Juneyoung Park

Lead of Gen.AI, OptAI Inc. • jyoung.park@opt-ai.kr • Website • GitHub

Research Interests

On-Device LLM, Quantization, Efficient LLM, Agentic AI

Education

Ajou University — Suwon, South Korea
Master of Science *in Medical Science* — 2022.09 – 2024.08

- Advisor: Prof. Tae-Joon Kim (Department of Neurology, Ajou University Medical Center)

Sangji University — Wonju, South Korea
Bachelor of Science *in Computer Engineering* — 2017.03 – 2024.08

- Relevant Coursework: Machine Learning, Deep Learning, Algorithms

Professional Experience

Gen.AI — OptAI Inc.
Team lead — 2024.09 – Present

- Led end-to-end on-device **LLM optimization and deployment** for EXAONE models, covering quantization, memory reduction, tokenizer/prompt pipeline design, and ONNX–QNN graph transformation.
- Successfully ported **LLM/VLM to Qualcomm Snapdragon 8 Gen2/Gen3/8 Elite (Gen4)** NPUs through hardware-aware model analysis, KV-cache optimization, and custom NPU offloading.
- Designed Telecom-specific **instruction tuning, persona tuning, and knowledge-grounding strategies** to optimize the model for LG U+ ixi-O customer-support scenarios.
- Built and optimized **custom tokenizer and input pipeline** (token merging, special tokens, vocabulary pruning, pre-tokenization → decoding pipeline) achieving real-time on-device latency.
- Developed the full **PyTorch → ExecuTorch → ONNX → QNN automated conversion pipeline**, including calibration dataset construction and mobile-level performance validation.

Publications

C: Conference, J: Journal, P: Preprint

[C5] **Stop Befor You Forget: NTK-Guided Early Stopping for Continual Learning**
Juneyoung Park, Seongbae Lee, Seongwan Kim, Jaeho Lee
Under review, ICML 2026

[P2] **Robust Meta-Optimization with Kernel-based Adaptation and Task Similarity**
Juneyoung Park, Minjae Kang, Seongwan Kim, Jaeho Lee
arXiv, Preprint

[C4] **Riemannian Optimization for LoRA on the Stiefel Manifold**
Juneyoung Park, Minjae Kang, Seongbae Lee, Haegang Lee, Seongwan Kim, Jaeho Lee
Conference of Empirical Methods in Natural Language Processing (EMNLP) Findings, 2025

[J3] **A novel machine learning model for screening the risk of obstructive sleep apnea using craniofacial photography with questionnaires**

Juneyoung Park et al,.
Journal of Clinical Sleep Medicine (JCSM), 2025, IF: 4.5

[C3] **Riemannian Geometric-base Meta Learning**
Juneyoung Park, Yumi Lee, Tae-Joon Kim, Jang-Hwan Choi
AAAI Conference Artificial Intelligence (AAAI), 2025

[P1] **EB-GAME: A Game-Changer in ECG Heartbeat Anomaly Detection**
Juneyoung Park, Da Young Kim, Yunsoo Kim, Jisu Yoo, Tae-Joon Kim
Preprint

[C2] **Skip-GANomaly++: Skip-Connections and Residual Blocks for Anomaly Detection**
Juneyoung Park, Jae-Ryung Hong, Min-Hye Kim, Tae-Joon Kim
AAAI Student Abstract and Poster Program (AAAI), 2024

[C1] **OSA-NET: An Efficient Convolutional Neural Network for OSA Diagnosis Screening Tool**
Juneyoung Park, Hye-Rim Shin, Tae-Joon Kim
International Conference on Image Processing Theory, Tools and Applications (IPTA), 2023

[J2] **SCLC-Edge Detection Algorithm for Skin Cancer Classification**
Juneyoung Park, Chang-Min Kim, Chan-Hong Park
Journal of Korea Institute of Convergence Signal Processing, 2022

[J1] **Segmentation of Skin Cancer Lesions using ResUNet++**
Juneyoung Park, Young-Hwan Han
Journal of Institute of Electronics and Information Engineers (IEIE), 2022

Industrial Projects

EXAONE-4.0 On-Device LLM Optimization for iOS | *Client: LG Uplus* 2025.07 – Present
Role: Lead Researcher, OptAI Inc.

- Optimized EXAONE 4.0 to run efficiently on Apple's **ANE (Apple Neural Engine)** within mobile memory constraints.
- Performed graph restructuring, operator replacements, and memory optimizations aligned with ANE limitations.
- Made **official contribution to the Anemll framework**.
- Validated performance and model quality for deployment in LG Uplus ixi-O mobile service.

LLM-based Document Analysis | *Client: Polaris Office* 2025.07 – Present
Role: Lead Researcher, OptAI Inc.

- Fine-tuned LLM to convert unstructured documents into structured JSON/Key-Value formats for enterprise search.
- Designed **prompt strategies** for document understanding, hierarchical parsing, and key information extraction.
- Converted models for Intel SoC environments and built a **custom inference engine** instead of using OpenVINO.
- Established unified output schemas suitable for enterprise-scale document retrieval systems.

EXAONE-4.0 On-Device LLM Optimization | *Client: LG Uplus* 2025.09 – 2025.12
Role: Lead Researcher, OptAI Inc.

- Extended previous work to EXAONE 4.0, performing architecture analysis and hardware-aware optimization for Snapdragon SoC.
- Developed **OptAttention**, a custom NPU-optimized attention kernel integrated with Qualcomm QNN.

- Applied structured **pruning techniques** to reduce parameters while maintaining accuracy.
- Achieved reliable long-context(8K) and multi-turn performance on Telecom service.

EXAONE-3.5 On-Device LLM Advancement | *Client: LG Uplus* 2025.02 – 2025.07
Role: Lead Researcher, OptAI Inc.

- Led **Instruction tuning and persona tuning** tailored for **LG U+ Telecom customer-service** domain (ixi-O)
- Designed domain-specific tokenizer optimization, including vocabulary pruning and token merging for Telecom terminology.
- Established knowledge-grounding strategies to reduce hallucinations in Telecom use cases.
- Refined system prompts and conversion templates specialized for Telecom call-center workflows.
- Conducted large-scale latency, quality, and robustness validation on mobile devices.

On-Device LLM Optimization | *Client: LG Uplus* 2024.09 – 2024.12
Role: Principal Researcher, OptAI Inc.

- Optimized the EXAONE3.5 model to run on Qualcomm Snapdragon **8 Gen 3 / 8 Elite(Gen 4) NPU**.
- Applied INT8/4 static quantization, KV-Cache quantization, and On-device modeling.
- Built and debugged the full **PyTorch -> ONNX -> QNN conversion pipeline** for mobile deployment.
- Redesigned model graph and memory layout to fit strict mobile memory constraints.

Invited Talks & Presentation

On Device AI Industry Review
Tech Seminar, DBRAIN September 2025

Tutorial for On-Device LLM using Executorch
Hands-on Tutorial, LG Uplus December 2024

What is the On-Device LLM
Machine Learning Guest Seminar, Ewha Womans University. December 2024

Professional Service

Session Chair
Empirical Methods in Natural Language Processing (EMNLP), 2025

Reviewer
International Conference on Learning Representations (ICLR), 2026
AAAI Conference Artificial Intelligence (AAAI), 2026
International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), 2024

Member
Mentor at Medical AI, Deep daiv, 2023.07 - 2024.04

Reference

Tae-Joon Kim, Advisor Professor
Department of Neurology, Ajou University School of Medicine
E-mail: tjkim23@ajou.ac.kr
Site: https://aunal.org