


# DAEKEUN KIM (김대근, 金垚權)

 <https://www.linkedin.com/in/daekeun-kim>

 <https://github.com/daekeun-ml>

 <https://huggingface.co/daekeun-ml>

 [housekdk@naver.com](mailto:housekdk@naver.com)

 +82-10-3847-7950

With 22 years of experience, including 19 years specializing in AI/ML, Daekeun has worked across startups, manufacturing, FSI, and cloud, gaining deep expertise in developing and deploying AI/ML products. Daekeun holds 6 first-author patents and have led AI/ML projects that contributed to the mass production of over 20 products.

As an AI/ML technical specialist, Daekeun has led over 150+ AI/ML workloads, delivered 80+ seminars as the ML community tech leader, and mentored 18 ML experts. While his career was deeply rooted in computer vision, his expertise now spans all AI/ML domains, including GenAI, with a strong focus SLM fine-tuning and SLM/LLM serving. Daekeun has a double major in computer science and mathematics, and a master's in computer science specialized in ML.

## PROFESSIONAL EXPERIENCE

**Amazon Web Services, South Korea**

**May. 2025 - Current**

### **Principal AI/ML Specialist Solutions Architect**

- Provide specialized technical leadership and guidance, collaborating with customers in Korea to implement cutting-edge AI/ML solutions on AWS, including FM (Foundation Model) pre-training/fine-tuning/serving, Agentic RAG, and evaluation-driven GenAIOps. As AWS' top AIML technical expert, solving the most challenging problems for Korea customers and contributing to the growth of internal AI/ML expertise.

**Microsoft, South Korea**

**Mar. 2024 - May. 2025**

### **Senior Technical Specialist - Machine Learning / AI Global Black Belt**

- Lead deep dive technical support for building AmorePacific's BFM (Beauty Foundation Model) and coached domain-specific data preparation and fine tuning of Phi-4 model. The customer journey is featured as the biggest highlight in Microsoft AI Tour in Seoul, and [the breakout session](#) was presented with the customer.
- Developed 24 reusable end-to-end technical assets, including 2 end-to-end hands-on labs, 3 Microsoft tech blog, 3 Hugging Face dataset/models, 5 L300-400 AIML decks, 11 workshop samples. The most representative technical asset is [SLM Innovator Lab](#), which directly performed narrative and technical asset development, customer delivery, and APJ scaling. Directly contributed to Azure adoption by delivering to 10+ Korea customers and indirectly contributed to \$600K revenue by reusing in China/Japan/India region. It was also highlighted at WW level, presenting assets and use cases directly.
- Directly contributed content for Korean customers to Microsoft's official Hugging Face model hub in collaboration with Microsoft Research. Registered the Korean benchmark results of the [Phi-3.5-Mini](#) and [Phi-3.5-MoE \(Mixture of Experts\)](#) models, and the Korean speech fine-tuning assets and experimental/benchmark results of the [Phi-4-Multimodal model](#).

**Amazon Web Services, South Korea**

**Jun. 2019 - Mar. 2024**

**Senior AI/ML Specialist Solutions Architect / Worldwide Specialist Organization**

- AWS's biggest contributor to the proliferation of Korean NLP/LLM workloads. 21 public speaking engagements on the topic, and has published 47 hands-on labs such as [GenAI-Korean-LLM](#), [Korean NLP use-cases](#), [NLP distributed training](#), and [AWS Inferentia](#). Hands-on labs have been reused by 100+ customers, of which 30+ have contributed to production launch. [In the Hugging Face Open Ko-LLM leaderboard](#), [broke the barrier of 50 scores](#) for the evaluation metric for the 1st time in Korea by fine-tuning a Llama2-based model with DPO (Direct Preference Optimization).
- Made the biggest contribution to win AIML adoption on AWS with an annual average of \$500K ARR. Led 100+ workloads represented by [Samsung Electronics' Bug Triage System Migration](#), [SK Telecom MLOps adoption](#), [Kakaostyle SageMaker Distributed Training](#), [Hyperconnect Inferentia adoption](#), [PoSCO GenAI disaster prevention system](#), two of which were presented by the customer in AWS re:Invent.
- As a Korea ML TFC (Technical Field Community) leader, nurtured AIML personnel and exerted the greatest influence in APJ as a tech lead. Directly promoted 18 ML AoDs (Aspiring Area of Depths) out of a total of 37 AoDs in Korea, delivered ML deep-dive seminars 67 times, the most at APJ.
- Has contributed the most to the revitalization of AIoT in Korea as APJ's AIoT tech leader. [Developed ML demo based on Amazon SageMaker and AWS IoT Greengrass 2.0](#) required for smart factory, and presented it to 4 external public speaking sessions and 2 internal sessions.

