# TAEMIN KANG

taemin.kang@gmail.com · +1 (814) 308-3329 · linkedin.com/in/taemin-kang · github.com/2002tae · taeminkang.com · artemiscareer.com

## EDUCATION

**Pennsylvania State University** University Park, PA *Expected May 2027*

B.S. Computational Data Science & Statistics (Double Major) · Minor in Mathematics · Major GPA 3.54 / 4.0

**Coursework:** Computer Structures and Algorithms · Big Data (MapReduce, Spark) · Database Systems · Machine Learning · Applied Data Mining (graduate) · Linear Algebra

## EXPERIENCE

**Solo Developer & Product Owner** Artemis — AI Career Platform *Mar 2026 – Present*

- Engineered a Postgres-native async ingestion pipeline (Procrastinate cron) that seeds ~830 company boards nightly across the major ATS platforms, normalizing heterogeneous schemas into a shared index of 300K+ postings.
- Designed a multi-tier cache (normalized URL → ATS API → metered LLM) with SHA-256 content-hash deduplication and cross-source identity merge, so duplicate records resolve on a $0 path instead of re-triggering paid extraction.
- Built record-lifecycle hygiene over that index (active/expired classification, dead-link and garbage-record detection, periodic revalidation) with a final liveness check before deletion; in a sample of 20 deletion candidates, 16 were still live.
- Built the backend on FastAPI, SQLModel and PostgreSQL 17 with row-level security on every table and contract-first Pydantic → OpenAPI → TypeScript codegen gated in CI, eliminating client/schema drift.
- Integrated Google and Microsoft OAuth with Gmail/Outlook and Calendar via Cloud Pub/Sub push receivers and Microsoft Graph webhooks, with a scope-isolated encrypted token store and automatic subscription renewal.
- Self-host Qwen3.5-9B (4-bit AWQ, vLLM) on a single 16 GB GPU for nightly batch parsing: 318K+ calls at no marginal cost versus an estimated $312–513 on a commercial API, with per-call cost metering, hash-versioned prompts, and a CI gate on prompt changes.
- Cut database size by a third (6,231 → 4,246 MB) by content-addressing 74 system prompts that had been stored 333,633 times.

**Intelligence Analyst** Korea Defense Intelligence Command *Dec 2023 – Jun 2025*

- Built a Python NLP pipeline that classified 150+ unstructured intelligence reports, cutting manual analysis time 60%.
- Developed an anomaly-detection system flagging high-priority events at 85% precision.

## PROJECTS

**Project Eureka — Korean-Market Data Infrastructure** solo *Feb 2026 – Present*

- Architected and operate a 78 GB PostgreSQL + TimescaleDB system (~160M rows across 89 tables and 18 hypertables) fed by 26 scheduled ingestion jobs over REST and SDMX APIs and three anti-bot crawlers, with a dead-letter retry queue and a ~30-second intraday event daemon.
- Implemented bitemporal point-in-time versioning (vintage-dated, UPSERT-only history) and an amendment-supersession resolver over 135K record versions that recovers each filing's original publication date, so a later amendment no longer masks when information first became public.
- Maintain a daily-reconciled slowly-changing entity dimension: ~2,800 active entities out of 3,900+ historical, with about 50 state transitions per day across an eight-state status model, rebuilt daily rather than snapshotted so retired entities keep their full history.
- Operate self-healing multi-day backfills (file lock, checkpoint, resume) that reconstructed 342K rows spanning 2008–2023 with no manual intervention, plus daily automated health and coverage alerting.
- Built an event-level corporate-action ledger after finding the vendor's "adjusted" flag was a constant default: market-cap discontinuities above 2× fell from 455/642 (raw) to 4/566, each event measured two independent ways.
- Run standing data-quality automation: daily cross-table reconciliation, staleness detection, and coverage-gap and dimension-diff audits.
- Deployed KR-FinBERT GPU inference over 929K Korean filings and news documents (BFloat16, torch.compile, CUDA-OOM-adaptive batching) and implemented 121 registered features across two layers.
- Built a read-only schema-and-lineage explorer rendering 97 table relationships, column-level lineage, and eight-band freshness into a single self-contained HTML view.

## RESEARCH

**Undergraduate Researcher — Bayesian Volatility & Tail-Risk Modeling** Penn State SURE Program *May 2026 – Jul 2026*

Department of Statistics · Mentor: Dr. ByeongJip Kim · PI: Prof. Xiaoyue Niu

- Built a conditional predictive-distribution pipeline over 30 years of index data (1996–2026), implementing GARCH and particle-filter models from scratch and validating each by synthetic parameter recovery.

**Undergraduate Researcher — Power-System Stability & Oscillation Prediction** Penn State EECS *Sep 2026 – Present*

EECS Undergraduate Research Program · Advised by Prof. Yan Li

- Selected for a funded research award; building learning-based models for power-system stability assessment from PMU time-series measurements.

## SKILLS

**LANGUAGES** Python · SQL · Scala · R

**DATABASES** PostgreSQL · TimescaleDB · pgvector · data modeling · bitemporal / slowly-changing dimensions · query optimization · partitioning

**PIPELINES** ETL / ELT · async Python · Procrastinate · systemd · dead-letter queues · idempotent, resumable backfills · Apache Spark

**DATA QUALITY** point-in-time correctness · reconciliation & coverage audits · staleness detection · data lineage · observability & alerting

**BACKEND** FastAPI · SQLModel · Alembic · OpenAPI codegen · row-level security · OAuth · Cloud Pub/Sub · GitHub Actions CI · Linux

**SPOKEN** Korean (native) · English (fluent) · Mandarin Chinese (HSK 5, 270/300)

## HONORS

Sung and Fumi Lee Scholarship — Penn State (Fall 2026) · Mu Sigma Rho, National Statistics Honor Society