

Jake Maly

jake.maly@mail.utoronto.ca | www.jakemaly.me | linkedin.com/in/jakemaly | github.com/jakemaly

EDUCATION

University of Toronto, St. George

Toronto, ON

Bachelor of Applied Science in Industrial Engineering, Minor in AI with PEY Co-op

Sep. 2025 – April 2029

- GPA: 3.6/4.0; Dean's Honours List (2x); Elliot M. Wilson Scholarship (\$5,000)

PROFESSIONAL EXPERIENCE

AI/ML Research Engineer

May 2026 – Present

University of Toronto

Toronto, ON

- Led a 3-person research team under Prof. Chi-Guhn Lee to design and evaluate a multi-agent AI system predicting Canadian criminal-appeal outcomes across 209 cases
- Reduced time to first token (TTFT) by ~800ms and eliminated all runtime embedding API calls by injecting active-case context directly, reserving BM25/vector RAG for static precedent
- Built a centralized evaluation and observability layer with Next.js and Langfuse, tracking accuracy, token cost, and latency across 100+ experiments
- Improved verdict-prediction accuracy from 40% to 67% through DSPy-based prompt optimization and A/B testing

Mathematics & Physics Tutor

Oct. 2022 – Present

Teenage Tutors & Toronto District School Board

Toronto, ON

- Guided students through difficult math & science concepts, encouraging critical thinking and problem-solving

Contract Software Engineer

EduTutor

May 2026 – Jul. 2026

- Built a full-stack quiz builder for an education platform, owning the user interface, versioned data design, and REST APIs (React, Vite, FastAPI, PostgreSQL); reduced quiz creation time from 15 minutes to 1 minute
- Engineered an asynchronous RAG validation pipeline for public question-database submissions, running formatting and factual-accuracy checks across 100% of ingested records

Blue Marble Academy

Feb. 2026 – Apr. 2026

- Built an AI-powered math & physics question generator using the Azure OpenAI SDK and Pydantic for an Ontario-based tutoring platform (React, FastAPI), including data validation & fallback handling

Applied Optimal

Jan. 2026 – Mar. 2026

- Collaborated on a 5-person team to build a B2B application for multiple client organizations (Flutter, Riverpod, FastAPI, Supabase); implemented role-based access control (RBAC) to protect restricted routes

Reframe Marketing

Nov. 2025 – Jan. 2026

- Built an AI document-analysis pipeline with vision-based extraction that creates Streamlit/Plotly reports, reducing analysis time for up to 50-page documents from 2+ hours to under 3 minutes

PROJECTS

Mementos, Deep Research & Notebook | *Python, LangGraph, LightRAG, Qdrant, Next.js, TypeScript*

- Built a research platform for parallel web research and searchable findings (LangGraph, GraphRAG, Qdrant); added a Next.js tracing interface and 3D TouchDesigner visualization

Socratic Python Tutor | *Gemma, Hugging Face Transformers, PEFT, LoRA*

- Built an AI Python tutor that guides learners without revealing answers; fine-tuned a 31B LLM using QLoRA and SFT across 400 examples to reduce data leakage from 8.33% to 0%

TECHNICAL SKILLS

Languages: Python, Java, JavaScript, TypeScript, SQL, Dart, HTML/CSS

Frameworks & Data: FastAPI, React, Next.js, Flutter, Streamlit, PostgreSQL, Azure Cosmos DB, Qdrant, Docker

AI/ML & Engineering: PyTorch, NumPy, Hugging Face Transformers, LangChain, LangGraph, LightRAG, DSPy, RAG, LLMs, PEFT/LoRA, SFT, Pydantic, MSAL, REST APIs, Git/GitHub, RBAC, Langfuse