**Hyundai Card & Capital, Co., Ltd., South Korea**

**Feb. 2017 - Jun. 2019**

**Senior Data Scientist and Team Leader / Data Science Division**

- Created a deep learning-based model for tabular data to extract off-the-shelf latent features. Achieved 2.4% better AUPRC (Area Under the Precision-Recall Curve) than the highest score of the existed models.
- Leveraged advanced feature engineering techniques and tree-ensemble models, including LightGBM and XGBoost, to enhance the churn prediction model, resulting in a 5.4% increase in predictive performance over baseline models.
- Designed and developed a vehicle repurchase prediction model in collaboration with various departments, leading to a 20% higher customer penetration rate compared to traditional models.
- Applied Multi-Armed Bandits (MAB) algorithms, such as Thompson Sampling and Contextual Upper Confidence Bound (UCB), to optimize customer-centric marketing services, improving decision-making and targeting efficiency.

**LG Display, Co., Ltd., South Korea**

**Mar. 2013 - Feb. 2017**

**Senior Research Engineer and Project Leader / OLED Algorithm Team / Circuit Research**

- As the R&D leader of three major projects, led a team of four members to develop innovative algorithms and successfully apply them to mass production applications.
- Designed and implemented a sophisticated MURA (Refers to uneven luminance and splotchiness) detection and compensation algorithm using inpainting, vignetting correction, and auto-encoder techniques. This algorithm was applied to the mass production of over 20 LCD/OLED models, significantly improving display quality.
- Created a cutting-edge viewing angle compensation algorithm by utilizing Higher-order Singular Value Decomposition (SVD) and regression techniques. This algorithm was successfully applied to the mass production of Panasonic 55"/65" LCD models, enhancing viewing experiences.
- Developed a CVD (Color Vision Deficiency) correction algorithm by implementing several Daltonization-based algorithms and representation learning.

**Atalgo, Inc, San Diego, CA, USA**

**Feb. 2008 - Jan. 2010**

**Research Engineer / Vision Core Development Department**

- Reduced the bit-rate required to stream 720p video by 45% through the prototyping and implementation of various computer vision algorithms, such as pre/post-processing, foreground detection, and object detection/tracking. This technology has since been integrated into the algorithms used by Samsung affiliates.
- Developed and evaluated multiple video compression algorithms, such as haze removal techniques using Shape Adaptive Discrete Cosine Transform (SA-DCT), DCT truncation, and Dark Channel Prior, to enhance video quality and reduce data size.
- Developed a high-speed face-detection engine utilizing Local Binary Pattern (LBP) and Adaboost algorithms, significantly improving the performance of face detection in three applications.

**Nara-i-net Co., Ltd., South Korea**

**Sep. 2001 - Sep. 2004**

**Software Engineer / R&D Department**

- Enhanced search functionality and user interface for Panrae, a legal precedent information system, implementing keyword extraction features that achieved 57% improvement in search speed and significantly enhanced query performance and user experience.
- Led front-end development and assisted with back-end programming for the website YesLaw (<http://www.yeslaw.com>), which provides legal information system.
- Developed the website CareerNet (<http://www.career.re.kr>), which helps people to find their talents and plan their future careers.

