Hongcheng Jiang

AI Engineer | GenAI, RAG & Agentic Systems | Multimodal AI
U.S. Permanent Resident | Citizenship Expected Dec. 2026
Open to Remote | Relocation Nationwide | (+1) 626-206-2285
hongcheng.jiang@outlook.com linkedin.com/in/hongchengjiang github.com/jianghongcheng

Professional Summary

AI Engineer with a PhD and 1+ year of industry ML engineering experience, building production-oriented LLM agents, RAG systems, and AI applications with Python, FastAPI, and LangGraph. Combines deep-learning research expertise with backend engineering to build reliable AI systems, improve model accuracy and retrieval quality, and optimize token usage and latency.

Certifications

Google Cloud: Professional Machine Learning Engineer; Professional Cloud Architect
Microsoft Certified: Azure AI Apps and Agents Developer Associate (AI-103)

Technical Skills

Programming: Python, SQL, NumPy, Pandas
LLM Agents: LangGraph, MCP, Structured Outputs, Tool Integration, Human-in-the-Loop
Retrieval / RAG: BM25/Vector Search, Embeddings, Reranking, Chunking, Schema Grounding
ML / Modeling: PyTorch, Hugging Face, Transformers, Diffusion Models, Multimodal Learning
Cloud: Google Cloud (GCP), Microsoft Azure
Backend / Systems: FastAPI, SQLite, PostgreSQL, Async Workers, Docker, Linux, Git, vLLM, Ollama
Evaluation: Retrieval Metrics, SQL Correctness, Ablation Studies, Token Usage, Latency, Recovery Testing

Work Experience

Research Associate **June 2026 – Present**
University of Missouri–Kansas City (R1 Research University) Kansas City, MO

- **Medical AI:** Collaborated with **Children's Mercy Kansas City** to develop GeoVIRA for automated foot and spine angle measurement from radiographs, combining predicted anatomical landmarks with frozen medical image features.
- **Model Engineering:** Designed a relational geometry encoder capturing landmark distances, orientations, and angular relationships; implemented **geometry-conditioned cross-attention** and measurement-aligned visual feature sampling.
- **Data Pipelines:** Prepared radiographic datasets and built training pipelines with patient-disjoint foot splits and **out-of-fold landmark predictions** to avoid in-sample detector outputs during downstream model training.
- **Model Evaluation:** Evaluated geometry-only and geometry–visual models through baseline comparisons, component ablations, and matched-versus-shuffled visual controls to assess the contribution of case-specific image features.

Machine Learning Engineer **June 2025 – May 2026**
NextTier – IT Solutions Consultancy Sacramento, CA (Remote)

- **Ownership:** Owned the end-to-end design and implementation of a natural-language SQL assistant for revenue analysis, order lookup, and grouped reporting, spanning retrieval, agent orchestration, service layer, and evaluation.
- **Agent Architecture:** Designed **LangGraph Planner–Verifier workflows** so generated SQL is checked by an independent verifier rather than returned unvalidated, with clarification turns for ambiguous requests and explicit retry budgets to bound repair cost.
- **Retrieval:** Combined **BM25/vector hybrid retrieval with cross-encoder reranking** to recover both exact schema terms and paraphrased business concepts, with database/table filtering to scope generation context.
- **Execution Controls:** Treated the database as an untrusted surface—read-only execution by default, table-level permissions, and human approval gating for any write, with transactional idempotency so retried requests cannot double-apply.
- **Reliability:** Built browser and **FastAPI/MCP interfaces** backed by durable asynchronous workers; persistent job state, renewable leases, and retries let in-flight requests survive worker restarts instead of failing silently.

Graduate Research Assistant **Jan 2020 – May 2025**
University of Missouri–Kansas City Kansas City, MO

- **Vision & Generative Modeling:** Developed **transformer- and diffusion-based models** for thermal image super-resolution, hyperspectral pansharpening and restoration, and NIR-to-RGB image translation, targeting spatial-detail recovery and spectral fidelity.
- **Clinical NLP:** Co-developed an **LSTM–CRF** model for medical concept extraction from clinical text, modeling contextual dependencies and structured label sequences.
- **ML Engineering:** Built end-to-end research pipelines for data preprocessing, augmentation, model training, hyperparameter experimentation, baseline comparisons, and ablation studies across multiple ML tasks.
- **Research Publications:** Authored and co-authored peer-reviewed papers on image reconstruction, cross-spectral translation, remote sensing, and clinical NLP.

