

Hancheol Park

Senior AI Research Engineer, Nota Inc. | Ph.D. in Computer Science, KAIST | Seoul, South Korea

hancheolp@gmail.com | [Google Scholar](#) | [GitHub](#) | [Hugging Face](#) | [LinkedIn](#) | [Homepage](#)

Research Areas & Technical Expertise

Foundation Models	LLM/VLM pretraining, instruction fine-tuning, retrieval-augmented generation (RAG), RAG-aware fine-tuning, vLLM-based serving, custom model architecture integration in vLLM, model evaluation
Efficient AI	Efficient LLMs/VLMs, lightweight neural architecture design, quantization, pruning, knowledge distillation, model porting, model graph optimization, ONNX/TFLite conversion, NPU/GPU-aware optimization, on-device deployment
Reliable NLP	Hallucination mitigation, ambiguity detection, uncertainty estimation, selective answering, uncertainty-aware generation, confidence calibration
Human-Centric CV	Image and video understanding, surveillance video analytics, event detection, human activity recognition, vision systems for user convenience and safety

Work Experience

Nota Inc. – KOSDAQ-listed in Nov. 2025

Senior AI Research Engineer

Seoul, Republic of Korea

Sep. 2020 – Present

Tech Lead, NetsPresso XPU Enabler

Sep. 2025 – Present

- Developed end-to-end model optimization, porting, and deployment workflows for heterogeneous AI accelerators, adapting and compressing AI models for hardware-specific execution across GPUs, NPUs, and diverse chipsets.

Selected Projects

- Sovereign AI Foundation Model Project, Upstage Consortium (MSIT/NIPA)** *2025 – Present*
Served as technical owner and lead developer for MoE-specific model compression, including INT4/NVFP4 quantization and expert pruning. Developed official compressed releases of Solar-Open models for efficient deployment.
- LLM Porting and Optimization for Qualcomm NPUs** *2025*
Developed optimization and porting workflows for diverse language models, including Llama, Qwen, and EXAONE, targeting Qualcomm NPU execution environments.

Team Lead, NetsPresso Application

Sep. 2022 – Dec. 2025

- Led the development and delivery of on-device AI applications for computer vision and natural language processing, serving enterprise and public-sector clients through competitive bids and custom solution deployments.
- Managed end-to-end application development, from model design and optimization to on-device deployment, field validation, and client-facing demonstrations.

Selected Projects

- Efficient VLM Development for On-device Industrial Safety Applications** *2024*
Developed lightweight vision-language models with fewer than 4B parameters by combining compact SLMs such as Phi and Qwen with vision encoders, followed by pretraining and post-training. Deployed the models on Qualcomm Snapdragon-based mobile devices and industrial QRB5165 platforms for on-device applications that detect diverse risk factors in industrial environments.
- Hybrid LLM System for SK Telecom** *2024*
Served as technical owner and lead developer for a hybrid LLM routing system that routes queries between mobile SLMs and server-side LLMs based on query difficulty. The system was showcased at MWC 2025.
- Competition-driven Validation for Edge AI Solutions** *2023 – 2024*
Led challenge-oriented development to validate technical competitiveness for client-facing proposals, achieving top results in NVIDIA AI City Challenge tracks on traffic and surveillance intelligence.

- **On-device Surveillance Systems for Local Governments**

2022 – 2024

Developed and deployed real-time CCTV-based surveillance systems for local governments including Daejeon and Incheon, enabling on-device detection of events such as fighting, smoking, and falling for immediate monitoring and response.

Team Lead, NetsPresso Performance

Sep. 2020 – Dec. 2022

- Led core technology development for NetsPresso, Nota AI's AI model compression and optimization platform.
- Developed efficient model architectures and compression techniques for production-ready object detection and semantic segmentation models.

Selected Projects

- **Efficient Architecture Design**

Developed lightweight object detection and segmentation architectures, including Lightweight YOLO variants and efficient SegFormer-style models, for real-time and resource-constrained deployment.

- **Pruning and Knowledge Distillation**

Developed proprietary pruning and knowledge distillation techniques for neural network compression, contributing to multiple patented technologies.

