王朝晖

机器学习系统研究员 / 工程师

zhaohui.geoffrey.wang@gmail.com 美国 洛杉矶 LinkedIn GitHub

// 个人简介

机器学习系统方向的研究员与工程师，专注于**可验证、高效的大模型基础设施**——从 CUDA/Triton 稀疏注意力内核与 GPU 密码学，到智能体安全、请求路由与边缘部署可行性。
2026 年以第一作者身份发表 **6 篇主会论文**（ICML、ECML PKDD、CIKM、ICICS、NLPCC ×2），另有 10 篇 NeurIPS / ICML / ICLR / AAMAS / ProbML / C3NLP workshop 论文；同时具备 AWS、Docker、Kubernetes 上的生产级后端与运维经验。现任 CMU Infini-AI Lab 研究助理。

// 技术技能

GPU 与推理 — CUDA、Triton 内核、GPU MSM、CUDA Graph、稀疏注意力、KV Cache、投机解码、量化（INT4/INT8）、显存池化；PyTorch、SGLang、vLLM、FlashInfer、TensorRT-LLM、ONNX Runtime
机器学习与智能体 — Transformers、DeepSpeed、LoRA/QLoRA/PEFT、scikit-learn、XGBoost、LangChain、多智能体编排、大规模红队测试、W&B/MLflow
理论方法 — 序贯决策理论、最优停止、多臂老虎机、机制设计、不确定性量化、统计检验（ANOVA/FDR）、零知识证明、Merkle 树、Sumcheck
安全与 DevOps — OWASP Top-10、SAST（bandit、semgrep、gosec、Clippy、Cppcheck）、pip-audit/trivy、密钥管理、JWT / RBAC / OAuth2（AWS Cognito）；Docker、Kubernetes、GitHub Actions CI/CD、基础设施即代码、Prometheus、Grafana、OpenTelemetry
云与后端 — AWS（EC2、S3、RDS、IAM、MediaConvert、CloudFront）、Nginx、Spring Boot、Node.js/Express、FastAPI、Django/Wagtail、Kafka、Redis、PostgreSQL、MongoDB
编程语言 — Python、C/C++、CUDA、Rust（已发布 crate `hanfei-shu`、`hanfei-fa`）、Java、TypeScript/JavaScript、Go、Bash、SQL

// 核心能力

独立确定研究方向——一年内完成 6 篇第一作者主会论文，从问题定义到 camera-ready 全流程

面向不同受众的书面表达：同行评议论文、公开技术笔记与内部工程文档

跨层次系统工作能力——能在同一项目中于 CUDA 内核、分布式后端与生产运维之间自如切换

可复现性纪律：基准测试框架、幂等部署命令，以及对负面结果的如实报告

// 工作经验

CMU Infini-AI Lab — 研究助理
导师：Beidi Chen 教授 · 美国匹兹堡（远程） · 2026 年 3 月 – 至今

稀疏注意力大模型推理框架 **Vortex** 的核心贡献者（基于 **SGLang/FlashInfer**）；设计定制稀疏注意力 **CUDA** 内核并配合 **CUDA Graph** 捕获，实现端到端 **2.7×** 加速。

实现页级跨层稀疏索引共享与 masked top-k 内核，在 **1B–70B** 模型上削减冗余注意力访存。

构建大规模可复现的**评测与基准测试框架**（多模型扫描、3 万余次运行、Docker/AWS、W&B 追踪），支撑 ICML 2026 红队研究。

Meetfood — 软件工程实习生
美国洛杉矶 · 2025 年 5 月 – 2025 年 8 月

使用 **Node.js/Express/MongoDB** 设计并交付 RESTful API——**JWT** 鉴权、**基于角色的访问控制**、**AWS Cognito**、媒体上传与跨设备同步接口。

