Yu Gu

- aidengu001@gmail.com
- +1-858-342-9893
- LinkedIn

Research & Applied Scientist specializing in large-scale foundation models, multimodal learning, and enterprise AI deployment. Led the development of PubMedBERT (20M+ downloads, 2000+ citations, *ACM HLTH Best Paper of the Year)*, BiomedParse (*Nature Methods*), and BiomedJourney (text-to-image generation for medical AI). Developed retrieval-augmented generation (RAG) and multi-agent AI systems for enterprise applications. Co-founded an AI startup acquired in Series C. Published in Nature, Cell, ICLR, ACL; reviewer for NeurIPS, ICML, and Nature journals. Featured in *Forbes, CNBC*, and the *World Economic Forum*.

Professional Experience

Senior Applied Scientist Microsoft **Jan 2020 – Present**

Building Scalable, Domain-Specific Models & Enterprise-Grade Systems

- **PubMedBERT**: Developed one of the first **domain-adaptive LLMs**, addressing tokenization and terminology gaps where general-purpose models fail. Now the **core of Azure's Text Analytics for Health**, enabling real-time entity extraction, relation detection, and medical coding across **10+ major institutions**. **20M+ downloads, 100+ enterprise adoptions**.
- **BiomedParse**: Built a **universal segmentation** foundation model for multimodal imaging (CT, MRI, pathology), solving deployment challenges across organizations. **Deployed in three major institutions**, with **50K+ monthly downloads** and growing industry adoption.
- **UniversalNER**: Designed a **scalable named entity recognition** framework, leveraging **LLM distillation and retrieval-enhanced learning**, achieving **13% higher accuracy** than leading industrial solutions. Now powering de-identification and automated data processing pipelines in large-scale first-party deployments.
- **BiomedJourney**: Developed a **text-to-image generation model**, enabling **realistic medical image synthesis** with nuanced control over disease progression. Used across projects for **data augmentation, model training, and enterprise demos**, addressing **scarce-data** challenges in AI model development.

Scalable Deployment, Retrieval & Efficient Model Inference

- Developed **multi-agent** systems leveraging specialist models for complex sub-tasks, addressing tumor board decision-making for challenging cancer treatment cases. Demoed at the **World Economic Forum**, showcasing LLM-driven collaborative reasoning.
- Led **cross-functional teams** (scientists, engineers, PMs, designers) to build and deploy scalable pipelines and cloud infrastructure, delivering **four enterprise-grade models** optimized for **production deployment and seamless public adoption.**
- Optimized LLM inference with FAISS-based retrieval, reducing computational overhead and improving real-time search efficiency. Accelerated model deployment with ONNX, achieving 35% faster inference speeds while maintaining high accuracy in production environments.

Machine Learning Scientist/Co-founder Med Data Quest Inc. **Jun 2017 – Jan 2020**

- **Co-founded a startup**, leading to a **Series C acquisition**, developing AI-driven solutions that transitioned from research to enterprise adoption. Led a 10+ person international team in designing a triaging and hierarchical AI system, powering an assistant annotation platform for domain experts.
- Developed full-stack NLP pipelines, securing **top-5** rankings in N2C2 Challenges (2016, 2018, 2019)

Education

Peking University

B.Sc in Microelectronics
B.Sc in Economics

Selected Publications (Scholar)

- *PubMedBERT: Domain-Specific Language Model Pretraining. ACM Health*(Best Paper of the Year 2022)
- *BiomedParse: Image Parsing for Everything, Everywhere. Nature Methods*
- *UniversalNER: LLM Distillation for Open NER. ICLR*
- *BiomedJourney: Temporal AI for Patient Outcome Simulation. arXiv*
- *GigaPath: A Whole-Slide Foundation Model. Nature*