Key Projects

RadMeasure — Medical Measurement Agent GitHub

Python · LLM Agents · Medical Imaging · Geometry Verification · Human-in-the-Loop

- Built a medical measurement agent with a **KEEP/REPAIR/STOP controller** that constrains LLM-proposed actions to registered measurement protocols and tools.
- Implemented geometry verification, bounded repair, human review for unresolved cases, and execution tracing for controlled and auditable measurement workflows.
- Evaluated a selective repair policy offline on 176 cases from **HVAngleEst**, applying corrections to **20.1% of prediction records** and reducing overall MAE from **2.68° to 2.54°** versus no repair.

SQL-Agent — Text-to-SQL Agent & Evaluation GitHub

Python · LangGraph · FastAPI · MCP · SQLite/PostgreSQL · Hybrid RAG · Docker

- Built a **natural-language SQL assistant** for reporting with schema grounding, bounded repair, advisory verification, and controlled execution; added durable FastAPI/MCP jobs, Docker deployment, and regression checks.
- Implemented optional **BM25/vector hybrid retrieval and cross-encoder reranking** over schema and business-rule documents, with database/table filtering and separate retrieval evaluation.
- Evaluated on 500 official **BIRD Mini-Dev** questions using full schema context, a Qwen3-4B Planner, and a Qwen2.5-Coder-7B advisory Verifier; achieved **39.2% execution accuracy versus 36.0%** for the local PV-SQL (Probe–Verify) configuration, with **64.5% fewer tokens per correct task** and **3.66 s p95 latency**.

SMLP-KAN — Diffusion-Based Hyperspectral Image Restoration GitHub

Python · PyTorch · Diffusion Models · Kolmogorov–Arnold Networks · Attention · Remote Sensing

- Developed an **unsupervised diffusion-based framework** for hyperspectral sharpening, denoising, and inpainting, combining spectral diffusion priors with panchromatic guidance without requiring paired clean-image training data.
- Designed a hybrid **MLP–KAN architecture** with attention-based feature refinement to model spectral–spatial relationships, suppress redundant features, and preserve spatial detail.
- Evaluated restoration across five hyperspectral datasets with component ablations; used **97.6% fewer parameters** than PLRDiff while improving **Botswana** denoising PSNR by **2.21 dB** and SSIM by **4.2%** at $\sigma = 0.1$.

SF-GPT — NIR-to-RGB Image Translation GitHub

Python · PyTorch · Transformers · Cross-Spectral Learning · Image Translation

- Developed a **spatial-frequency guided transformer** combining pixel-domain and frequency representations to reconstruct visible-color imagery from near-infrared inputs.
- Designed Spatial-Frequency Guided Multi-head Self-Attention (**SFG-MSA**) to fuse DCT- or DWT-derived features with pixel representations, with channel and spatial attention for texture and structural refinement.
- Evaluated SF-GPT on the **IEEE VCIP 2020 NIR Image Colorization Challenge** dataset; outperformed the **first-place method**, ATcycleGAN, with **1.53 dB higher PSNR**, **2.6% higher SSIM**, and **4.1% lower angular error**.

FW-SAT — Guided Thermal Image Super-Resolution GitHub

Python · PyTorch · Transformers · Window-Based Attention · Multimodal Image Fusion · Super-Resolution

- Developed a **transformer-based thermal super-resolution model** combining low-resolution thermal images with aligned high-resolution visible imagery to reconstruct thermal images with finer structural details.
- Designed Flexible Window-based Self-Attention (**FWA**) to aggregate information across neighboring windows, integrating channel and spatial attention with window-based self-attention to capture global, regional, and local features.
- Evaluated thermal super-resolution on the **PBVS 2024 GTISR Track 2** validation set, achieving **4.6%** and **6.9%** higher SSIM than HAN at 8× and 16× upscaling, respectively, with attention ablations.

Education

Ph.D., Electrical & Computer Engineering (GPA: 4.0/4.0) **Jan 2020 – May 2025**
University of Missouri–Kansas City Kansas City, MO

M.S., Computer Science (GPA: 3.8/4.0) **Aug 2017 – Dec 2019**
University of Missouri–Kansas City Kansas City, MO

Selected Publications (9 First-Author)

H. Jiang et al., "GeoVIRA: Geometry–Visual Interfaces for Radiographic Angle Measurement," *IEEE/CVF WACV Workshop (FABRIC)*, 2027. (Accepted)

H. Jiang et al., "Does Predicting Anatomical Axes Improve Foot Angle Measurement? Recovery, Measurement Utility, and Reference-Axis Substitution," *IEEE/CVF WACV Workshop (LVR)*, 2027. (Accepted)

H. Jiang et al., "SMLP-KAN: Spectral MLP-KAN Diffusion Prior for Hyperspectral Image Restoration," *IEEE/CVF CVPR Workshop (PBVS)*, 2026.

H. Jiang and Z. Chen, "Spatial-Frequency Guided Pixel Transformer for NIR-to-RGB Translation," *Infrared Phys. Technol.*, vol. 148, Art. no. 105891, 2025.