Dongwhi Kim

Email • Website • LinkedIn • Code • Publications

Education

University of Notre Dame — Notre Dame, IN
B.S. Computer Science & Applied Mathematics (double major) — 2022 - 2026
Advisors: Prof. Nuno Moniz, Prof. Meng Jiang — **GPA**: 3.97

Relevant Courses

All A: Neural Networks (graduate) • Maching Learning for Embedded Systems (graduate) • Machine Learning for Engineers • Intro to AI • Theory of Computing • Advanced Databases • Databases • Operating Systems • Computer Architecture • Datastructures • Algorithms • Linear Algebra & Differential Equations • Calculus I, II, III • Discrete Mathematics • Probability and Statistics

Research Interests

Trustworthy AI: (Multimodal) Language Model Safety • Mechanistic Interpretability • Machine Unlearning
Natural Language Processing: Natural Language Programs • Evaluation of Foundation Models
Human-Computer Interaction: Human-AI Interaction • Agentic Systems • AI tools for Research

Publications

1. **Kim, Dongwhi***, Liu, Zheyuan*, Wan, Y., Yuan, X., Tan, Z., Mo, F., Jiang, M. MTMCS-Bench: Evaluating Contextual Safety of Multimodal Large Language Models in Multi-Turn Dialogues. In submission to *ACL* (2026)

2. Nogueira, Brenda, Geyer, W., Anderson, A., Li, T., **Kim, D.**, Moniz, N., Chawla, N.. From Verification Burden to Trusted Collaboration: Design Goals for LLM-Assisted Literature Reviews. In *AAAI Workshop on AI for Scientific Research* (2026)

3. **Kim, Dongwhi***, O'Roark, Katherine*, Liu, Z., Jiang, M. Probe-Guided Machine Unlearning with Adversarial-Robust Representation Steering. Poster presentation in *RISE AI* (2025)

4. **Kim, Dongwhi**, Moniz, N. Relevance-Aware Algorithmic Recourse. In *IDA* (2025)

* denotes equal contribution

Research Experience

Multimodal Contextual Safety Benchmark — University of Notre Dame
Research Assistant, advised by Prof. Meng Jiang — July 2025—Present

- Designed a benchmark with 30,080 multi-turn, image-grounded dialogue to evaluate safety and refusal alignment in MLLMs.
- Created two types of contextual harm, one that builds up and another that carries harm discretely from the original query.
- Built multi-agentic workflows in DSPy to generate the dialogues for the benchmark, including a generator and judge program.
- Optimized generator LLM dialogues through GEPA (prompt optimization) from 62% to 92% accuracy on generation metrics.

Design Goals for LLM-Assisted Literature Reviews — University of Notre Dame
Research Assistant, advised by Prof. Nuno Moniz — Aug 2025—Present

- Improved research paper analysis system with structured extraction of key insights, methods, and results from with citation tracking and contextual chat interface for research exploration.
- Helped build end-to-end research workflow automation tool to create literature reviews on research papers.

Probe-Guided Machine Unlearning with Adversarial-Robust Representation Steering — University of Notre Dame
Research Assistant, advised by Prof. Meng Jiang — Nov 2024—Present

- Developed a more robust two-stage machine unlearning pipeline including probe training and LLM unlearning so harmful content is internally indistinguishable from harmless for safe or confusing outputs.
- Trained adversarially-robust MLP probe with exponentially moving average prototypes, FGSM hardening, and geometric margin constraints for robustness.

- Poster presentation at RISE AI 2025.

Relevance-Aware Algorithmic Recourse University of Notre Dame

Research Assistant, advised by Prof. Nuno Moniz Aug 2023—May 2025

- Introduced the first framework to provide counterfactual explanations in regression settings on tree-ensemble models through Bayesian optimization on relevance functions mapping user preferences to a continuous scale.
- Our approach reduced iterations up to 8.38% and reduced costs up to 2.17% when compared to our baseline without a relevance function across 15 regression datasets.
- Published and presented work at IDA 2025 in Konstanz, Germany.

Cancer Detection Grid Search University of Notre Dame

Research Assistant, advised by Prof. Nuno Moniz Aug 2022—May 2023

- Pre-processed data for Random Forest, Logistic Regression, and XGBoost in imbalanced data domains of cancer datasets.
- Benchmarked 10 resampling strategies including SMOTE, random over/undersampling, and hybrid methods on imbalanced lung cancer datasets, where random under and oversampling had the best ROC, AUC, and F1-scores.
- Increased model metrics through hyperparameter tuning using grid search to optimize model performance.

