Jie Zhu

jiezwork@gmail.com ♦ 6 Jessica Dr, Woburn, MA, 01801 ♦(540) 257-4581

EDUCATION

Virginia Polytechnic Institute and State University — ***VA, United States***
M.Eng. in Machine Learning, Computer Engineering — *Sep 2019-Dec 2020*
Member of *IEEE HKN* (Virginia Tech, Beta Lambda Chapter)

Huazhong University of Science and Technology — ***Wuhan, China***
M.S. in Microelectronics and Solid-State Electronics — *Sep 2016-Jun 2019*

University of Science and Technology of China — ***Hefei, China***
B.S. in Applied Physics — *Sep 2012-Jun 2016*

TECHNICAL SKILLS

Programming Skills: Java, Python, JavaScript, C#, PowerShell, Shell, C/C++, HTML/CSS, Ruby, Scala, Groovy, Perl
Tools: vLLM, PyTorch, Transformer, Amazon SageMaker, LangChain, LangGraph, OpenCV, TensorFlow, Pandas, SciPy, NumPy, SK-Learn, Kubernetes, Spring, Guice, .Net Framework, Jenkins, Splunk, Terraform, Tavern, JVM, JUC, CAS, ReactJS, Redux, Thunk, JQuery, Axios, Swagger UI, Postman, Nginx, Docker, MySQL, MongoDB, MATLAB, Mathematica

WORK EXPERIENCE

Amazon, Machine Learning Engineer II — ***Boston, MA, United States***
Focus on Amazon AGI Customization Service — *Aug 2025 - present*

- Conduct research for cutting-edge Large Language Models (LLMs) and develop Generative AI customization solutions
  - Nova LLM Training techniques and optimizations
  - Customization capabilities through fine-tuning and distillation
  - Multimodal Large Language Models' leveraging heterogeneous data sources
  - Developing high efficiency deployment platform for Large Language Model

Amazon, Software Development Engineer II — ***Boston, MA, United States***
Focus on Amazon People Experience Technology — *Dec 2023 – Aug 2025*

- Designed and developed new features in world largest Human Capital Management system to handle million level of high volume hiring
  - Designed and implemented new hiring model for automatic merge and acquisition branch, reducing acquisition process time **from weeks to hours**
  - Optimized the hiring documentation process from handwriting to electronic file, reducing **80% paperwork** during hiring process
  - Improved the contingent worker onboarding experience, reducing the start date delay rate **by 20%**
  - Integrated the generic hiring process with datasets from multiple domains including background screening, drug testing and labor eligibility verification, **improving the precision of contingency check**
- Designed and implemented the GenAI solution for advanced hiring automation
  - Optimized the error resolution process by **AI self-correction**, reducing the manual process to resolve the data error during hiring process
  - Created a new **Generative AI people searching tool** for information searching, improving the efficiency of human resource work
  - Created a new **Generative AI chat bot** based on human resource knowledge base data model to provide AI consultation solution, reducing the average consultation procedure **from day level to hour level**

Amazon, Software Development Engineer — ***Boston, MA, United States***
Focus on Alexa Audio Voice Experience Growth — *May 2022 – Dec 2023*

- Developed new features in Alexa Audio Voice infrastructure (**Java Spring Based server** and **AWS architect solution**) to increase user music experience on voice assistance device (Monthly Active Users, Audio Experience Penetration, High Impact Action)
  - Alexa Audio **Out-Of-Box Experience**: monthly user completion rate increased by **30%**

- Alexa Audio **Content Injection**: Achieved integration goal with different music providers (**Amazon Music**, **Spotify**, **Apple**)
- Alexa Audio **Routines**: Enhanced Alexa Routines with music action (Voice, Touch, End of Play Queue)
- Alexa Audio **Multi-Device Enhanced experience**: Enabled user audio experience with multiple devices like home concert.

- Participated in the enhanced Skill Connections interface development (backend is based on a combination of Pipeline Pattern and Strategy Pattern) to provide music experience for 1p/3p skills (**AmazonMusic Spotlight**, **Alexa Smart Home**, **Hey Disney**, etc.)
- Designed and implemented **Skill Connection** Callback signal solution to enhance music experience with external domains and enable 1p/3p skills continuing user interaction
- Designed and implemented **Alexa Routines** Music burst traffic solution (based on a **distributed cache strategy**) with external teamwork to reduce high peak hours traffic load by **30%** hit rate and latency by **50%**
- Participated the **LLM (Large Language Model)** integration with Alexa Music Voice Experience

Citrix Systems, Software Development Engineer ***Burlington, MA, United States***
Focus on Application Layering and Cloud Service (ALCS) *Jan 2021- May 2022*

- Participated the **SAFe** process to work with team on planning and developing the enterprise-level project, Citrix Image Portability Service (Citrix IPS), which gives a practical solution to migrate server images from on-premises resource location to different cloud platforms including **Azure**, **GCP**, **Vsphere** and **AWS**
- Developed new features in multiple backend services of Citrix IPS based on different languages including **C# .Net Core**, **Python** and **PowerShell**, and deployed the features to the **Kubernetes clusters** using **Helm Chart**.
- Contributed on integration test by developing API testing cases using **Tavern** framework
- Implemented new virtualization features on **Splunk** dashboard for support team to monitor the health status of services
- Designed and implemented **RESTful API** features using **C# .Net 5** and documented the features by **Swagger UI** for tech preview and release
- Designed and implemented **PowerShell Module** package, including a group of **PowerShell Cmdlets** to conduct multiple phases of image migration task, as an alternative user interface of IPS product with demonstration and documentation.
- Improved a **Terraform** solution to achieve the automation of multi-geo infrastructure deployment including **Azure Kubernetes clusters**, **Azure Frontdoor** service and monitoring service
- Improved **Jenkins** pipeline by a parallel structure using **Groovy** to reduce the time cost of automation deployment by > **50%**

PUBLICATIONS AND CONFERENCE PRESENTATIONS

- J. Lin, Y. Zeng, **J. Zhu**, et al, The 14th International Conference on Learning Representations **(ICLR 2026)**.
- **J. Zhu**, K. Yang, The 6th International Conference on Electrical Computer and Energy Technologies **(ICECET 2026)**.
- C. Yu, F. Liu, **J. Zhu**, et al, The 5th Asia-Pacific Conference on Communications Technology and Computer Science **(ACCTCS 2025)**.
- **J. Zhu**, J. He, et al, **IEEE Electron Device Letters**, Vol 40, 9 (2019).
- J. He, **J. Zhu**, C.C. Ma, et al, **Applied Physics Letters** 115, 7 (2019).
- C.C. Ma, J. He, J. Lu, **J. Zhu**, et al, **Applied Science** 8, 1238 (2018).
- C.C. Ma, J. He, J. Lu, **J. Zhu**, et al, **Japanese Journal of Applied Physics** 58, 10 (2018).
- J. Lu, C.C. Ma, J. He, **J. Zhu**, et al, Physica B: Condensed Matter 566, 116-120 (2019).
- **J. Zhu**, C.C. Ma, et al, **MATEC Web Conf.**, 201, 06002 (2018).
- International Conference on Smart Science, 2018, Osaka, Japan.
- The 4th International Conference of Advances in Functional Materials, 2018, Nanjing, China.