Zhihao Lin

Seattle, WA zhihaolinsde@gmail.com GitHub LinkedIn

Summary

Software Engineer with 3+ years at AWS building production distributed systems supporting 20M+ daily customer interactions. Experienced in backend architecture, cloud infrastructure, retrieval-grounded LLM workflows and tool integration, with hands-on experience building LLM inference and evaluation pipelines using open-weight models.

Skills

Programming Languages: Java, Python, TypeScript

Cloud & Data: AWS (EC2, Fargate, Lambda, DynamoDB, OpenSearch, S3, SNS/SQS, CloudWatch, CloudFormation)

Systems & Tools: Distributed systems, data pipelines, CI/CD, observability, Claude Code

AI/ML: OpenAI API, Llama, PyTorch, vLLM, RAG, LLM inference, tool calling, LLM evaluation

Industry Experience

Applied AI Engineer May 2026 – Sep 2026
Binhao Technology *Remote*

- Embedded with business and engineering users to map end-to-end OEM/ODM workflows and translate operational requirements across customer intake, quotation, BOM management, tooling, sampling, procurement, and production preparation into scoped AI automation projects.
- Built Python-based unified project data pipeline to ingest Excel, Word, PDF, PPT, image, and system-export data, normalizing product information while tracking field-level provenance, versions, confidence, and conflicts.
- Developed LLM-powered document intelligence workflows to extract structured customer requirements, retrieve similar historical products and BOMs, and generate grounded quotation and sample-BOM drafts while identifying missing and conflicting information for review.
- Automated downstream document and workflow generation for quotations, design requests, tooling applications, T1/T2 sample notices, procurement requests, and BOM updates, incorporating human-in-the-loop approval and validation guardrails.

Software Development Engineer Feb 2023 – Apr 2026
Amazon Web Services — Amazon Connect *Seattle, WA*

- Built and operated customer-facing metrics platform powering near-real-time contact-center analytics, handling **280K TPS** and supporting metric computation, data storage, validation, and API-based metric access as daily customer interactions grew from 10M+ to **20M+**.
- Led end-to-end delivery of outbound-campaign metrics, translating product requirements into metric definitions and driving data contract alignment across upstream and backend systems; resolved launch-blocking schema gaps and extended the platform for new data structures. Worked backward from launch to coordinate dependencies, drive launch-readiness reviews and org-wide bug bashes, and execute rollout across 10 AWS Regions.
- Owned and productionized a metrics validation service using AWS Lambda, DynamoDB, OpenSearch, and S3, expanding pre-launch validation coverage to over **90%**, reducing customer complaints related to metric definitions by approximately **70%**, and cutting investigation time for related support tickets from one day to two hours.
- Extended the validation platform with retrieval-grounded LLM workflows and structured tool interfaces for metric onboarding, validation, and debugging, using DynamoDB and OpenSearch context with validation guardrails.
- Led service expansion into **six** new AWS regions, coordinating cross-team dependencies and launch readiness; introduced an AI-assisted SOP-to-workflow prototype that reduced expansion cycles from two weeks to one.
- Owned the design and delivery of an asynchronous cross-region OpenSearch data-deletion workflow with load-aware throttling, idempotent execution, and verification to meet compliance requirements without impacting availability.
- Owned CI/CD, monitoring, and on-call operations for services maintaining 99.95%+ availability.

Software Development Engineer Intern Jun 2022 – Sep 2022
Amazon Web Services — Amazon Connect *Seattle, WA*

- Prototyped a metrics validation service using AWS Lambda, DynamoDB, OpenSearch, and S3; demonstrated it across the organization and later productionized it as a full-time engineer.

AI Systems & Research Experience

AI Systems Research Assistant Sep 2026 – Present
University of Virginia *Remote*

- Conducting systems research on efficient LLM inference, focusing on KV-cache management, serving efficiency, and runtime behavior in vLLM-based workloads.
- Profiling concurrency and inference bottlenecks in agentic workloads and evaluating system optimization strategies.

Independent LLM Researcher Jan 2026 – Sep 2026
Independent Research *Seattle, WA*

- Co-developed Scenario-Actor for adaptive decoding on Llama-3.1-8B-Instruct; led all experiments in Python and PyTorch, including hidden-state extraction, decoder-component training, inference, and evaluation.
- Built an experimental pipeline for training and validation on 37K+ prompts, decoding benchmarks, and ablations; paper accepted at **AACL-IJCNLP** 2026 Main Conference.

Sentiment Analysis for Email Interaction and Fake News Research Jan 2022 – Apr 2022
Research Assistant *Santa Clara, CA*

- Contributed to an industry-collaborative NLP research project on sentiment analysis by reviewing related literature, participating in research discussions, and deploying experimental models on AWS EC2.
- Built a Chrome/Gmail sentiment-analysis plugin prototype that integrated NLP inference into email workflows for lightweight sentiment interpretation.
- Developed reusable research software that crawled web images associated with news reporters and combined Amazon Rekognition with DeepFace for face attribute analysis, producing structured reporter data for downstream fake-news detection research; packaged the software for use by a PhD researcher.

Publications

Read the Room Before Your Model Generates 2026
*Muyao Kong**, ***Zhihao Lin**** *— AACL-IJCNLP 2026 Main Conference (Accepted)*
A scenario-adaptive decoding framework that derives prompt-level sampling temperatures from interpretable scenario representations learned from Transformer hidden states. *Equal contribution.

Education

Santa Clara University Sep 2020 – Dec 2022
Master of Science in Computer Science — GPA: 3.77/4.0 *Santa Clara, CA*

University at Buffalo, The State University of New York Aug 2016 – Jun 2020
Bachelor of Science in Computer Science — Honors: cum laude; Dean's List (2019) *Buffalo, NY*