**AWARDS**

- 2025 FYQ3 CSA (Commercial Solution Areas) award (2025). Developed reusable technology assets and disseminate them cross-area. <https://azure.github.io/slm-innovator-lab>
- 2023 AWS Q4 Tech blog awards for most unique visitors, 9.46% of total unique visitors.
- 2022 AWS Q2 AIML TFC awards for providing the most technical and accurate answers to AWS internal community.
- 1st place on the Hugging Face Open Ko-LLM leaderboard ([1st Korean to achieve 50+ scores](#)).
- Nominated for Korea Solutions Architect (SA) of the Year 2021 at AWS, out of over 150 SAs.
- 2nd place award in the 1st and 2nd data science competition at Hyundai Card (out of 32 participants, out of 14 participants, respectively).
- Received the Outstanding Poster Paper Award for "Automatic Micro Defect Detection in Non-repetitive Patterns for High-resolution TFT-LCDs" at IMID 2015.
- Honored with the LG Vision Incentive Award for outstanding performance at LG Display for 3 consecutive years (2014-2016)

**EDUCATION**

**University of Michigan, Ann Arbor, MI**

Master of Science in Computer Science and Engineering, Apr. 2012

**University of Wisconsin, Madison, WI**

Bachelor of Science, Double major in Computer Science and Mathematics, Dec. 2007

**Selected Academic Projects:**

Emotion Classification using Discriminative RBM (2012): A novel classification for classifying highly ambiguous emotions using Emotion Profiles and Discriminative Restricted Boltzmann machines.

Mortality prediction for ICU patients (2011): A novel approach to classify mortality using supervised

learning based on keyword/chart-event feature vectors from MIMIC-II database.

Background Subtraction using 3-way RBM (2011): Multiple visible units in 3-way Restricted Boltzmann machines, allowing machines to capture the correlation of frames on moving camera.

## **PATENTS**

**Daekeun Kim** and Jinwoong Shin. Method and Device of LCD Device for Generating Compensation Data. KR-10-2018-0074907.

**Daekeun Kim**, Euiyeol Oh, and Jihoon Park. Display Device and Method for Driving the same. KR-10-2018-0073296.

**Daekeun Kim**, Sanglin Lee. Prediction Method and System for Predicting a Luminance Decline of Display Device. KR-10-2017-0026974.

**Daekeun Kim**, Euiyeol Oh, and Kipyoo Hong. Method and Device for Generating Compensation data of Display Device. KR-10-2016-0016214.

**Daekeun Kim** and Jinkoo Kim. Inspection Method and Device for Flat Panel Display Device. KR-10-2017-0004811.

Jungwan Bae, Namseok Choi, **Daekeun Kim**, and Byungkyu Dan. Apparatus and Method for Compensating of Brightness Deviation. KR-10-2016-0004136.

## **PAPERS**

**Daekeun Kim** and Youngjoon Choi (2023). AWS 2nd gen. Inference Accelerator Analysis for Serving High-Performance and Cost-Effective Generative AI Models. *KICS Summer conference*, 2023.

Youngjoon Choi and **Daekeun Kim** (2023). Comparison and Analysis for the Performance of Deep Learning-Based Time Series Prediction Algorithms According to Increasing Model Size. *The Journal of KICS*, 2023-01 Vol.48, pp.123-128. doi:10.7840/kics.2023.48.1.123.

**Daekeun Kim** and Youngjoon Choi (2022). Optimizing Cost-Effective Deep Learning Inference Performance in the AWS Cloud Environment. *KICS Summer conference*, 2022.

Soonam Lee and **Daekeun Kim** (2018). Background Subtraction using the Factored 3-way Restricted Boltzmann Machines. arXiv:1802.01522.

Euiyeol Oh, **Daekeun Kim**, Jong Sang Baek, and Yoonsik Choe (2016). A Method on Choosing the Preferred Camera Resolution for Mura Compensation. *IMID*, 2016.

Euiyeol Oh, **Daekeun Kim**, Jong Sang Baek, and Yoonsik Choe (2015). Automatic Micro Defect Detection in Non-repetitive Patterns for High-resolution TFT-LCDs. *IMID*, 2015.

## **OPEN SOURCED CONTRIBUTIONS**

### **[Selected GitHub repositories]**

**SLM Innovator Lab**: Structured hands-on lab covering 1) QnA generation from raw data, 2) synthetic data augmentation, 3) fine-tuning and model serving, 4) Evaluation-driven LLMops with Azure AI.

