Pengyuan Li, Ph.D.

Research Staff Member, MIT-IBM Watson AI Lab
Adjunct Faculty, Data Science Institute, University of Delaware

☎ *(302)-501-0291*
✉ *pengyuan@ibm.com*
in *pengyuanli*

Professional Summary

AI researcher with end-to-end experience across the full AI development lifecycle — from large-scale data acquisition, to training foundation models, and deploying real-world applications. Lead developer of **Granite Vision Models**, with deep expertise in multimodal learning, scalable ML systems, and enterprise-grade AI solutions. Proven impact through open-source releases, publications, patents, and organizational awards at IBM Research.

Selected Highlights

Model Impact — **Granite Vision Models**: Lead developer; 200k+ usage on HF, 250k+ usage on Ollama, 1M+ usage on IBM platforms

Data Collection — Data acquisition lead; collected 10+ PB of text, code, and multimodal data; contributed to **Granite language models** and **Granite code models**

Awards — IBM **O-Level** (2024), **Special** (2023), and **A-Level** (2022) Accomplishment awards

Mentorship — Mentored 5 PhD interns; Adjunct Faculty at Univ. of Delaware (BINF601); Capstone advisor (UCSC-IBM HCI271)

Leadership — Organizer of the ”**AI and Biodata Resources**” workshop at Biocuration2025

Industry Experience

2021–Present **Research Staff Member**, *IBM Research*

- Lead developer of Granite Vision Models; built and contributed to every stage of development — from data collection and filtering to model training, onboarding, and customer integration
- Data acquisition lead for developing large language, code, and multimodal models; collected 10PB of data (**IBM Special Accomplishment 2023, O-Level Accomplishment 2024**)
- Led large-scale GitHub data mining, including repositories, pull requests, and issues for code model training
- Developed pipelines for parsing and extracting PDF documents to create high-quality datasets
- Built a search engine to match business client requirements to IBM solutions (**IBM A-Level Accomplishment 2022**)
- Conducted business document analysis for information extraction and downstream applications

Summer 2022, 2024 & 2025 **Intern Mentor**, *IBM Research*

- Mentored 5 PhD interns on research topics in machine learning, LLMs, and scientific document understanding
- Co-developed innovative project ideas, leading to publications and patent filings

Spring & Fall 2023 **Research Advisor**, *UCSC-IBM HCI271 Capstone*

- Advised student teams building an LLM training platform
- Provided technical feedback on backend training workflows, UX, and evaluation methods

Jun–Aug 2019 **Research Intern**, *IBM Research*

- Performed NLP analysis and topic modeling on customer reviews to identify service insights

Academic Experience

2023–Present **Collaborator**, *Sternberg Lab, Caltech*

- Image manipulation detection in scientific literature, collaborated with the microPublication Journal (www.micropublication.org)
- Leveraging LLM for accelerating gene summarization work at Alliance of Genome Resources (www.alliancegenome.org)

2023–Present **Adjunct Faculty**, *University of Delaware, Data Science Institute*

- Class design for BINF601: Introduction to Data Sciences
- Delivered lectures on biomedical image analysis and data science fundamentals

2015–2021 **Research Assistant**, *University of Delaware*

- Biomedical document classification utilizing image and text information
- Figure and caption extraction from scientific documents (FDFigCapX)
- Compound image separation of published figures (FigSplit)
- Biomedical image classification for supporting the bio-image annotation process
- Heart disease detection using ECG signals and ultrasound images

2018 **Visiting Student**, *University of British Columbia*

2011–2015 **Research Assistant**, *Harbin Engineering University*

2013 **Visiting Student**, *UCLA School of Medicine*

2012 **Visiting Student**, *Tongji University*

2009–2010 **Lab Member**, *ACM-ICPC Lab, Zhengzhou University*

Awards & Honors

2024 Corporate O-level Accomplishment, IBM Research

2023 Corporate Special Accomplishment, IBM Research

2022 Corporate A-level Accomplishment, IBM Research

2021 Frank A. Pehrson Graduate Student Award, University of Delaware

2020 Distinguished Graduate Student Award, Dissertation Fellowship

2013 National Scholarship for Graduate Students, China

2009 Silver Medal, ACM-ICPC Henan Province

Service & Activities

Invited Talk *Granite Vision Models*, OpenCV Live Talk, 2025

Conference Reviewer EMNLP 2025 (Area Chair), ACL 2025, ICML 2025, NeurIPS 2024–2025, SIGKDD 2023–2024, SIGIR 2024, WWW 2022–2025, BIBM 2020–2025 (Session Chair)

Conference Organizer *AI and Biodata Resources Workshop*, Biocuration 2025

Organizing Committee IBM Research - Almaden Spirit Team – Academic talks, social events, return-to-work activities

Selected Publications

[1] *Granite Vision: a lightweight, open-source multimodal model for enterprise Intelligence*. arXiv:2502.09927, 2025

[2] *Decay Pruning Method: Smooth Pruning With a Self-Rectifying Procedure*. ICCV 2025

[3] *Granite Code Models: A Family of Open Foundation Models for Code Intelligence*. arXiv:2405.04324, 2024

[4] *Don't be my Doctor! Recognizing Healthcare Advice in Large Language Models*. EMNLP 2024

[5] *Long-form information retrieval for enterprise matchmaking*. ACM SIGIR 2023

[6] *Utilizing image and caption information for biomedical document classification*. ISMB/ECCB 2021; also in Bioinformatics, 2021

[7] *Extracting figures and captions from biomedical documents*. Bioinformatics, 2019

[8] *Compound image segmentation of published biomedical figures*. Bioinformatics, 2018

[9] *Segmenting compound biomedical figures into their constituent panels*. CLEF 2017 (Best of Lab Paper)

[10] *Brain CT image similarity retrieval method based on Uncertain Location Graph*. IEEE JBHI, 2014

...Full list available upon request or via Google Scholar.

Patents

[1] Generation of graphical icons for taxonomy nodes. (Filed)

[2] Generating diagrams for visualizing structured documents. (Filed)

[3] Navigation guide using different vehicle components. (Filed)

[4] Medical Image Similarity Retrieval Based on Uncertain Location Graph (CN103226582A)

Education

2015–2021 **Ph.D., Computer Science**, *University of Delaware*
Advisor: Prof. Hagit Shatkay
Dissertation: Utilizing Image Information for Biomedical Document Classification

2011–2014 **M.E., Computer Software and Theory**, *Harbin Engineering University*
Advisor: Prof. Haiwei Pan
Dissertation: Medical Image Retrieval Based on Uncertain Location Graph

2007–2011 **B.E., Computer Science and Technology**, *Zhengzhou University*