Sijia Wang

Machine Learning Engineer — Production AI & Agentic Systems

scarlett.95.wang@gmail.com — github.com/swang1024 — Google Scholar — LinkedIn
Graduation: May 2026 (PhD, Duke University) — Available: June 2026 — Work-authorized (OPT/STEM-OPT)

Professional Summary

Machine learning engineer who ships production AI systems used by real users. Built and A/B-tested **retrieval and ranking** systems at Pinterest (GraphSAGE candidate retrieval, Faiss ANN, multitask BERT) with measurable revenue and CTR impact, and designed **SAGE**, an agent-memory system that grounds LLM agents with the right context under tight cost and latency budgets while beating Mem0 on the LoCoMo benchmark. Comfortable across the full loop of building with models — training, evaluation, serving, and deployment — and the eval and feedback machinery that catches regressions before they reach users. Experience in domains where wrong answers carry real consequences (credit-risk lending, model reliability and hallucination detection). PhD, Duke ECE (2026).

Technical Skills

Languages	Python, SQL, C++, Bash (also Matlab)
Agent & LLM Systems	Agent memory & grounding, RAG / retrieval pipelines, eval & benchmarking harnesses (LoCoMo), tool / function calling, hallucination & confidence scoring, prompt & context management, cost & latency optimization, LoRA, quantization, distillation, vLLM, SGLang
ML Foundations	Representation learning, LLMs, multimodal (VLM), retrieval / ranking, continual & federated learning, domain adaptation, interpretable ML
Infra & Tooling	Docker, AWS, Spark, Git, Unix/Linux, PyTorch, Hugging Face, Slurm, online A/B testing, Spinner workflow (Pinterest)
AI-assisted Dev	Claude Code, Codex, ChatGPT, Copilot

Industry Experience

Pinterest Labs — ML Engineer / Research Intern, Ads Retrieval & Targeting **Summer 2022, Summer 2023**

- Designed and shipped an **advertiser-similarity retrieval pipeline** (**GraphSAGE** embeddings + **Faiss ANN** search) into Pinterest's production **Spinner** auto-targeting system, enriching ad candidate supply; ran the end-to-end online experiment and measured **∼1% revenue uplift in A/B testing**.
- Built the offline KNN workflows and wrote tested, reviewed Python/Spark code that passed internal production review; owned features end-to-end — data ingestion, embedding indexing, candidate scoring, online serving, and offline/online evaluation.
- Earlier, developed a **multitask BERT broad-match model** to improve ad–query relevance and retrieval coverage; deployed to production with measurable **CTR improvement**, partnering across Ads Retrieval, Ranking, and Infra teams on experiment design and launch.

Agent Systems & Applied Research

SAGE — Agent Memory & Grounding System (lead author), DUKE — ADVISOR: PROF. RICARDO HENAO **2025 — 2026**

- Built **SAGE**, a persistent memory layer that decides what an LLM agent can recall and when to write or update memory — grounding responses with the right context at the right time under explicit **cost and latency budgets**, the core problem in serving agents over real workloads.
- Designed the **novelty-gated update policy** as a sequential decision problem, cutting redundant memory writes and serving cost while preserving answer quality.
- Built the evaluation harness benchmarking SAGE against **Mem0** on the **LoCoMo** long-conversation benchmark — **7/7 token-F1 wins** with lower cost and latency — packaged as a single-command reproducible benchmark; as a drop-in gate it skips ∼16–18% of LLM calls with minimal quality change. Serving on ollama/vLLM/SGLang (**arXiv preprint + open-source repo** github.com/swang1024/SAGE).

Model Reliability & Failure Handling — VLM Hallucination Detection, DUKE — ADVISOR: PROF. RICARDO HENAO (TMLR 2026) **2025**

- Built a **cross-modal consistency check** comparing a model's visual vs. textual reasoning paths to flag low-confidence and "unknown" outputs — deciding when a model should **abstain, escalate for human review, or self-correct** rather than answer wrongly.
- Built an end-to-end multimodal evaluation pipeline benchmarking **GPT-4V, Qwen-VL, and LLaMA-VL**; quantified epistemic uncertainty and characterized failure modes directly relevant to agent trust boundaries and mission-critical workflows.

Earlier Research Experience

Samsung Semiconductor, SoC R&D Lab — Research Fellow / ML Research Intern **2019 — 2022**

- **Led research on continual and federated learning**, producing **3 patents** and **2 publications**.
- **Communication-efficient federated learning**: compressed the global model via quantization to cut downlink cost while preserving accuracy; server-side refinement via quantization-aware training without access to client data — efficient model serving under real constraints.
- **Sustainable continual learning**: task-similarity detection and encoder reuse to bound memory growth — the same scaling problem as bounded memory in long-running agents.

Selected High-Stakes-Domain Project

FICO Explainable Machine Learning Challenge — Credit-Risk Decisioning, ADVISOR: PROF. CYNTHIA RUDIN **2018**

- Built **interpretable credit-risk models** on FICO's lending dataset — a domain where wrong answers carry direct financial and regulatory consequences — combining a traditional scoring system with a neural model with no accuracy loss for full interpretability.
- Led development of an interactive web interface for case-level decision auditing; work accepted at **NeurIPS 2018** (workshop) and published in **Decision Support Systems (2022)**; FICO Recognition Award.

Education

PhD, Electrical and Computer Engineering, Duke University **Aug 2019 — May 2026**
M.S., Electrical and Computer Engineering, Duke University **Aug 2017 — May 2019**
B.E., Telecommunication Engineering, Communication University of China **Sep 2013 — Jun 2017**

Publications (Selected)

- *SAGE: A Novelty Gate for Efficient Memory Evolution in Agentic LLMs*, arXiv preprint 2026 (under review, ACL ARR). Code: github.com/swang1024/SAGE.
- *Fallback-Enabled Closed-Set Classification: Cross-Modal Consistency in Vision-Language Models*, **TMLR 2026**
- *GAN Memory with No Forgetting*, **NeurIPS 2020**
- *Model Recycling Framework for Multi-source Data-free Supervised Transfer Learning*, **IEEE MLSP 2025 (Oral)**
- *Toward Sustainable Continual Learning: Detection and Knowledge Repurposing of Similar Tasks*, **IEEE MLSP 2025**
- *A Holistic Approach to Interpretability in Financial Lending*, **Decision Support Systems Journal 2022**

Patents (Filed)

- Sustainable continual learning with detection and knowledge repurposing of similar tasks — 2023 (App. No. 18/099,631)
- Method and apparatus for communication-efficient federated learning with global model compression — 2023 (App. No. 17/824,558)
- Method and apparatus for continual few-shot learning without forgetting — 2022 (App. No. 17/156,126)

Awards & Service

- AISTATS Top 10% reviewer – 2022
- SOC R&D lab of Samsung Semiconductor Fellowship – 2019-2022
- The Pratt Peers Student Support Network – 2025 - 2026
- FICO Recognition Award for FICO Explainable Machine Learning Challenge – 2019
- Duke Academic Scholarship – 2017-2018
- Outstanding Student Scholarship in Communication University of China – 2013-2016
- Teaching Assistant – Intro to Deep Learning (PhD), Modern Analytics (MBA) – Duke