Publications

1. **Hancheol Park**, Geonho Lee, Tae-Ho Kim. "DREAM-MoE: Downstream Routing Error-Aware Margin-Preserving Quantization for Mixture-of-Experts Large Language Models." *AdaptFM @ ICML*, 2026.
2. Geonho Lee, **Hancheol Park**, Seunghyun Lee, Jungwook Choi, Tae-Ho Kim. "SRA-MoE: Output-Aware Selective Router Alignment for MoE Quantization." *AdaptFM @ ICML*, 2026.
3. **Hancheol Park**, Geonho Lee, Tairen Piao, Tae-Ho Kim. "Value-and-Structure Alignment for Routing-Consistent Quantization of Mixture-of-Experts Models." *arXiv preprint*, 2026.
4. **Hancheol Park**, Jaeyeon Kim, Geonmin Kim, Tae-Ho Kim. "Nota AI at GenAI Detection Task 1: Unseen Language-Aware Detection System for Multilingual Machine-Generated Text." *GenAIDetect @ COLING*, 2025.
5. **Hancheol Park**, Geonmin Kim. "Where do LLMs Encode the Knowledge to Assess the Ambiguity?." *COLING (Industry Track)*, 2025.
6. Geonmin Kim, Jaeyeon Kim, **Hancheol Park**, Wooksu Shin, Tae-Ho Kim. "Assessing the Answerability of Queries in Retrieval-Augmented Code Generation." *arXiv preprint*, 2024.
7. **Hancheol Park**, Soyeong Jeong, Sukmin Cho, Jong C. Park. "Self-Knowledge Distillation for Learning Ambiguity." *arXiv preprint*, 2024.
8. Jeongho Kim, Wooksu Shin, **Hancheol Park**, Donghyuk Choi. "Cluster Self-Refinement for Enhanced Online Multi-Camera People Tracking." *AI City Challenge Workshop @ CVPR*, 2024.
9. Wooksu Shin, Donghyuk Choi, **Hancheol Park**, Jeongho Kim. "Road Object Detection Robust to Distorted Objects at the Edge Regions of Images." *AI City Challenge Workshop @ CVPR*, 2024.
10. Donghyuk Choi, **Hancheol Park**, Geonmin Kim, Wooksu Shin. "Computationally Efficient Decoders for Semantic Segmentation Models." *KIISE Transactions on Computing Practices*, 2023.
11. Donghyuk Choi, Wooksu Shin, **Hancheol Park**, Ki Min, Youngjun Yoo. "Efficient Semantic Segmentation Models with Weighted Sum-based Feature Fusion Decoders." *KCC*, 2023. **[Best Presentation Award]**
12. Matej Kristan et al. (including **Hancheol Park**). "The First Visual Object Tracking Segmentation VOTS2023 Challenge Results." *VOTS Workshop @ ICCV*, 2023.
13. **Hancheol Park**, Jong C. Park. "Deep Model Compression Also Helps Models Capture Ambiguity." *ACL (Long Paper)*, 2023.
14. Fitsum Gaim, Wonsuk Yang, **Hancheol Park**, Jong C. Park. "Question-Answering in a Low-resourced Language: Benchmark Dataset and Models for Tigrinya." *ACL (Long Paper)*, 2023. **[Outstanding Paper Award]**

15. Jeongho Kim*, Wooksu Shin*, **Hancheol Park***, Jongwon Baek. “Addressing the Occlusion Problem in Multi-Camera People Tracking with Human Pose Estimation.” *AI City Challenge Workshop @ CVPR*, 2023. (* Equal contribution)
16. Bo-Kyeong Kim, Jaemin Kang, Daeun Seo, **Hancheol Park**, Shinkook Choi, Hyoung-Kyu Song, Hyungshin Kim, Sungsu Lim. “A Unified Compression Framework for Efficient Speech-Driven Talking-Face Generation.” *On-Device Intelligence Workshop @ MLSys*, 2023.
17. Bo-Kyeong Kim, Shinkook Choi, **Hancheol Park**. “Cut Inner Layers: A Structured Pruning Strategy for Efficient U-Net GANs.” *HAET Workshop @ ICML*, 2022.
18. **Hancheol Park**, Kyo-Joong Oh, Ho-Jin Choi, Gahgene Gweon. “Constructing a Paraphrase Database for Agglutinative Languages.” *Data & Knowledge Engineering (SCI(E) Journal)*, 2019.
19. Huije Lee, **Hancheol Park**, Wonsuk Yang, Jong C. Park. “Detection of Non-Standard Meaning Usage with Word Embedding.” *HCIK*, 2018.
20. Hoyun Song, **Hancheol Park**, Wonsuk Yang, Jong C. Park. “Predicting Symptoms of Depression for Social Media Users via Linguistic Patterns.” *KSC*, 2017.
21. Wonsuk Yang, **Hancheol Park**, Jong C. Park. “Neural Theorem Prover with Word Embedding for Efficient Automatic Annotation.” *Journal of KIISE*, 2017.
22. **Hancheol Park**, Jung-Ho Kim, Jong C. Park. “Addressing Low-Resource Problems in Statistical Machine Translation of Manual Signals in Sign Language.” *Journal of KIISE*, 2017.
23. Jung-Ho Kim, Najoung Kim, **Hancheol Park**, Jong C. Park. “Enhanced Sign Language Transcription System via Hand Tracking and Pose Estimation.” *Journal of Computing Science and Engineering*, 2016.
24. Wonsuk Yang, **Hancheol Park**, Jong C. Park. “Neural Theorem Prover with Word Embedding for Efficient Automatic Annotation.” *HCLT*, 2016. **[Best Paper Award]**
25. **Hancheol Park**, Jung-Ho Kim, Jong C. Park. “Addressing Low-Resource Problems in Statistical Machine Translation of Sign Language.” *KCC*, 2016. **[Best Paper Award]**
26. **Hancheol Park**, Gahgene Gweon, Jeong Heo. “Affix Modification-Based Bilingual Pivoting Method for Paraphrase Extraction in Agglutinative Languages.” *BigComp*, 2016. **[AFNLP Best Asian Paper Award]**
27. Jong Myoung Kim, **Hancheol Park**, Gahgene Gweon, Jeong Hur. “The Correlation between Search Quality and Query Popularity.” *BigComp*, 2016.
28. **Hancheol Park**, Gahgene Gweon. “Initiating Moderation in Problematic Smartphone Usage Patterns.” *CHI Extended Abstracts*, 2015.
29. Jong Myoung Kim, **Hancheol Park**, Young-Seob Jeong, Ho-Jin Choi, Gahgene Gweon, Jeong Hur. “Measuring Popularity of Machine-Generated Sentences Using Term Count, Document Frequency, and Dependency Language Model.” *PACLIC*, 2015.
30. **Hancheol Park**, Gahgene Gweon, Ho-Jin Choi, Jeong Heo, Pum-Mo Ryu. “Sentential Paraphrase Generation for Agglutinative Languages Using SVM with a String Kernel.” *PACLIC*, 2014.
31. **Hancheol Park**, Gahgene Gweon, Ho-Jin Choi. “An Automatic Evaluation Metric for Korean Paraphrase via Semantic Frame.” *KIPS*, 2014.

