Scott Lai

✉ scott.lai@duke.edu   📞 +1 352-339-8629   🖥 https://scottll.github.io/scottlai.github.io/| linkedin.com/in/scottlaiq/   🪪 Work Authorization: Green Card

EDUCATION ([Education Page](#))

Duke University | Durham, NC   *Aug 2022 – May 2024*
Master of Science in Interdisciplinary Data Science
- Relevant Courses: Statistical Modeling, Data Engineering, NLP, Data Analysis in Cloud, Practicing Machine Learning, Blockchain Development

University of Wisconsin | Madison, WI   *Aug 2016 – Dec 2018*
Bachelor of Science in Statistics and Economics
- Relevant Courses: Calculus, Linear Regression, Data Visualization, Machine Learning in Python, Statistics and Probability.

TECHNICAL SKILLS

- **Tech Skills:** Python, Rust, Node.js, HTML, CSS, JavaScript, R, MySQL, LangChain, LlamaIndex, GPT-api, Claude-api, Tableau, Scikit Learn, TensorFlow, PyTorch, Machine Learning (ML), Deep Learning (DL), Reinforcement Learning (RL), AB Testing. AWS, GCP, Azure, Git, Hugging Face, Bash, Spark, Hadoop, Docker, Kubernetes, beanstalk, OpenAI, Stable Diffusion, Claude 2, Midjourney, CodeWhisperer, Solidity, Web3.js, Truffle, Remix, Flask, Django.
- **Professional Certifications**: ESG Certification (in progress), AWS Technical Essentials Certification, AWS Cloud Practitioner Essentials Certification, AWS Cloud Foundations Certificate (in progress)
- **Language**: Mandarin (Native), English (Native)

EXPERIENCE ([Experience Page](#))

ScholarAI | Houston, Texas   *Oct 2023 – present*
- *Founding Software Engineer*
  - Architected and deployed a RAG-based document retrieval system using AWS DynamoDB and LangChain, improving search accuracy by 35%
  - Developed a full-stack web application using React, FastAPI, and PostgreSQL to handle 10,000+ daily patent processing requests
  - Engineered a scalable ETL pipeline processing 2TB+ of patent data daily using Apache Spark and Airflow
  - Built an AI-driven OCR system with 98% accuracy for complex technical documents
  - Trained and optimized an AI model for patent claim generation, achieving 17% performance improvement over ChatGPT and Claude 3.5
  - Implemented a distributed multi-agent system architecture serving 500+ concurrent users

The Classroom Company | Durham, NC   *Oct 2023 – Oct 2024*
- *Co-founder & Chief Development Officer & Artificial Intelligence Engineer*
  - Developing cloud-based LLMs architecture for artificial intelligence applications in education industry.

Duke University Graduate School | Durham, NC   *May 2023 – May 2024*
- *Teaching & Research Assistant*
  - **Rust Candle LLM Development:** Collaborating with HuggingFace on the development and deployment large language models (LLMs) using Rust Candle.
  - **Cloud Research and Development:** Managing relationships with AWS and Azure, and actively participating in their cloud research and development projects.
  - **Course Management & Content Design:** Managed courses on Microsoft Teams, designed weekly modules aligned with GitHub syllabus, graded assignments, and conducted sync meetings with professors. Offered weekly office hours for student support.
  - **AWS Cloud Club Captain:** Coordinated with AWS and Duke to manage the Cloud Club, facilitating access to cloud resources and certifications for over 200 Duke students.

Duke University Fuqua Business School | Durham, NC   *Sep 2022 – May 2024*
- *Research Assistant*
  - Assist professor Campbell R. Harvey focuses on the Blockchain research including IPFS interacting with NFTs, stable coin and CBDC.
  - Developed Blockchain research on file storage.

Spigot, Inc. | Virtual / Fort Myers, Florida   *Nov 2022 – Aug 2023*
- *Data Scientist & AI Developer*
  - Developing AI tools to support various aspects of the company's operations.
  - Conducting comprehensive research on Big Data to extract valuable insights and trends.
  - Engaging in research and development activities related to Blockchain technology.
  - Collaborating on strategic projects within the FinTech and Blockchain domain to help achieve business objectives.

PROJECTS ([Project Page](#))

ScholarAI education webapp (https://scholarai.io/)   **Oct. 2023**
Tech: LLMs, Python, Docker, LangChain, LlamaIndex, Claude, AutoGen, GenAI, Nodejs, etc.
- Simplifies academic, scientific, and legal research by providing instant answers, insights, and analysis from papers, patents, and more—all powered by cutting-edge AI.

Ari AI Law webapp (https://www.ari.law/#lp-pom-block-32) **May. 2024**
Tech: LLMs, Python, Docker, LangChain, LlamaIndex, Claude, AutoGen, GenAI, Nodejs, USPTO, TensorFlow, PyTorch, etc.
Better strategic decision making based on AI insights compiled across millions of patents and scientific papers.

NeuroComm (https://colab.duke.edu/project/neurocomm-bridging-eeg-and-ai-cognitive-insight/) **Jan. 2024**
Tech: LLM, LLLM, EEG, Python
- Bridging EEG and AI for Cognitive Insight

Candle Cookbook (https://nogibjj.github.io/candle-cookbook/) **Oct. 2023**
Tech: LLM, Azure, AWS, GitHub Action, Jenkins, Rust
- Using cloud service to deploy the GPU LLM binary to making large language models more accessible and open-source.

Personal AI Avatar Assistant (https://scottai.azurewebsites.net/) **July 2023**
Tech: Machine Learning, AWS ECR, Azure App Service, OpenAI, Hugging Face
- Created an innovative personal AI avatar assistant designed to provide users with an enhanced interactive experience.