Pranshu Gupta

pranshug258@outlook.com • +1 (425) 679-2402 • https://linkedin.com/in/pranshug258/ • https://pranshug.com

Professional Experience

I am a senior software engineer with 7 years of experience, specializing in building highly available, low latency, distributed systems. I have worked on systems that power Microsoft Azure and LLM Inference on Azure AI.

Microsoft Azure OpenAI • Senior Software Engineer • Jan 2025 – Present

Led the development and architecture for data plane APIs serving all large language model inference on Azure.

- Led the design and implementation of a scalable, distributed rate limiter to serve 2 billion requests/day.
- Contributed to the design and implementation of Azure AI's prompt router, launched at Microsoft Build 2025.

Microsoft Azure Core • Senior Software Engineer • Mar 2024 – Jan 2025

Software Engineer II • Jun 2020 – Mar 2024

Led the development and architecture for resource provisioning in Azure Resource Manager, the unified platform for deploying and managing all resources on Azure.

- Led and delivered improvements in bulk delete API, reducing P90 latency by 50% and P99 by 65%.
- Designed and led the development of background job quota-based throttling in Azure control plane.
- Implemented a critical security patch and helped more than 200 teams in Azure onboard to it.
- Implemented callback design for resource actions, which reduced the number of polling requests by 60%.
- Implemented dependency handler in resource group deletion, which reduced failures by 98%.

Microsoft Core Services • Software Engineer • Jun 2017 – Aug 2019

Designed and implemented modules for Customer Data Enrichment Service, which enriches customer data from marketing campaigns to help marketers create better sales opportunities for Microsoft.

- Migrated to serverless architecture, reducing operation costs by 90% and improving performance by 3x.

Technical Skills

- Programming Languages : C#, Python, C++, PowerShell, Azure Bicep, Kusto QL, HTML/CSS/JS, Markdown
- Cloud and Infrastructure : Microsoft Azure, .NET Core/Framework, Redis, Azure CLI, Service Bus
- AI and ML : Generative AI Inference, OpenAI API, GitHub Copilot, PyTorch, Deep Reinforcement Learning

Education

Georgia Institute of Technology, Atlanta • May 2020 • MSCS (ML specialization)

Indian Institute of Technology, Kanpur • May 2017 • B.Tech. CS

Publications

- Saha, K., Gupta, P. et al. Observer Effect in Social Media Use | Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems - https://dl.acm.org/doi/10.1145/3613904.3642078
- El Sherief, M., Saha, K., Gupta, P. et al. Impacts of school shooter drills on the psychological well-being of American K-12 school communities – https://www.nature.com/articles/s41599-021-00993-6

Selected Projects

- Image completion with statistics of patch offsets – 44 stars on GitHub.
- Crowd behavior analysis – detect outliers in the trajectories of entities in a moving crowd using the minimum description length principle – 39 stars on GitHub.
- Deep image captioning – generate captions for an image using deep reinforcement learning model with embedding reward – 26 stars on GitHub.