Sijia Wang

Research Scientist — Machine Learning

scarlett.95.wang@gmail.com — github.com/swang1024 — Google Scholar — LinkedIn
Graduation: May 2026 (PhD, Duke University) — Available start: June 2026 onward (flexible through end of 2026)

Professional Summary

Duke ECE PhD (May 2026) with a **multi-venue publication record** (**NeurIPS 2020, TMLR 2026, IEEE MLSP 2025 Oral**) and **3 patents**. Research spans continual / federated learning, transfer learning and domain adaptation, multimodal reliability (cross-modal hallucination detection), and **agent memory** (SAGE: public arXiv + reproducible code). Translates research into shipped systems — two Pinterest Labs production deployments with measured business impact. Comfortable owning a problem from formulation and evaluation design through to a runnable, reproducible artifact.

Technical Skills

ML Foundations	Representation learning, generative models, continual / federated learning, transfer learning & domain adaptation, interpretable ML, large-scale recommendation
LLM & VLM	Prompt engineering, in-context learning, quantization, LoRA, distillation, self-supervised learning, vLLM / SGLang, VLMs (LLaMA, Qwen-VL, GPT-class)
Reliability & Eval	Hallucination detection, selective prediction / unknown rejection, calibration & epistemic uncertainty, cross-modal consistency, evaluation under distribution shift, A/B testing, failure-mode analysis
Agent & Memory	Persistent memory for LLM agents, memory lifecycle (write/update/compress/forget), novelty gating, retrieval-grounded generation, evaluation frameworks, tool/function calling, context management
Retrieval & Ranking	Large-scale embedding retrieval, two-tower / GraphSAGE embeddings, ANN search (Faiss IVF/HNSW), KNN candidate generation, ranking models, RAG pipelines, online serving
Programming	Python, C++, SQL, Bash; PyTorch, Hugging Face Transformers, Spark, AWS, Git, Linux

Research Experience

SAGE — Memory Management for Agentic LLMs, DUKE — ADVISOR: PROF. RICARDO HENAO **2025 — Present**

- Designed **SAGE**, a cost-efficient **novelty-gated memory layer** for LLM agents that governs the write/update/compress/forget lifecycle, formulating memory updates as a policy decision driven by a **von Mises–Fisher novelty gate**.
- On the **LoCoMo** long-conversation benchmark, SAGE beats **Mem0** on average token-F1 across open-weight backbones while reducing add-phase **API cost ∼3.4×** and **latency ∼2.5×**; as a drop-in gate for A-Mem it skips ∼16–18% of LLM calls with minimal quality loss.
- Released as a public **arXiv preprint with single-command reproducible code** (github.com/swang1024/SAGE); under review at ACL ARR 2026.

Cross-Modal Consistency & Hallucination Detection in VLMs, DUKE — ADVISOR: PROF. RICARDO HENAO **2025**

- Designed a **cross-modal consistency framework** that flags hallucinations and "unknown" predictions in vision-language models by comparing visual- vs. text-grounded reasoning paths, enabling **selective prediction / unknown rejection** with calibrated trust signals. **(TMLR 2026)**
- Built an end-to-end multimodal evaluation pipeline benchmarking **GPT-4V, Qwen-VL, and LLaMA-VL**; quantified epistemic uncertainty and characterized failure modes relevant to trust-and-safety and content-moderation workflows.

Multi-Source Data-Free Transfer Learning, DUKE — ADVISOR: PROF. RICARDO HENAO **2022 — 2024**

- Developed a **multi-source model-recycling framework** that selects and reuses pretrained models under white-box and black-box settings, enabling Model-as-a-Service customization with limited data access. **(IEEE MLSP 2025, Oral)**

Toward Sustainable Continual Learning, DUKE — ADVISOR: PROF. RICARDO HENAO, PROF. LAWRENCE CARIN **2020 — 2022**

- Built **task-similarity detection and parameter-reuse** to address superlinear model growth in continual learning, reducing memory footprint while maintaining accuracy across sequential tasks. **(IEEE MLSP 2025)**

Industrial Research Experience

Pinterest Labs — Research Intern, Ads Retrieval & Targeting **May–Aug 2022, May–Aug 2023**

- **2023**: Designed and shipped a graph-based **advertiser-similarity retrieval pipeline** using **GraphSAGE embeddings** and **Faiss ANN search**, enriching ad candidate supply for Pinterest's auto-targeting system. Built offline KNN workflows, integrated into the production **Spinner workflow**, wrote tested Python/Spark code under internal review, and ran end-to-end online experiments achieving **∼1% revenue uplift in A/B testing**.
- **2022**: Developed a **multitask BERT model for broad match** improving ad-query relevance and retrieval coverage; deployed to production with measurable **CTR lift**. Owned the large-scale retrieval feature lifecycle end-to-end — ingestion, embedding indexing, candidate scoring, online serving, and offline/online evaluation — across Ads Retrieval, Ranking, and Infra teams.

Samsung Semiconductor, SoC R&D Lab — Research Fellow / Deep Learning 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-aware training** to cut downlink cost while preserving accuracy; enabled server-side refinement through mix-up of reconstructed feature maps without client data access.
  - **Sustainable continual learning**: built **task-similarity detection and encoder-reuse** to bound memory growth in continual learning systems.
  - **Continual few-shot learning**: incrementally updated models for new tasks via a weight generator derived from prior-task representations.
- **GAN Memory without Forgetting**: introduced a parameter-efficient generative-replay method for continual learning without catastrophic forgetting (NeurIPS 2020).

Selected Collaborative Project

FICO Explainable Machine Learning Challenge, ADVISOR: PROF. CYNTHIA RUDIN **2018**

- Co-developed an **interpretable global model** combining traditional credit scoring with a feed-forward network (no accuracy loss for full interpretability) plus an interactive web interface for case-level explanations. NeurIPS-accepted; FICO Recognition Award.

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: Models, Visualizations, and Summary-Explanations*, **Decision Support Systems 2022**
- *An Interpretable Model with Globally Consistent Explanations for Credit Risk*, **NeurIPS 2018 Workshop** (AI in Finance)

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)

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**

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