Professional Experience

Amazon (SageMaker) May 2025—Aug 2025

AI/ML SDE Intern Palo Alto, CA

- Built admin and data AI agents in LangGraph with MCP/A2A protocols, long-term memory, and human-in-the-loop
- Reduced customer service time by 80% with admin agent and translated natural language to SQL with data agent
- Provisioned zero-downtime agent infrastructure by AWS CDK on ECS/Fargate clusters with auto-scaling/load balancers
- Synced agent memory with serverless Aurora PostgreSQL database and stored SOPs in a Graph RAG knowledge base
- Organized the 1st Bay Area AWS intern hackathon for 200 interns, raising $3,000 in sponsorships and enlisted 8 judges

Humana May 2024—Aug 2024

Data Analyst Intern Louisville, KY

- Eliminated 2,500 redundant mobile lines, saving $2.4MM annually from Power Apps and Power BI dashboards insights
- Updated internal LLM chatbot in C#/ASP.net, receiving 20,000 daily queries, and increasing user productivity by 15%

Alora Finance Startup Sept 2023—May 2024

Lead SWE Notre Dame, IN

- Launched financial education platform used by 800 students and teachers to introduce financial education in schools
- Led nine undergrads, taught Git/agile development, coded in React, Django REST API, SQL, tested APIs in Postman

Projects

DSPy (Arbor) Notre Dame, IN

Open-source contributor for RL optimizations Arbor library (174 GitHub Stars) May 2025—Present

- Built NVIDIA-GPU monitoring tools for training, and support for single GPU use for vLLM for reinforcement learning
- Added CLI/Python module for config management, API startup, example scripts, and JQlang query logging for debug

Handwritten Math Equation Transcriber Notre Dame, IN

Neural Networks Final Project Feb 2025—May 2025

- Designed bidirectional LSTM encoder-decoder architecture with multi-head attention mechanism for seq2seq learning
- Pre-processed equations to LaTeX, then corrected with LLM using in-context learning with 92% validation accuracy

Food Tinder (1st Place Advanced Databases Final Project) Notre Dame, IN

1st Place Advanced Databases Final Project ($1050) Feb 2025—May 2025

- Architected React Native app with Expo/Redux, Python (FastAPI), Oracle database, AWS EC2 hosting and S3 buckets
- Coded collaborative filtering recommendation system using PL/SQL procedures, indexes, views and job scheduler

Trip It! Notre Dame, IN

Databases Final Project Sep 2024—Dec 2024

- Built entire backend with Flask, PostgreSQL (pgvector for embeddings), Redis/Celery for task scheduling, AWS hosted
- Generated 10k synthetic trips, automated clustering/vectorizer model updates, designed content-based recommender

ML Pickup Line Generator and Classifier London, UK

London Study Abroad Project May 2023—June 2023

- Implemented text-generating recurrent neural network in Python using TensorFlow with 1.71 test loss for pickup lines using gated recurrent units, batch normalization, dropout, and L2 regularization.
- Presented and articulated technical challenges and solution to over 60 students and interested professors.

Leadership Experience

AI/ML Project Club Aug 2025—Present

Founder Notre Dame, IN

- Launched project series with 10 projects and 52 students in its first year for hands on ML experience.
- Held invited speaker sessions, office hours, and competition for end of year.

Theory of Computing Course Aug 2024—Dec 2024

Teaching Assistant Notre Dame, IN

- Held weekly office hours for 150 students on topics like computational theory from Turing machines to NP-completeness.
- Simplified complex concepts, provided feedback and graded assignments, and guided students on semester-long projects.

LikeLion US July 2023—May 2024

Co-Founder / Co-President Notre Dame, IN

- Mentored students with no background in coding to build successful software engineering careers
- Guided students to launch competitive startup businesses using programming to solve real-world issues

Student International Business Council Sept 2022—May 2024

Amazon Project Leader / Microsoft + Uber + PWC Project Analyst Notre Dame, IN

- Spearheaded weekly meetings with ten students and company contacts to devise innovative business solutions
- Tackled ambiguous problems through abstract thinking, implementation timelines, and risk assessments
- Developed BNPL feature for Xbox game purchases with implementation timeline and risk analysis
- Proposed optimal QR code solution for truckers to identify shipments, improving efficiency and reducing detention time

Asian American Association (AAA) Sept 2022—May 2024

Freshman Representative / Treasurer / Audio Director Notre Dame, IN

- Managed organization's finances and budget of $10,000+, maintaining accurate financial records
- Led groups, performed speeches, and organized activities for 100+ students at events including Winter Retreat

Awards and Honors

1st place Advanced Databases Projects Oracle Sponsor ($1,050)	2025
1st place IBM Hackathon ($3,000)	2024
iTREDs Scholar (University of Notre Dame)	2024—present
Sorin Scholar (University of Notre Dame)	2023—present
Deans List (University of Notre Dame)	2022—present