Lead contributor. <https://github.com/Azure/slm-innovator-lab>

**Agent Innovator Lab**: Best practices for developing Agentic AI covering 1) Agentic design pattern, 2) Evaluation design pattern, 3) Optimization design pattern with Azure AI.

Key contributor. <https://github.com/Azure/agent-innovator-lab>

**Generate Synthetic QnAs from Real-world Data**: Create/augment a QnA dataset from complex unstructured data, assuming a real-world scenario.

Lead contributor. <https://github.com/Azure/synthetic-qa-generation>

**Fine-tune/Evaluate/Quantize SLM/LLM using the torchtune on Azure ML**: Starter for quickly and easily applying SLM/LLM fine-tuning, evaluation, and quantization with torchtune on Azure ML.

Lead contributor. <https://github.com/Azure/torch-tune-azureml>

**SLM/LLM Fine-tuning on Azure:** Fine-tuning/serving an open source SLM/LLM on Azure ML.

Lead contributor. <https://github.com/Azure/azure-llm-fine-tuning>

**Korean Language Proficiency Evaluation:** Performs benchmarking on KMMLU, CLiCK, and HAE-RAE Korean dataset with minimal time and effort.

Lead contributor. <https://github.com/daekeun-ml/evaluate-llm-on-korean-dataset>

**Azure GenAI Utils:** A set of utilities for working with Azure GenAI.

Lead contributor. <https://github.com/daekeun-ml/azure-genai-utils>

**KoSimCSE Training on Amazon SageMaker:** Hands-on for ML beginners to perform SimCSE step-by-step. Implemented both supervised SimCSE and unsupervised SimCSE on Amazon SageMaker.

Lead contributor. <https://github.com/daekeun-ml/KoSimCSE-SageMaker>

**Korean NLP downstream tasks Hands-on Labs on Amazon SageMaker:** A collection of Korean NLP hands-on labs on Amazon SageMaker, converging classification, NER (Named Entity Recognition), QnA, Sentence BERT, Natural Language Inference, Summarization, Translation, and TrOCR.

Lead contributor. <https://github.com/aws-samples/sm-kornlp>

**AutoGluon on AWS:** Example codes for various AutoML tasks using AWS AutoGluon.

Lead contributor. <https://github.com/aws-samples/autogluon-on-aws>

**AWS Generative AI/Machine Learning Samples – Korea:** A collection of localized (Korean) AWS AI/ML workshop materials for hands-on labs.

Key Contributor. <https://github.com/aws-samples/aws-ai-ml-workshop-kr>

### [Selected Hugging Face models]

**Phi-4-multimodal-finetune-ko-speech** (2025): A fine-tuned model with added Korean speech data achieved CER 1.61% and CER 3.54% on the zeroth test dataset. This repo is registered in Appendix B of the official Microsoft HuggingFace Phi-4-multimodal repo.

<https://huggingface.co/daekeun-ml/Phi-4-multimodal-finetune-ko-speech>

**Phi-2-Ko** (2024): A model that performed continual pretraining using a 5GB Korean corpus for PoC purposes.

<https://huggingface.co/daekeun-ml/phi-2-ko-v0.1>

**Llama-2-ko-DPO-13B** (2023): 1st Korean LLM model to exceed the average metric of 50 percent.

<https://huggingface.co/daekeun-ml/Llama-2-ko-DPO-13B>

**TrOCR for Korean Language** (2022): A model trained using sentence separator library (Kiwi Python wrapper) for PoC purposes.

<https://huggingface.co/daekeun-ml/ko-trocr-base-nsmc-news-chatbot>

**Sentiment Binary Classification** (2021): Fine-tuning with KoELECTRA-Small-v3 model and Naver Sentiment Movie Corpus dataset)

<https://huggingface.co/daekeun-ml/koelectra-small-v3-nsmc>

### [Hugging Face dataset]

**Korean GLAN (Generalized Instruction Tuning) Instructions Dataset** (2025): 300K seedless synthetic data using Phi 3.5 MoE, which is useful for reinforcing in-depth knowledge at the undergraduate/graduate level.

