

Xinye Yang

Seattle, WA | +1-401-499-3747 | [Email](#) | [LinkedIn](#) | [GScholar](#) | [Website](#)

Software engineer at AWS Kiro building **agentic AI** developer tools end-to-end — LLM context/memory systems, **multi-agent orchestration** and automation frameworks, **inference optimization and model serving**, and full-stack product engineering.

EDUCATION

- **Brown University** Providence, RI
Master of Science in Computer Science (GPA: 4.0/4.0, straight A's) Sep. 2024 – May 2026
 - Courses: Software Engineering, Database Systems, Multi-Agent Systems, Self-Supervised Learning, etc.
- **University of Science and Technology Beijing** Beijing, China
Bachelor of Engineering in Artificial Intelligence (GPA: 3.83/4.0) Sep. 2020 – Jun. 2024
 - Courses: Algorithms and Data Structures, Operating Systems, Computer Networks, Linear Algebra, etc.

EXPERIENCE

- **Software Development Engineer** Jul. 2026 – Present
*Amazon Web Services (AWS) | Agentic AI (**Kiro**) Team* Seattle, WA
 - **Design an AI automation framework for agent-driven workflows**, standardizing multi-step **tool calling and orchestration (MCP)**, execution state management, and failure recovery to make long-running agent sessions reliable and observable.
 - **Evolve LLM context and memory systems** (compaction, retrieval, context scheduling under token budgets), **reducing inference cost and latency** while preserving response quality for long-running developer workflows.
 - **Built CI pipelines and integration-test infrastructure** for AI features, adding **automated LLM evaluation (evals)** and regression gates that raise launch confidence and release velocity.
 - **Contributor to KiroCrew** (open-source **multi-agent** workspace built on **kiro-cli** and the **Agent Client Protocol (ACP)**): improved gateway-restart resilience for **persistent, checkpointed agent sessions** (message redelivery / duplicate-frame handling) and hardened **redaction and sandboxing** across WebSocket and Slack egress paths.
- **Research Assistant (Part-time)** Jul. 2026 – Present
University of Colorado Anschutz School of Medicine Remote
 - Research **AI agents and reliable model serving for medical imaging**: reliability stress testing and **decision-time routing** for chest X-ray vision-language models, and **confidence-gated cloud-edge cascade inference** for cost- and latency-aware triage (CHASE 2026; *Smart Health* 2026).
- **Software Development Engineer Intern** Jun. 2025 – Aug. 2025
*Amazon Web Services (AWS) | Agentic AI (**Kiro**) Team* Seattle, WA
 - **Designed and implemented the core distributed context compaction system** for Kiro IDE using TypeScript and Node.js, combining hierarchical recursive compression with RAG powered by Amazon S3 Vectors and OpenSearch Serverless, achieving 80–90% compression while ensuring scalability and reliability.
 - **Drove the production launch serving 100K+ users**, improving distributed system performance and context relevance through rigorous benchmarking, strategy evaluation, and close collaboration with applied scientists.
 - **Built a full-stack distributed solution** integrating the VS Code Extension API with backend RESTful services and a vector retrieval pipeline using Bedrock Embeddings and LangChain; resolved critical UI/UX issues, **enhancing developer productivity and overall experience**.
- **Software and Machine Learning Engineer** Sep. 2023 – May 2024
Chinese Academy of Sciences, Institute of Automation Beijing, China
 - Developed a **real-time patient monitoring system** with **high availability** and **scalable architecture**, securing a software patent and national grant, and significantly improving ICU response speed.
 - Designed and deployed a hybrid solution combining **PyTorch-based Key Point Detection, Optical Flow Analysis**, and a **React web app** for **real-time data processing and visualization**, integrating **MySQL** and **MQTT** to achieve high performance and scalability.

- **Reliability Stress Tests and Decision-Time Routing for Chest X-ray Vision–Language Models.** [Xinye Yang](#), Zhusi Zhong, Scott Collins, Grayson Baird, Xuyu Wang, Zhicheng Jiao. *IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE 2026)*.
- **Confidence-Gated Cloud-Edge Cascade Triage via Variational Risk Minimization for Medical Imaging.** [Xinye Yang](#), Zhusi Zhong, Scott Collins, Michael Bernstein, Grayson Baird, Terrence Healey, Michael Atalay, et al. *Smart Health*, 100689, 2026.
- **U.S. Provisional Patent:** An Integrated Platform for Intelligent Radiology Reporting and Evaluation. Inventors: [Xinye Yang](#), Zhusi Zhong, Scott Collins, Michael Bernstein, Grayson Baird, Zhicheng Jiao. *Filing Date: January 28, 2026*.
- **A Unified Platform for Radiology Report Generation and Clinician-Centered AI Evaluation.** Zhuoqi Ma, [Xinye Yang](#), Zach Atalay, Andrew Yang, Scott Collins, Harrison Bai, Michael Bernstein, Grayson Baird, Zhicheng Jiao. *medRxiv*, 2025. [[Link](#)]
- **The AI Challenge: A Turing Test Pilot Study of Attendings and Residents in Identifying AI-Generated Content.** Zhuoqi Ma, [Xinye Yang](#), Zachary Atalay, Zhusi Zhong, Scott Collins, Harrison Bai, Michael Bernstein, et al. *Meta-Radiology*, 100199, 2025.
- **Understanding the Brain with Attention: A Survey of Transformers in Brain Sciences.** Cheng Chen, Huilin Wang, Yunqing Chen, Zihan Yin, [Xinye Yang](#), Huansheng Ning, Qian Zhang, Weiguang Li, et al. *Brain-X*, 1(3), e29, 2023.
- **Optimization of CNN for Diagnosis on Lung Disease by Lung Segmentation and Rib Suppression.** [Xinye Yang](#), Yuhang Liu, Zhiwei Lin, Lingyu Zhong, Teoh Teik Toe. *2022 15th International Symposium on Computational Intelligence and Design (ISCID)*, pp. 254–259, IEEE, 2022.

TECHNICAL SKILLS

Languages: TypeScript, JavaScript, Python, Java, C/C++, Go, SQL, HTML/CSS

AI & LLM Systems: LLM inference optimization (SGLang), model serving & routing, RAG, vector search, multi-agent systems, agent orchestration (LangChain, LangGraph, CrewAI, MCP, ACP), tool calling / function calling, context engineering, prompt engineering, LLM evals, fine-tuning, PyTorch, TensorFlow

Cloud & Infrastructure: AWS (Bedrock, S3 Vectors, OpenSearch Serverless), Docker, CI/CD (GitHub Actions), microservices, distributed systems, observability, performance profiling, Linux, Git

Frameworks: React, Vue, Node.js, Flask, FastAPI, SpringBoot, MySQL, PostgreSQL, Redis