Supreet Singh Dhillon

☎ 604-344-0493 ✉ supreet.singhdhillon@icloud.com in supreetsdhillon SupreetSinghDhillon

Professional Summary

AI Engineer specializing in agentic system design, LLM orchestration, and private-by-default architectures. Proven track record building real-world applications with LangChain, LangGraph, OpenAI, Gemini, Mistral, and LLaMA—reducing manual workload by over 90% in domains like logistics, legal tech, and healthcare. Experienced in deploying scalable backends using Python and Azure Functions, with deep focus on semantic retrieval, prompt engineering, and structured memory management for production AI systems.

Education

Simon Fraser University, Burnaby BC **Aug 2024**

Bachelor of Applied Science in Computer Science

Technical Skills

AI/ML Tooling: LangChain, LangGraph, RAG, OpenAI, Gemini, Ollama, HuggingFace, scikit-learn, Mistral, LLaMA
LLM Ops: Prompt chaining, agent orchestration, semantic search, vector stores, privacy-first AI architectures
Languages: Python, JavaScript, TypeScript, C#, C++, Java, C, SQL
Backend & Infrastructure: Node.js, Azure Functions, AWS (Lambda, S3, EC2, DynamoDB, SQS), Docker, Kubernetes
Databases: PostgreSQL, DynamoDB, Redis, MS SQL, Azure SQL, Realm (on-device)
Frontend: React, Angular, React Native, HTML/CSS
Developer Tools: Git, GitHub Actions, Jira, Postman, Insomnia, Android Studio, VS Code, IntelliJ, Cursor, Windsurf
Methodologies: Agile/Scrum, Test-Driven Development (TDD), CI/CD pipelines
API Design: REST, GraphQL, JSON-based APIs

Work Experience

Founder, CEO, Head of Agentic AI Development **May 2024 – May 2025**

LoadMinds (Acquired by MG Transport Ltd.)

- **Built LoadMinds from scratch** as a full-stack AI dispatcher to eliminate the need for human dispatchers across quoting, customs, invoicing, and email workflows.
- **Designed a routing system** that parsed 400+ daily emails, classified intent (quote, customs, status update, billing, etc.), and passed them to specialized agents that could act without human input.
- **Engineered the core agent runtime using LangChain and LangGraph**, building memory-augmented agents with structured state transitions, fallback logic, and real-time observability across asynchronous task queues.
- **Deployed specialized agents** for quoting, customs, and invoicing—each capable of executing multi-step workflows such as load negotiation, customs document generation (ACE/ACI), and invoice issuance with payment tracking.
- **Built all backend services in Python** using Azure Functions and durable queues. Used Blob Storage for document management and tracked every agent action through a central log system I built myself.
- **Frontend:** Built a fast React interface showing real-time agent activity, confidence levels, and status flags. Dispatchers could override any agent at any time, but **less than 5% of tasks required intervention.**
- **Processed over 1,500 customs clearances and 1,000 invoices in production** with 95%+ accuracy. **Reduced dispatcher workload by 85%**, saving SMB carriers 20+ hours/week.
- **Acquired by MG Transport Ltd. in May 2025** after proving agentic AI could reliably run real cross-border dispatch.

Software Developer Intern **Jan 2023 – Dec 2023**

Geotab

- **Improved web-based route visualization** using polylines in C# and TypeScript, reducing map load times by upto 28% (7s → 5s).
- **Optimized vehicle telemetry storage algorithm,** reducing time complexity from O(n) to O(log n) through data structure redesign.
- **Wrote and maintained 25+ Selenium-based tests under TDD workflows,** improving regression coverage and deployment confidence.
- **Built backend APIs and telemetry data handlers** using C# and .NET Framework, powering real-time vehicle tracking features.
- **Collaborated across Agile teams** (developers, testers, PMs) to align backend engineering work with sprint deliverables and product goals.

Publications

Chaudhury, R., Huffman, C., Kwan, I., Deol, G. S., **Dhillon, S.**, & Chilana, P. K. (2025).

***MILESTONES: The Design and Field Evaluation of a Semi-Automated Tool for Promoting Self-Directed Learning Among Online Learners*.**
Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '25), ACM.
https://doi.org/10.1145/3706598.3714295

Contribution: Led AI development for categorizing learning resources and modeling user behavior to support adaptive feedback in self-directed learning systems.

Research Projects

PixIndexer — *LangChain, LangGraph, OpenAI, Gemini, On-device vector DB* **Nov 2023 – Mar 2024**

- **Built a privacy-first, AI-powered mobile app for semantic photo retrieval**, enabling users to search local image galleries using vague natural language prompts like "that party with blue lights" or "sunset trip last year".
- **Orchestrated a multi-agent pipeline with LangChain and LangGraph** to generate captions, extract object-context embeddings, perform semantic similarity ranking, and decompose fuzzy queries into structured search directives.
- **Leveraged OpenAI and Gemini models for LLM inference at ingestion time only**, transforming raw image metadata into semantic tags—then storing all indexed representations locally to enable real-time, offline retrieval.
- **Ensured all user queries and photos remained on-device post-indexing**, enforcing a strict no-cloud runtime policy and offering opt-in reprocessing for new media via LLMs if users enabled it.
- **Achieved sub-10ms vector retrieval** through incremental index updates in Realm, with dynamic memory management for low-power mobile environments.
- **Designed a local-first, edge-deployed AI architecture** with no cloud dependencies at query time, aligning with zero-trust principles in secure multimodal AI design.
- **Built a cross-platform React Native interface**, optimized for low-memory usage with real-time image preview, search refinement, and seamless UX across Android/iOS.
- **Executed as a solo, self-directed applied research project** exploring edge deployment of LLMs, LangGraph-based agent routing, and private semantic media search systems.

LegalEase — *LangChain, RAG, OpenAI + Gemini APIs, Python, Azure Functions* **Jun 2023 – Dec 2023**

- **Built a legal tech platform that delivers AI-generated answers to legal questions,** routed through a human-in-the-loop review pipeline to ensure factual and jurisdictional accuracy.
- **Engineered an LLM orchestration layer** that merged outputs from OpenAI and Gemini models using legal prompt chaining, retrieval-augmented generation (RAG), and fallback logic for ambiguous queries.
- **Implemented real-time legal query classification, context-aware prompt construction, and relevance scoring** to provide users with case-specific, well-reasoned responses across multiple legal domains.
- **Built an Angular frontend and Python backend** connected via Azure Functions, enabling session-persistent legal form submissions and scalable LLM-based response generation.
- **Architected a privacy-first data flow** with encrypted data transmission, field-level access control, and secure Azure SQL storage—ensuring legal query data remained fully confidential.

Cradle — *TypeScript, JavaScript, React, Python, Jest* **May 2022 – Aug 2022**

- **Co-developed a digital health platform for Uganda's clinical ecosystem,** enabling online doctor visits, health record storage, and patient management in low-connectivity regions.
- **Led frontend testing implementation with Jest,** achieving 90% test coverage and ensuring long-term stability of user-critical features.
- **Resolved 10+ accessibility and usability issues** across the React interface, improving task completion flow and boosting engagement by 20%.
- **Collaborated with a team of 5 students under the supervision of Dr. Brian Fraser (SFU),** applying Agile practices to deliver functional components for real-world deployment.

Academic Projects

Heart Disease Prediction Model — *Python, scikit-learn, Joblib* **Jan 2022 – Apr 2022**

- **Built a machine learning model** to predict heart disease using logistic regression and KNN; deployed using Joblib.
- **Performed end-to-end preprocessing and feature engineering** with Pandas, NumPy, and SciPy to maximize model accuracy.