<https://huggingface.co/datasets/daekeun-ml/GLAN-qna-kr-300k>

**Naver News Summarization** (2022): A custom dataset by crawling Naver News for the Korean NLP model hands-on. 27K articles on IT and economics from July 1-10, 2022.

<https://huggingface.co/datasets/daekeun-ml/naver-news-summarization-ko>

## TECH BLOG CONTRIBUTIONS

**Daekeun Kim** (2024). Fine-tune/Evaluate/Quantize SLM/LLM using the torchtune on Azure ML. *Microsoft Tech Community*. [Link to article](#).



**Daekeun Kim** (2024). Generate Synthetic QnAs from Real-world Data on Azure. *Microsoft Tech Community*. [Link to article](#).

**Daekeun Kim** (2024). Fine-tuning Florence-2 for VQA (Visual Question Answering) using the Azure ML Python SDK and MLflow. *Microsoft Tech Community*. [Link to article](#).

Manoranjan Rajguru and **Daekeun Kim** (2024). Fine-tune Small Language Model (SLM) Phi-3 using Azure Machine Learning. *Microsoft Tech Community*. [Link to article](#).

**Daekeun Kim** (2023). Quickly build high-accuracy Generative AI applications on enterprise data using Amazon Kendra, LangChain, and LLMs. *AWS Korea Tech Blog*. [Link to article](#).

**Daekeun Kim** (2023). Build a powerful question answering bot with Amazon SageMaker, Amazon OpenSearch Service, Streamlit, and LangChain. *AWS Korea Tech Blog*. [Link to article](#).

**Daekeun Kim** (2023). Interactively fine-tune Falcon-40B and other LLMs on Amazon SageMaker Studio notebooks using QLoRA. *AWS Korea Tech Blog*. [Link to article](#).

**Daekeun Kim** (2023). Deploy Falcon-40B with large model inference DLCs on Amazon SageMaker. *AWS Korea Tech Blog*. [Link to article](#).

Hyundoo Jin, **Daekeun Kim**, Daeyeol Shim, and Daehoon Oh (2023). Using Amazon SageMaker Distributed Training with KakaoStyle to Model a Category Automated Classification System. *AWS Korea Tech Blog*. [Link to article](#).

**Daekeun Kim** and Hyeonsang Jeon (2023). Train a Large Language Model on a single Amazon SageMaker GPU with Hugging Face and LoRA. *AWS Korea Tech Blog*. [Link to article](#).

Sungwon Han, Heewon Ko, Hyojung Kang, Kyungdae Cho, Sanghwa Na, and **Daekeun Kim** (2023). SK Telecom's Case Study of Building a ML Pipeline Using AWS Inferentia and AWS Step Functions. *AWS Korea Tech Blog*. [Link to article](#).

## **TECH BOOK TRANSLATION**

**Daekeun Kim** and Daeyeol Shim (2023). Co-translated “[Machine Learning System Engineering in Action](#)”, authored by Ben Wilson.

**Daekeun Kim** and Youngmin Kim (2023). Co-translated “[Designing Machine Learning System](#)”, authored by Chip Huyen.

## **PUBLIC SPEAKING**

Led the way in thought leadership activities among all AWS Solution Architects in Korea and across the WWSO APJ Specialist Solutions Architects. Not only participated in numerous public speaking engagements, but contents were also cited by 50+ Korea customers, particularly in the Enterprise, Digital Native, Startup, Manufacturing, and FSI segments. Achieved the highest CSAT at AWS tier-1 events three times.

### **AWS Seoul Summit**

- Track Owner (2024), AWS Summit 2024 - GenAI Track
- Demo booth Leader (2023), [AWS Summit 2023 - GenAI Ask the expert](#) (Live QnAs with 222 in-person attendees.)
- Workshop Speaker (2022), [AWS Summit 2022 - Model Serving patterns](#) (250 online attendees, CSAT 4.7. Highest CSAT of all sessions at the AWS Seoul Summit 2022.)
- Conference Speaker (2021), [AWS Seoul Summit 2021 - SageMaker Distributed Training](#) (CSAT 4.65. Highest CSAT of all sessions at the AWS Seoul Summit 2021.)
- Conference Speaker (2020), [AWS Seoul Summit 2020 - Reinforcement Learning on AWS](#) (CSAT 4.48)
- Demo Speaker (2020), [AWS Seoul Summit 2020 - AWS Smart Factory](#). Collaborated with 6 people to create a smart factory demo. (CSAT 4.64)

