

Weifeng Hu

Software Development Engineer • 7 Years of Experience • U.S. Permanent Resident (Green Card)

📍 Seattle, WA | ✉ weifhu@umich.edu | 📞 weifhu0124 | 🌐 weifeng-hu

Experience

Amazon Prime Video - PV Personalization & Discovery
Software Development Engineer II

Seattle, WA
Sep 2026 - Present

- Member of the Prime Video Data Platform team.

Amazon - Luna & Prime Gaming
Software Development Engineer II

Seattle, WA
Dec 2025 - Sep 2026

- Architected a Bedrock Knowledge Base with hybrid semantic and keyword retrieval, unifying access to 100K+ enterprise documents from Confluence, SharePoint, and other organizational sources.
- Designed and deployed an AgentCore-hosted MCP server, providing enterprise AI agents with secure, standardized access to organizational knowledge under a sub-second latency SLA.
- Integrated agentic RAG workflows into employees' daily tools, reducing manual information retrieval effort by 2+ hours per employee per week.

Amazon Web Services - Amazon Software Builder Experience
Software Development Engineer II

Seattle, WA
Apr 2022 - Dec 2025

- Architected and delivered an LLM-powered enterprise data platform, powering 3 Amazon GenAI services used by 1 million+ employees.
- Built production-grade Java APIs for retrieval and LLM orchestration, supporting 600K+ distinct queries within three months.
- Designed event-driven metadata enrichment pipelines using Step Functions, Lambda, and Python, automating metadata processing for 20 million+ internal documents.
- Designed and implemented a decoupled, three-stage Glue ETL pipeline, scaling batch computation of search-relevance metrics to petabyte-scale user event data.
- Incorporated NoSQL ingestion workflows for high-throughput processing, sustaining hundreds of transactions per second with low-latency performance.
- Mentored junior engineers and led cross-organizational design reviews, establishing scalable engineering practices across teams.

Software Development Engineer I

Aug 2020 - Apr 2022

- Refactored a React application serving 30K daily active users, cutting page load times by 40%.
- Productionized a document similarity model on SageMaker and redesigned its serving APIs, enabling more comprehensive content discovery features.
- Developed a distributed ingestion system with REST APIs, processing millions of daily document operations while eliminating reliability bottlenecks.
- Scaled and hardened production Elasticsearch clusters, reducing operational overhead by 20% and eliminating critical node failure modes.

Amazon
Software Development Engineer Intern

Seattle, WA
Jun 2019 - Sep 2019

- Built syntax analysis tooling that transforms abstract syntax trees into JSON representations, enabling standardized downstream text analysis.
- Integrated the analysis pipeline with DynamoDB, improving performance and expanding its data processing capabilities.

Didi Chuxing
Software Development Engineer Intern

Beijing, China
May 2018 - Aug 2018

- Developed an XGBoost-based forecasting system to predict regional ride-completion rates, enabling localized driver-supply planning.

- Built a new data visualization interface and hardened the supporting codebase to improve application usability and reliability.

Skills

Languages: Java, Python, Spark, React, TypeScript, C/C++

AI/ML: LangChain, LangGraph, RAG, LLM, MCP, Neural Networks, Recommender Systems

AWS: Bedrock, DynamoDB, ECS, S3, Lambda, Glue, Elasticsearch/OpenSearch, SQS, SNS, SageMaker

Data: Event-Driven Architectures, Batch & Streaming ETL Pipelines, Data Analytics & Visualization

System Design: Distributed Systems Design, API Design, Scalability, Availability, Performance Optimization

Education

University of California, San Diego

M.S. Computer Science (GPA: 3.90/4.00)

Graduate Student Association Conference Presentation Travel Grant Recipient

San Diego, CA

Aug 2018 – May 2020

University of Michigan, Ann Arbor

B.S. Computer Science (GPA: 3.83/4.00)

College of Engineering Class of 1935 Scholarship Recipient

Ann Arbor, MI

Aug 2014 – May 2018

Licenses & Certifications



**AWS Certified Generative AI
Developer – Professional**

Amazon Web Services

Issued Aug 2026 – Expires Aug 2029



**Claude Certified Developer –
Foundations**

Anthropic

Issued Jul 2026 – Expires Jul 2027



**Claude Certified Architect –
Foundations**

Anthropic

Issued Apr 2026 – Expires Apr 2027

Projects

Adaptive Self-RAG for Amazon Agentic AI Offerings

Seattle, WA

- Built a self-refining RAG system with LangGraph using a Chroma vector store to answer technical questions on Bedrock AgentCore, Amazon Q Business, and Amazon Q Developer.
- Designed a multi-agent evaluation loop where specialized LLM agents check answers for hallucinations, factual grounding, and alignment with retrieved context.
- Implemented a router agent that classifies out-of-scope queries and automatically triggers web search via Tavily for questions unrelated to Amazon's Agentic AI offerings.

Reflexion Agent on Stock Picking Strategy

Seattle, WA

- Built an AI agent that performs Tavily search to draft an essay describing stock-picking strategies.
- Implemented a Reflexion-style self-improvement loop using a critic agent that evaluates reasoning quality and iteratively revises the draft for clarity and accuracy.

LinkedIn Agent with LangChain

Seattle, WA

- Built an AI agent that uses Tavily Search and Scrapin.io to retrieve LinkedIn profiles and generate concise summaries and personalized “fun facts.”
- Developed a Flask-based frontend to visualize agent outputs, including structured Pydantic objects and intermediate reasoning steps.

Research

- **Hu, W.**, A Data-Driven Study of Cross-Cultural Social Impressions on Faces. UCSD Thesis 2020.
- Song, A., **Hu, W.**, Yadav, D. P., Wen, F., Zuo, B., Vul, E., & Cottrell, G. (2020). Do you see what I see? A Cross-cultural Comparison of Social Impressions of Faces. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 42.
- Song, A., Yadav, D. P., **Hu, W.**, Cottrell, G., & Vul, E. (2020). To Dye or Not to Dye : The Effect of Hair Color on First Impressions. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 42.
- **Hu, W.**, Yang, R. (2020). Predicting the Success of Kickstarter Projects in the US at Launch Time. In: Bi, Y., Bhatia, R., Kapoor, S. (eds) Intelligent Systems and Applications. IntelliSys 2019. Advances in Intelligent Systems and Computing, vol 1037. Springer, Cham.

Teaching

- Neural Networks & Pattern Recognition (UCSD, Winter 2020)
- Recommendation Systems (UCSD, Fall 2019)