通过响应缓存与 MongoDB 查询/索引优化，将核心搜索与信息流接口的 p95 延迟降低 **约 40%**（约 500 ms → 300 ms）。

在 **AWS** 上搭建视频流水线：S3 上传、**MediaConvert** 转码、**CloudFront** CDN 分发，平均起播时间降低 **约 50%**；服务部署于 EC2（自动扩缩）与 RDS，遵循最小权限 IAM。

搭建 **GitHub Actions CI/CD**，将发布周期从约 1 小时的人工操作压缩到 **10 分钟以内**。

腾讯 — Java 后端开发实习生
中国深圳 · 2019 年 7 月 – 2019 年 8 月

基于 **Spring Boot** 与 **MyBatis** 开发内部消息平台后端模块；**Redis** 热点会话缓存使热路径数据库读取减少 **约 60%**。

将同步服务调用重构为异步 **CompletableFuture** 流水线，平均响应时间降低 **约 30%**（约 200 ms → 140 ms）。

通过声明式清单（**基础设施即代码**）将团队微服务部署至 **Kubernetes**（TKE），配置存活/就绪探针与 Prometheus/Grafana 监控，人工部署耗时从约 30 分钟降至 **5 分钟以内**。

// 项目经验

Interruptr — 多智能体漏洞扫描器
Python、LangGraph、AutoGen · 2025–2026

可作为 **CI/CD 安全门禁**部署的多智能体静态分析流水线，检测 C/C++ 内存安全缺陷、竞态条件等 CWE 分类漏洞：三个云端专家 + 一个*零边际成本*的本地对抗验证器。

在 NIST Juliet 上达到 **77.2% F1**、**$0.002/样本**、100% 召回；仅本地验证器一项即带来 **+10.3 个百分点**的精确率提升（McNemar 检验 $p<10^{-6}$）。

Timeline-Feed — 高流量加固后端
TypeScript/Node.js、Kafka、Docker · 2025

生产级信息流后端，完成 **OWASP Top-10** 加固（XSS 防护、HMAC 签名游标）与优雅降级熔断。

P95 低于 15 ms、可用性 **99.99%**；9 个服务的 Docker 部署中包含 30+ Prometheus 指标、Grafana、OpenTelemetry 与 80% 测试覆盖率。

geoffreywtech.me — 自建研究作品集网站
Django 5.2、Wagtail、Docker、Nginx、AWS EC2 · 2025–2026

独立设计、部署并运维本双语 Wagtail CMS（Nginx TLS 终止 + EC2），内容通过幂等的 Django 管理命令发布，而非手工编辑。

对线上环境执行完整 **DevSecOps 审计**（bandit/semgrep/pip-audit/trivy 结合运行时探测）：修复生产环境 `DEBUG=True` 暴露与 **CWE-601** 开放重定向，两轮审计共修补 **132 个 CVE**，最终 Django `check --deploy` 零告警、`pip-audit` 无已知漏洞。

// 教育背景

南加州大学（USC） — 计算机科学硕士（STEM）
美国洛杉矶 · 2024 年 6 月 – 2025 年 12 月
中国科学院大学（UCAS） — 计算机科学博士研究生（未完成）
中国北京 · 2020 年 8 月 – 2023 年 3 月
中南大学 — 计算机科学与技术学士
中国长沙 · 2016 年 9 月 – 2020 年 6 月

// 论文发表

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

// 获奖荣誉

2026 年获得 6 篇第一作者主会录用——ICML、ECML PKDD、CIKM、ICICS 与 NLPCC ×2，覆盖智能体安全、决策论驱动的服务调度、推荐系统评测、可验证推理与引用完整性。

发布开源 Rust crate `hanfei-shu` 与 `hanfei-fa`，用于 GPU 加速的零知识证明。

自建并运维边缘–雾–云 GPU 测试平台，支撑全部 2026 年论文实验：2× Jetson Orin Nano（8 GB）+ RTX 3080 Ti（16 GB）+ 工作站 GPU 服务器。