DONGEUN KIM

Seoul, South Korea • dongeunk@umich.edu • github.com/dkim1112 • linkedin.com/in/kimdongeun

EDUCATION

University of Michigan – Ann Arbor *Ann Arbor, MI | Expected May 2030*

B.S. Data Science (Major) & Statistics (Minor)

Relevant Courses: *Program & Data Structures, Discrete Math, Elementary Programming Concepts, Intro to Data Science, Multivariable Calculus*

EXPERIENCE

Undergraduate Researcher | University of Michigan (Medical School) *June 2025 - Present*

- Constructed variant × phenotype beta matrix from PheWeb GWAS inputs across 1M+ variants × 1,400+ disease endpoints, handling missing values and allele alignment over 38,000+ records, to surface cross-disease genetic architecture systematically missed by single-trait analysis
- Applied truncated SVD (k = 50) via scikit-learn to decompose association matrix, retaining 87% of variance in latent biological axes
- Validated embedding space against 200+ known comorbidity pairs, enabling multi-phenotype prioritization without trait-specific retraining

Undergraduate Researcher | University of Michigan (Cutaneous Lab) *Sep. 2024 – Dec. 2025*

- Addressed dermatologists' bottleneck navigating large, fragmented literature collections where existing LLM tools hallucinate and lack source traceability, developing a multi-stage RAG-based alternative enabling 10+ clinicians to cut literature review time by ~75%
- Devised pipeline via hierarchical chunking, EnsembleRetriever, CrossEncoder Reranking, and MultiQueryRetriever in LangChain, reducing memory overhead by 97% and query latency vs. vector similarity baseline when tested across 55+ dermatology manuscripts
- Published preprint on bioRxiv (https://www.biorxiv.org/content/10.1101/2025.08.14.670384v1); presented abstract at KSID X ISID APAC 2026

Data Engineer Intern | Maetel *May 2025 - Sep. 2025*

- Spearheaded AI-native copywriting (Syndy.ai) infrastructure, integrating LLM APIs and few-shot prompt engineering, boosting stylistic accuracy by ~90% confirmed via user surveys; shipped on Product Hunt, ranking 5th among unfeatured products, driving 500+ signups in 48 hours
- Mined behavioral data from 3000+ user sessions within conversion funnels to identify onboarding and note-creation friction points; pinpointed features most correlated to retention, driving a 22% uplift in activation-to-retention conversion and directly shaping product prioritization
- Engineered Persona Chat data pipeline – 3-layer document ingestion architecture, auto-diff logic uncovering schema-level persona changes, and structured JSON schemas (persona_state, outreach_strategy, content_strategy) – eliminating manual B2B customer profiling
- Automated LinkedIn lead qualification after identifying inefficiency in outreach targeting, building a Lead Scoring Engine with structured LLM prompting to rank Sales Navigator prospects against ICP criteria; validated with BD team, improving response rate by 60%

Computer Vision Engineer | UM::Autonomy *Sep. 2024 - June 2025*

- Calibrated LiDAR using ArUco markers and linear transformations, compensating ±10cm inter-sensor displacement with 95% spatial precision
- Trained YOLO-based object detection model on 2,000+ Blender-simulated images via 3D data augmentation, placing 2nd at RoboBoat 2025

PROJECTS

CDC Event Pipeline | github.com/dkim1112/CDC-Event-Pipeline

- Realized intermediate order states vanish between poll cycles; architected WAL-based CDC pipeline capturing transition with 0 source DB load
- Orchestrated idempotent event processing via UUID5 deduplication and UPSERT snapshots, sustaining 0 duplicates over 533-event bursts
- Introduced dead letter queue, per-table out-of-order detection, end-to-end lag monitoring; verified core processing logic across 21 tests

Price Monitor Pipeline | github.com/dkim1112/Price-Monitor-Pipeline

- Observed Korean grocery prices diverging from official CPI reporting, building a raw-to-mart ETL pipeline to quantify the gap by integrating KOSTAT product-level prices and ECOS CPI indices, ingesting ~600K product listings per-run across 124 item categories
- Developed 6 pre-aggregation quality gates: recalibrated CPI anomaly threshold after 143 false positives, eliminating all erroneous alerts
- Designed idempotent mart layer with UPSERT, monthly-partitioned raw tables, and cron-scheduled collection with adaptive date probing

UMich KISA Official Website | www.umichkisa.com

- Integrated Next.js frontend with Python Flask backend; deployed on Vercel, supporting 100+ daily users spanning 600+ Korean students
- Built MySQL bulletin board (post, comment, like) and secured authentication via Google OAuth, increasing user engagement
- Redesigned UI in Figma and implemented with TailwindCSS, achieving 85% user satisfaction during a 3-week user survey

SKILLS

Languages & Frameworks: Python, C/C++, Java, JavaScript, TypeScript, HTML, CSS, SQL, React, Next.js, Node.js, Flask, TailwindCSS

Tools & Technologies: AWS, Kafka, Debezium, WebSocket, Git, Matplotlib, LangChain, LLM, RAG, Streamlit, Excel, LaTeX