Abhinav Kumar Singh

abhinavsingh.dev
Email: abhinav@abhinavsingh.dev
Mobile: +61-474766857

Education

- **The University of Melbourne** Melbourne, Victoria
  Master of Information Technology; WAM: 84.00 *Feb 2019 – Jan 2021*
- **SRM Institute of Science and Technology** Chennai, India
  Bachelor of Technology in Information and Telecommunication; GPA: 9.8 *Aug 2012 – July 2016*

Experience

- **WooliesX** Melbourne, Australia
  Senior ML Engineer *Mar 2022 – Present*
  - **Graph-Based Real-Time Personalization**: Built a high-throughput in-session recommendation system using GNNs, Neo4j, and feature stores, delivering low-latency, scalable recommendations for millions of users. Optimized real-time serving with Kubernetes, Redis, and async microservices, improving engagement and reducing inference latency.
  - **Scalable ML Inference Platform**: Designed an ML serving system with FastAPI, Kubernetes, and gRPC microservices, achieving 10x faster inference throughput. Integrated Redis caching and MLflow for seamless model management, A/B testing, and zero-downtime rollouts.
  - **AI-Powered Customer Support Assistant**: Developed an Agentic AI assistant using LangGraph and LangSmith, enabling structured reasoning and real-time debugging. Integrated RAG with vector databases (Weaviate, FAISS) and knowledge graphs for instant contextual retrieval, boosting issue resolution by 10% and customer satisfaction by 5%.
- **Servian** Melbourne, Australia
  ML Engineer *Oct 2021 – Mar 2022*
  - **Real-Time ML Serving Infrastructure**: Engineered a high-throughput, low-latency ML backend using Python, FastAPI, and Kubernetes on Azure, enabling real-time inference at scale. Designed a containerized microservices architecture with gRPC, Redis caching, and async processing, reducing response times 10x. Automated CI/CD with Terraform and GitHub Actions, ensuring zero-downtime rollouts with real-time monitoring via Prometheus and Grafana.
- **6Clicks** Melbourne, Australia
  AI Engineer *Feb 2021 – Oct 2021*
  - **Hailey AI Engine**: Designed and optimized the AI core of 6Clicks' GRC platform, leveraging NLP and transformer-based models to automate compliance analysis—reducing manual effort from days to seconds.
  - **Scalable AI Integration**: Built a real-time, multi-tenant AI service with API-driven insights, ensuring high availability and scalable inference for regulatory automation.
- **EY** Bengaluru, India
  Associate Software Engineer *Aug 2016 – Sep 2018*
  - **Enterprise-Grade Backend Development**: Developed scalable, high-performance microservices in a distributed, event-driven architecture, ensuring low-latency, fault-tolerant operations. Optimized database performance through query tuning, indexing, and connection pooling, while implementing structured logging, API versioning, and CI/CD automation for seamless deployments at scale.

Technical Skills

- **Programming Languages and Frameworks**: Python, C++, Golang, TensorFlow, PyTorch, Kubeflow, Kubernetes, FastAPI, LangChain, Javascript
- **Database**: PostgreSQL, MongoDB, Google BigQuery, Google BigTable, Google Firestore
- **Tools**: Git, Jira, Jenkins, Github, Docker, Terraform, VS Code, Postman, Cursor
- **Cloud**: Google Cloud, Microsoft Azure, AWS

Honors and Awards

- **AI Katas Champion** Sydney, Australia
  O'Reilly Media *Nov 2024 – Dec 2021*
- **Melbourne Engineering Scholarship** Melbourne, Australia
  University of Melbourne *Feb 2019 – Jan 2021*
- **Top Graduate in Department** Chennai, India
  SRM Institute of Science and Technology *Aug 2012 – July 2016*
- **SRM Engineering Scholarship** Chennai, India
  SRM Institute of Science and Technology *Aug 2012 – July 2016*