### **AWS Tier-1 events (AWS Innovate and AWS Special Webinar)**

- Conference Speaker (2023), AWS Innovate 2023 Data Edition – GenAI PEFT and RAG deep dive. (404 online attendees, CSAT 4.6)
- Track Owner (2023), [AWS Innovate 2023 - Modern App GenAI Track owner](#) (2,085 online attendees, CSAT 4.5)
- Conference Speaker (2022), [AWS Innovate 2022 AIML Edition – Hugging Face on SageMaker: Training and Deploying Cloud-Native Korean NLP models for everyone](#) (998 online attendees, CSAT 4.76. Highest CSAT among all tracks.)
- Webinar Speaker (2022), AWS Special Webinar - Presented a [SageMaker serving overview](#) and [hands-on lab sessions](#) during the two day SageMaker workshop. (623 online attendees, CSAT 4.4)
- Conference Speaker (2021), [AWS Innovate AIML Edition – SageMaker model serving](#) (533 online attendees, CSAT 4.5)
- Webinar Speaker (2020), [AWS Special Webinar - SageMaker Live Demo](#) (486 online attendees, CSAT 4.53)

### Conferences and Webinars hosted by AWS

- Webinar Speaker (2023), [GenAI Technical webinar – LLM Fine-tuning](#) (415 online attendees, CSAT 4.6)
- Workshop Speaker (2023), [2023 AWS Enterprise boost](#) – Special topics in Distributed training
- Technical Lead Advisor (2023), 2022 AWS Enterprise boost – AIML
- Webinar Speaker (2023), AWS Builders L300: AutoGluon on AWS (455 online attendees, CSAT 4.44)
- Workshop Speaker and Supporter (2022), MLOps and Distributed Training Workshop
- Workshop Speaker (2022), AWS Enterprise boost – AIoT edge device computing
- Technical Lead Advisor (2022), 2022 AWS Enterprise boost – AIML
- Webinar Speaker, AWS Builders L200: Overview of AWS's Computer Vision services covering from business analysts to ML engineers (160 online attendees, CSAT 4.2)
- Workshop Speaker (2022), IoT Developers day (21 online attendees, CSAT 4.44)
- Conference Speaker (2022), AWS re:Invent 2022 recap for Korea Enterprise customers
- Conference Speaker (2022), AWS re:Invent 2022 recap (431 online attendees, CSAT 4.72)
- Workshop Speaker (2021), Machine Learning bootcamp – SageMaker Studio introduction (281 online attendees, CSAT 4.76)
- Workshop Speaker (2021), Machine Learning bootcamp – SageMaker L300 deep dive (425 online attendees, CSAT 4.69)
- Workshop Speaker (2021), 2021 AWS Manufacturing boost – ML Inference on IoT Greengrass
- Technical Lead Advisor (2021), 2021 AWS Manufacturing boost – AIML
- Conference Speaker (2020), AWS Builders L300: KoBERT fine-tuning on AWS (CSAT 4.53. Highest CSAT of all AWS Builders sessions in June 2020.)
- Conference Speaker (2020), AIoT and AI on Edge workshop with LG AIoT board (CSAT 4.62. Highest CSAT among the public workshops in Q2.)
- Conference Speaker (2020), AWS SageMaker Deep Dive (CSAT 4.72)
- Session Supporter (2019). AWS Developers day - Created and demonstrated a distributed training demo that trains ResNet with ImageNet data in 45 minutes.

