — RLxF @ ICML 2026. Dynamics-based uncertainty penalties *increase* collisions from 26% to 34%; replacing the model-internal proxy with world-feedback signals cuts them to 1–14% without retraining.

From Risk Scoring to Risk Allocation — NExT-Game @ ICML 2026 (non-archival). Monitoring in multi-agent systems recast from per-state scoring to combinatorial allocation: a QUBO whose Pareto frontier contains greedy top-K, MMR and k-DPP as fixed operating points. +24% to +66% diversity over greedy from n=15 to n=200; a learned λ policy reaches 99.5% of an oracle grid search. OpenReview

TabletCraft — C3NLP @ ACL 2026 · bidirectional Akkadian NMT and cuneiform rendering; 49.1 / 48.5 BLEU, the first published quantitative result in the reverse direction. pip install cuneiscribe

When Rank Rises as LLMs Degrade — CL4FMAgents @ NeurIPS 2026 (non-archival). The directional assumption behind spectral health monitors fails under LLM post-training: a data-duplication regime that degrades held-out loss by 75% drives RankMe *above* healthy (13.5 pooled s.d.) and covariance effective rank to nearly twice healthy, so a one-sided monitor scores the worst checkpoint as the healthiest. A two-sided multi-channel monitor detects all three damage regimes within 10–60 steps of the fork, but leads held-out probe loss in only one of nine folds — reported as a negative result. arXiv:2610.09647

// Awards & Recognition

Six first-author main-track acceptances in 2026 — ICML, ECML PKDD, CIKM, ICICS and NLPCC ×2 — spanning agent security, decision-theoretic serving, recommendation evaluation, verifiable inference, and citation integrity.

Published open-source Rust crates `hanfei-shu` and `hanfei-fa` for GPU-accelerated zero-knowledge proving.

Built and operate a self-funded edge–fog–cloud GPU testbed used for every 2026 paper: 2× Jetson Orin Nano (8 GB) + RTX 3080 Ti (16 GB) + workstation GPU server.

// Additional Information

Open to relocation; availability immediate / negotiable. Reach me at zhaohui.geoffrey.wang@gmail.com — a PDF resume is linked at the top of this page.