Yuming Dai
Senior Senior Machine Learning Engineer

SUMMARY

Experienced Senior Machine Learning Engineer with over 8 years in AI research and system architecture, developing and scaling AI systems and machine learning models for use in retail, conversational AI, and writing assistance platforms. Distinguished for leading multidisciplinary teams and aligning AI with business strategies, achieving substantial performance improvements and cost savings. Proficient in AI/ML (e.g., deep learning, reinforcement learning, computer vision), data pipeline design, and MLOps (e.g., Kubernetes, Docker), with a strong background in full-stack engineering (Python, C++, Go, Node.js, React) and cloud platforms (AWS, GCP, Azure). Holds a Master's degree in Artificial Intelligence from New York University, complemented by a Bachelor's in Computer Science, validating comprehensive expertise in algorithmic research, technical publications, and significant contributions to domain-specific AI applications.

TECHNICAL SKILLS

Main Technical Skills	Generative AI, Azure VM
Programming Languages	Go, Python, TypeScript
AI & Machine Learning	Amazon Machine learning services, AWS ML (Amazon Machine learning services), AWS SageMaker, AWS SageMaker (Amazon SageMaker), Kubeflow, Mlflow, PyTorch, TensorFlow, Vertex AI
Java Frameworks	Apache Spark
Scala Frameworks	Apache Spark
C++ Libraries and Tools	C/C++/C#
JavaScript Frameworks	Node.js, React
Python Libraries and Tools	PyTorch, TensorFlow
Data Analysis and Visualization Technologies	Apache Airflow, Apache Spark, ETL Pipelines
Databases & Management Systems / ORM	Apache Hadoop, Apache Spark, NoSQL

Cloud Platforms, Services & Computing	Azure ML
Amazon Web Services	AWS ML (Amazon Machine learning services), AWS SageMaker, AWS SageMaker (Amazon SageMaker)
Azure Cloud Services	Azure VM
Google Cloud Platform	Google BigQuery
Methodologies, Paradigms and Patterns	Clean Architecture, microservices
Logging and Monitoring	Datadog, Prometheus
Virtualization, Containers and Orchestration	Docker, Kubernetes
SDK / API and Integrations	GraphQL
Deployment, CI/CD & Administration	Helm
Message/Queue/Task Brokers	Kafka
Other Technical Skills	GDPR, Google PageSpeed Insights, Horovod, Legacy Application, REST & gRPC API, SOC 2

WORK EXPERIENCE

1. **Company** : Mashgin
   Title : Senior Machine Learning Engineer
   Project : Autonomous Self-Checkout Kiosk Intelligence
   Duration : 11/2024 – Present
   Summary : Contributed to the transformation of Mashgin's self-checkout kiosks by creating an autonomous cognitive 'agent', which significantly reduced exceptional transactions and improved checkout throughput and sales.
   Responsibilities : Architecting AI system, engineering multi-modal agents, deploying continuous-learning pipelines, leading 'agent lifecycle' definition, delivering business outcomes, orchestrating integrations, mentoring the engineering team, steering research to production, and ensuring the agent architecture remained cutting edge.
   Technologies : 3D point-clouds, computer vision, policy modules, microservices architecture, edge-model deployments, reinforcement learning, CI/CD, telemetry, A/B testing

- **Company** : EliseAI
  Title : Machine Learning Engineer
  Project : Agentic Conversational AI System
  Duration: 10/2023 – 10/2024
  Summary: Led the development and deployment of EliseAI's high-throughput agentic conversational AI system for housing and healthcare clients, achieving 24/7 operations and increased conversion rates.
  Responsibilities: Architecting AI platforms, building intent-recognition engines, fine-tuning LLM pipelines, integrating with CRM/PMS systems, delivering A/B testing frameworks, launching product features, building predictive modeling pipelines,

collaborating across teams, mentoring engineers, and driving a vertical-first AI strategy.
Technologies: Python, microservices, natural language processing, LLM, Snowflake, Amazon S3, ensemble and deep-learning models, voice/NLP libraries, CI/CD

- **Company** : Grammarly
  Title : AI Researcher
  Project : Advanced Writing Assistance EngineComp
  Duration: 10/2020 – 10/2023
  Summary: Spearheaded the enhancement of Grammarly's writing-assistance engine, advancing the capabilities in grammar, style, tone, and correction with impactful features such as 'compose' and 'rewrite'.
  Responsibilities: Developing deep-learning models, crafting a hybrid architecture with classic NLP and transformer-based models, launching generative-AI features, constructing datasets for various writing tasks, and collaborating with teams to create a responsible AI framework.
  Technologies: Deep learning, NLP, transformer-based encoders/decoders, generative AI, dataset construction, real-time inference stacks, bias-monitoring
- **Company** : Lily AI
  Title : Machine Learning Engineer | Full Stack
  Project : Product-Content Enrichment Platform
  Duration: 09/2017 – 09/2020
  Summary: Led the full-stack development of Lily AI's product-content enrichment platform, which provided deep-learning powered attribute extraction, significantly enhancing clients' e-commerce personalization and conversion.
  Responsibilities: Creating the architecture for ML pipelines and full-stack services, deploying product-attribute extraction engines, designing backend services and front-end dashboards, integrating with retailer systems, and instituting MLOps practices.
  Technologies: Python, Flask, Docker, Kubernetes, React, Redux, ResNet, Inception, transformer-based text models, microservices, backend and frontend development

EDUCATION

- **Master of Artificial Intelligence**
  New York University
  09/2015 – 08/2017
- **Bachelor of Computer Science**
  New York University
  09/2011 – 08/2015