### External Conferences

- Microsoft DDD (Developer! Developer! Developer!) Speaker, <https://dddseoul.kr>, [YouTube Recording](#) (100 in-person attendees)
- COEX FastCampus 2024 GenCon AI Speaker (2024), [https://fcaing.co.kr/ai\\_gencon2024/](https://fcaing.co.kr/ai_gencon2024/), [<News 1>](#), [<News 2>](#) (900+ in-person attendees)
- COEX GenAI Summit 2023 Speaker (2023), [<News 1>](#) [<News 2>](#) [<News 3>](#) (250+ in-person attendees)
- Conference Speaker (2023), [AWS Korea User group presentation – LLM Fine-tuning](#) (36 in-person attendees)
- Conference Speaker (2022), [AWS X eTech] Smart Factory Predictive Maintenance (136 online attendees. Contributed to the discovery of 5 new opportunities.)

- Conference Speaker (2021), AWS Korea User group presentation - Sharing 15 years of Machine Learning experience (58 online attendees)
- Guest Speaker (2021), Smart Factory Predictive Maintenance - AIoT at Bioplus-Interphex 2021 conference
- Conference Speaker (2021), [AIPLUS 2021 – On-device Machine Learning on Cloud.](#)

### **Selected AWS Internal Conferences**

- Conference Speaker (2023), Gen'd AI Technical Workshop (167 online and in-person attendees, CSAT 4.4)
- Workshop Leader (2023), Korea GenAI Bootcamp (88 online and in-person attendees, CSAT 4.6)
- Workshop Leader (2023), APJ TechSummit RAG Workshop (123 online and in-person attendees, CSAT 4.89)
- Workshop Leader (2023), APJ TechSummit PEFT (Parameter-Efficient Fine-tuning) Workshop (59 online and in-person attendees, CSAT 4.78)
- Conference Speaker (2023), AIML WW TFC Summit 2023 – LLM Fine-tuning and Serving (154 in-person and online attendees, CSAT 4.89)
- Conference Speaker (2023), AIML WW TFC Summit 2021 – Machine Learning Inference on Edge-device Using SageMaker Neo and IoT Greengrass (126 online attendees, CSAT 4.82)
- Conference Speaker (2022), Amazon Personalize Scale Play L400 Deep Dive (46 in-person and online attendees, CSAT 4.8). Contributed to training 12 Amazon Personalize experts.
- Conference Speaker (2022), Amazon Personalize Scale Play L200 Tech Enablement (76 in-person and online attendees, CSAT 4.86)

## **SKILLS**

- 19 years of experience in AI/ML and GenAI with a total of 22 years of professional experience.
- 12 years of project leadership experience, including 1 year as a team leader at a major Korean conglomerate.
- Over 10 years as an AI/ML technical leader, delivering 80+ internal technical seminars and 26 company-wide tech talks.
- 6+ years of experience with distributed training using PyTorch, SageMaker DDP (Distributed Data Parallel), DeepSpeed, Megatron-LM, and AWS Trainium.
- Work experience in developing AI/ML services for 7 million customers in finance domain.
- 41 public speaking engagements and 200+ internal seminar experiences.
- One of two honorees in the AWS Korea LightningTechTalk Hall of Fame. This honor is given to those who have presented 10 or more times for AWS Korea employees.
- Comprehensive understanding across academic, engineering, and business domains, with exceptional skills in writing papers, writing patents, applying solutions to production, and providing customer support.
- Lead engineer for more than 20 mass production and production applications in the domains of startup, manufacturing, and finance.
- Recognized as top expert and top achiever in Cloud AI/ML. Achieved top tier at AWS for 3 years in a row.
- Supported AI/ML initiatives for various customer segments, including Enterprise, Digital Native Businesses, Retail, Manufacturing, and Startups, leading the successful production launch of over 60 PoCs.
- Languages and Toolkits: Python, PyTorch, Hugging Face, LangChain, LangGraph, AutoGen, and OpenCV.

## **CERTIFICATIONS**

- Microsoft Certifications: Azure AI Engineer Associate (2024-2026), Data Scientist Associate (2024-2026), Azure Data Fundamentals (2024), Azure Fundamentals (2024), and Azure AI Fundamentals (2024).
- AWS Certifications: AWS Machine Learning Specialty (2019-2028), AWS Big Data Specialty (2020-2023), AWS Developer Associate (2019-2022), and AWS Solutions Architect Associate (2019-2022).
- Korea National Certifications: Industrial Engineer Information Processing (2001), Word processor Advanced-level (2001), and Abacus Standard-level (1987).