Patents

1. **Hancheol Park**, “Method and System for Computationally Efficient High-Performance Object Detection,” KR Patent No. 10-2931775; Application No. 10-2022-0132409, Granted, 2026.
2. **Hancheol Park**, “Method and System for Compressing Natural Language Understanding Models via Layer Pruning,” KR Patent No. 10-2909777; Application No. 10-2022-0132410, Granted, 2026.
3. Geonmin Kim, Jaeyeon Kim, Wooksu Shin, **Hancheol Park**, “Technique for Reducing Hallucinated Answers from AI-Based Language Models,” KR Patent No. 10-2916056; Application No. 10-2024-0116795, Granted, 2026.

4. Jeongho Kim, **Hancheol Park**, “Method and Apparatus for Early Fire Detection,” KR Patent No. 10-2873458; Application No. 10-2023-0126929, Granted, 2025.
5. **Hancheol Park**, Geonmin Kim, “Method and Apparatus for Determining Ambiguity in an Input Prompt,” KR Patent No. 10-2841030; Application No. 10-2024-0127486, Granted, 2025.
6. **Hancheol Park**, “Method for Estimating Crowd Count, Method for Training a Model for Crowd Count Estimation, and Electronic Device for Performing the Same,” KR Patent No. 10-2779724; Application No. 10-2023-0145565, Granted, 2025.
7. Hyungjun Lee, **Hancheol Park**, “Method of Lightweighting a Neural Network for Object Recognition, Method of Recognizing an Object Using the Lightweighted Neural Network, and Electronic Device for Performing the Same,” US Patent No. 12,443,828; Application No. 18/932,549, Granted, 2025.
8. Hyungjun Lee, **Hancheol Park**, “Method of Lightweighting a Neural Network for Object Recognition, Method of Recognizing an Object Using the Lightweighted Neural Network, and Electronic Device for Performing the Same,” KR Patent No. 10-2740182; Application No. 10-2023-0154173, Granted, 2024.
9. **Hancheol Park**, Tae-Ho Kim, “Knowledge Distillation Method and System Specialized for Pruning-Based Deep Neural Network Compression,” KR Patent No. 10-2597184; Application No. 10-2022-7041865, Granted, 2023.
10. Jong C. Park, **Hancheol Park**, Heeje Lee, Wonsuk Yang, “Statement Reliability Evaluation System and Method Using Commonsense Knowledge and Linguistic Patterns,” KR Patent No. 10-2439165; Application No. 10-2020-0161606, Granted, 2022.
11. Jong C. Park, **Hancheol Park**, Jin-Woo Chung, Huije Lee, “System and Method for Constructing Emotion Lexicon by Paraphrasing and Recognizing Emotion Frames,” KR Patent No. 10-2398683; Application No. 10-2017-0106064, Granted, 2022.
12. Jong C. Park, **Hancheol Park**, Hoyun Song, Heeje Lee, “Method and System for Personality Recognition from Dialogues,” KR Patent No. 10-2319013; Application No. 10-2020-0011541, Granted, 2021.
13. Jong C. Park, Heeje Lee, **Hancheol Park**, Wonsuk Yang, “Apparatus for Detecting Non-standard Meaning Usage of Words, Method for Detecting Non-standard Meaning Usage of Words, and Recording Medium,” KR Patent No. 10-2204341; Application No. 10-2019-0025645, Granted, 2021.
14. Jong C. Park, Jinah Park, Jung-Ho Kim, Youngjin An, Wonsuk Yang, **Hancheol Park**, “System and Method for Communication Training Program over Virtual Reality and Continued Feedback via Mobile Device,” KR Patent No. 10-1998482; Application No. 10-2017-0154022, Granted, 2019.

