Milandu Keith Moussavou Boussougou

Dongjak-gu, Seoul, South Korea

mbmk92@gmail.com

milandukeith.com — linkedin.com/in/mbmk92 — github.com/selfcontrol7 — Google Scholar

Summary

AI/LLM Engineer and Ph.D. candidate with 6+ years of industry and research experience building and deploying AI systems. Experience includes model fine-tuning, LLM agents and evaluation, RAG, OCR, and multimodal learning. Recently fine-tuned a Korean OCR vision-language model (LoRA SFT and DPO) that made 46% fewer character errors than the production model on held-out client pages. Hands-on experience with model serving, monitoring, and data pipelines. Like to resolve complex production issues.

Experience

Enaible - Too Much AI (TMAI) **May 2025 - Present**

AI-LLM Engineer / Data Scientist

OCR Model Fine-Tuning and Release

- Fine-tuned LightOnOCR (2-1B, then 3-1B), a 1B vision-language OCR model, with LoRA SFT and DPO on Korean business reports to replace HunyuanOCR in a live pipeline. On 418 held-out pages the 3-1B model made 46% fewer Korean character errors and 71% fewer digit errors.
- Halved looping pages (6 to 3) with DPO on 1,409 preference pairs (GRPO did not improve on it), and recovered layout-box quality with a second LoRA stage (OmniDocBench box F1 0.515 to 0.557).
- Served the models with vLLM and Docker on shared GPUs; released an OCR gateway with 3 models behind one API for side-by-side tests, Grafana monitoring and a scripted rollback.

LLM Fine-Tuning for a Korean-First AI Assistant

- Led the LoRA fine-tuning of Qwen3.8-27B for a Korean-first assistant that calls tools, on privacy-scrubbed production conversations up to 64k tokens. On 151 held-out real conversations a blind LLM judge scored the best run 6.10 vs 5.46 for the base model ($p = 0.002$), with half the rule violations.
- Found and fixed a tool-calling regression caused by the training mix: BFCL v4 parallel-call accuracy went from 0% in the first run to 92.0% (base 82.5%).
- Built the session-split data pipeline (13,382 production records) and the evaluation harness (BFCL v4, FunctionChat-Bench, IFEval-Ko, LLM judge), with AI coding agents under my direction.

Thai Language Model Fine-Tuning

- Fine-tuned GPT-4, Qwen, and SolarPro (with pruning stage) models for Thai language using synthetic and augmented training data, and evaluated their performance on ThaiMTBench and ThaiExam.

LLM Evaluation, Optimization and RAG Systems

- Built a four-tier web-retrieval cascade (Jina Reader → FireCrawl → Crawl4AI → HTTP fallback) with per-domain routing for the agent's tools.
- Built browse_web, a tool that lets the agent use interactive sites (click, scroll, type, submit forms), and wrote the agent's tool-use instructions.
- Optimized Gemini Flash parameters on a 32-case evaluation set and revised the evaluation criteria and reference answers, increasing the average score from 0.64 to 0.88.

Freelance - Independent project **2025 - Present**

AI-LLM Engineer

Thai Fruit Garden - Conversational Order Agent

- Built a Facebook Messenger order agent (Gemini 3.1 Flash-Lite with tool calling, FastAPI, SQLite, Cloud Vision OCR) for shops selling in Korean, Thai, Vietnamese and English. Production ran in a silent observe mode that turns staff chats into orders; an auto-reply mode was built and tested.

- For a Thai grocery shop in Korea (Jun–Jul 2026) it captured 5,964 orders; its share of shipped orders rose from 29–30% to 66%, checked against the courier waybills.
- Matched transfer receipts to bank-deposit alerts: 52% of eligible receipts (7,411 of 14,260) were verified with no human, and 5 suspected forgeries were caught.
- Designed the agent tools (catalog, pricing, Kakao address check, order history, payment info) and a three-layer order check; a retry queue recovered 4,953 calls after Gemini outages.

Soongsil University **Mar 2021 - Feb 2027**

AI Researcher / Ph.D.

Korean Copyright Metadata Extraction System - AI pipeline for government digital content registration (한국문화정보원, 한국저작권위원회)

- Built a pipeline that extracts metadata from Korean contracts, consent forms, images and videos: OCR, then LLM extraction (Qwen3.5) and BERT-CRF NER (KLUE-RoBERTa-Large) in parallel; vision LLMs for images and video.
- Reached 93.6% macro-F1 on 22 work attributes and 100% on 15 contract fields in 4,000-item rehearsals of the official test (target 85%).
- Built an LLM arbiter that compares and reconciles the LLM and NER outputs field by field and labels each field (agreed, conflict, LLM-only, NER-only) with a short justification in Korean.
- Added OCR and vision fallback chains (Alibaba → Mistral → Naver; Qwen Omni → Qwen Plus → Gemma) with retries, keeping extraction running through outages, rate limits and model retirements.

Korean Voice Phishing Detection

- Developed text, acoustic, and multimodal detection models using KoBERT, KR-SBERT, attention CNN–BiLSTM, and back-translation augmentation.
- Compared audio–text fusion (concatenation, FiLM, cross-attention) with single-modality models on 1,188 calls: all setups scored 0.993–0.998 call macro-F1.
- Built compact acoustic classifiers (MFCC, eGeMAPS, Wav2Vec2); the 23K-parameter eGeMAPS model reached 99.71% segment-level vishing F1 and runs on a Galaxy S24 Ultra.
- Built and documented Korean vishing corpora with 2,927 transcripts and 1,417 calls, combining real FSS scam recordings with legitimate conversations.

Daewoong Foundation **Jul 2021 - Mar 2022**

Data Scientist / ML Engineer

- Trained time-series forecasting models for inventory management and out-of-stock prediction.
- Built web-based security log monitoring systems using Django and PostgreSQL.
- Deployed reproducible machine learning pipelines in Linux environments.

WeCrest **Sep 2020 - Oct 2021**

Data Scientist / Full-Stack Developer

- Led the design of a data pipeline, combining Django and Scrapy to scrape four major attorney-directory websites (TTIPAB, CITMA, FICPI, and WIPO Madrid) for the Madrid Protocol trademark intelligence system.
- Deduplicated and preprocessed approximately 50,000 attorney and firm records from the Madrid Protocol database using TF-IDF over character n-grams and cosine similarity.
- Developed and deployed a web application using Django, MySQL, and Linux servers.
- Managed cloud deployments on Microsoft Azure and Google Cloud Platform.

Huawei Technologies, CFAO Technology **2014 - 2017**

- Configured and deployed LTE base stations, CCTV, and 900+ IP phones for government and hospital sites.

Education

Soongsil University Mar 2021 - Feb 2027 (expected)
Ph.D. in Computer Science (Artificial Intelligence)

Soongsil University Sep 2018 - Mar 2021
M.S. in Computer Science (Artificial Intelligence)

Institut Supérieur de Technologie de Libreville Sep 2011 - Aug 2014
B.S. in Computer Science (Database - Network - Web)

Publications

Journal papers

- S.-H. Hwang, **M. K. Moussavou Boussougou**, and D.-J. Park, "Performance and Stability Analysis of a Fine-Tuned BERT-BiLSTM-CRF NER Model for Automated Information Extraction in Openly Licensed Works," *Journal of Software Assessment and Valuation (JSAV)*, vol. 21, no. 4, pp. 53–61, 2025. doi:10.29056/jsav.2025.12.05
- **M. K. Moussavou Boussougou**, P. Hamandawana, and D.-J. Park, "Enhancing Voice Phishing Detection Using Multilingual Back-Translation and SMOTE: An Empirical Study," *IEEE Access*, vol. 13, pp. 37946–37965, 2025 (SCIE). doi:10.1109/ACCESS.2025.3545250
- **M. K. Moussavou Boussougou** and D.-J. Park, "Attention-Based 1D CNN-BiLSTM Hybrid Model Enhanced with FastText Word Embedding for Korean Voice Phishing Detection," *Mathematics*, vol. 11, no. 14, art. 3217, 2023 (SCIE). doi:10.3390/math11143217
- **M. K. Moussavou Boussougou** and D.-J. Park, "Exploiting Korean Language Model to Improve Korean Voice Phishing Detection," *KIPS Transactions on Software and Data Engineering*, vol. 11, no. 10, pp. 437–446, 2022. doi:10.3745/KTSDE.2022.11.10.437

Under review

- **M. K. Moussavou Boussougou** and D.-J. Park, "Acoustic Representation Comparisons for Korean Voice-Phishing Detection with Codec Sensitivity and On-Device Feasibility Analysis," *IEEE Access*, under review.

Conference papers and patents

- **11 conference papers** (3 international: EEECS 2025 and 2026; 8 Korean), 7 as first author.
- **3 Korean patents**: 1 granted (KR 10-2640315, voice phishing detection) and 2 pending.

Skills

GenAI, LLMs and Agentic Systems: LLM Evaluation and Benchmarking; LLM/VLM Fine-Tuning (LoRA, DPO, GRPO); Prompt Engineering; Evaluation Criteria Design; Parameter Optimization; RAG Pipelines; Tool Calling.

ML and Data: MLOps; Deep Learning; Data Science; NLP; OCR; Vision-Language Models; Multimodal Learning; Model Evaluation; Synthetic Data Generation and Augmentation.

Tools and Frameworks: Python; PyTorch; SQL (MySQL, SQLite, Postgres); Hugging Face (PEFT-LoRA, QLoRA); Docker; vLLM; Prometheus/Grafana; FastAPI; Django; Scrapy; Crawl4AI; TensorFlow; Pandas; Matplotlib; Seaborn; scikit-learn; Langfuse; AI coding agents (Claude Code, Codex).

Platforms and Infrastructure: Linux, Oracle Cloud; Google Cloud Platform; Microsoft Azure; Weights & Biases (W&B); CI/CD.

Languages: Korean (Advanced, KIIP Level 5); English (Fluent); French (Native)

Honors and Awards

- **Ambassador's Award, AGA × SAF Awards**, 9th Seoul Africa Festival: distinguished contribution to the development of Gabon, *Oct 2026*
- **우수논문상 (Outstanding Paper Award)**: Korean Institute of Information Scientists and Engineers (KIISE), *2022 and 2025*
- **장려 논문상 (Encouragement Paper Award)**: Korea Information Processing Society, *2021*
- **Global Korea Scholarship (GKS/KGSP)**: National Institute for International Education (NIIED), *2017*

Leadership and Community

Samba-Corée (Non-profit, Gabonese community in Korea) Mar 2021 - Feb 2025

- **President (2024–2025):** Led the organization, secured funding and partnerships, worked with the embassy, and organized programs to help Gabonese students and workers settle into life in Korea.
- **General Secretary (2022–2023):** Coordinated meetings, prepared meetings and official correspondence, and established the organization's document and recordkeeping practices.
- **Treasurer (2021–2022):** Managed bookkeeping, budgets, and financial reporting for several events.