Awards & Honors

Competition Results

- NVIDIA Nemotron Hackathon Seoul, Track C: Nemotron for Synthetic Data Generation (SDG), **1st Place in Track C and Overall Winner**, NVIDIA, 2026.
- GenAI Content Detection Task 1: Binary Multilingual Machine-Generated Text Detection, Subtask B: Multilingual MGT Detection, **3rd Place**, GenAIDetect Workshop, COLING, 2025.
- NVIDIA AI City Challenge 2024, Track 4: Road Object Detection in Fish-Eye Cameras, **2nd Place out of 53 teams**, CVPR, 2024.
- NVIDIA AI City Challenge 2024, Track 1: Multi-Camera People Tracking, **3rd Place out of 15 teams**, CVPR, 2024.
- NVIDIA AI City Challenge 2023, Track 1: Multi-Camera People Tracking, **Top 10**, CVPR, 2023.
- Visual Object Tracking 2023 Challenge, **2nd Place out of 47 teams in accuracy**, ICCV, 2023.

Awards

- **Best SK AI Partner**, SK Tech Summit, 2023.
- **Outstanding Paper Award**, Association for Computational Linguistics (ACL), 2023.
- **Best Presentation Award**, Korea Computer Congress (KCC), 2023.

- **Best Paper Award**, Annual Conference on Human & Cognitive Language Technology (HCLT), 2016.
- **Best Paper Award**, Korea Computer Congress (KCC), 2016.
- **AFNLP Best Asian Paper Award**, IEEE International Conference on Big Data and Smart Computing (BigComp), 2016.

Education

Korea Advanced Institute of Science and Technology (KAIST)

Daejeon, Republic of Korea

Ph.D. in Computer Science

Feb. 2024

Thesis: Capturing Ambiguity in Natural Language Understanding Tasks with Information from Internal Layers.

Advisor: Jong C. Park

Teaching & Mentoring Experience

- **AI Tech Mentor**, boostcamp AI Tech, NAVER Connect Foundation, 4th Cohort, 2023.
- **AI Tech Mentor**, boostcamp AI Tech, NAVER Connect Foundation, 3rd Cohort, 2022.
- **Teaching Assistant**, CS612: Social Network-Aware Ubiquitous Computing, KAIST, Spring 2017.
- **Teaching Assistant**, CS206: Data Structures, KAIST, Fall 2016.

Academic Service

Program Committee

- **International Conference on Computational Linguistics (COLING)**, Machine Learning Track — 2020.
- **The 1st International Workshop on Spatial/Temporal Information Extraction from Unstructured Texts (WSTIE)** — 2016.

Conference/Workshop Reviewing

- **AAAI Conference on Artificial Intelligence (AAAI)** — 2020, 2017.
- **International Joint Conference on Artificial Intelligence (IJCAI)** — 2019, 2018, 2017.
- **Pacific Asia Conference on Language, Information and Computation (PACLIC)** — 2019.
- **International Conference on Web Intelligence, Mining and Semantics (WIMS)** — 2016.
- **Louhi: Workshop on Health Text Mining and Information Analysis**, EMNLP Workshop — 2016.
- **Language and Information** — 2016.

Journal Reviewing

- **ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)** — 2016–2019.
- **Journal of Computing Science and Engineering (JCSE)** — 2016–2019.

Public Highlights

Media & Presentations

- [MTN News Broadcast Interview on Nota's AI Model Optimization Technology](#), 2026.
- [Two MoE Quantization Papers Accepted to ICML 2026 Workshop](#), 2026.
- [NVIDIA Nemotron Dev Days Seoul – YouTube Live Session / Hackathon Highlights](#), 2026.
- [NVIDIA Korea Blog – NVIDIA Developer Days Seoul 2026 Recap](#), 2026.
- [KAIST News – ACL 2023 Outstanding Paper Award for Low-Resource Tigrinya Question Answering](#), 2023.
- [NVIDIA GTC 2023 Posters Spotlight: Computer Vision](#), 2023.

Technical Writing

- [LLM Model Quantization Techniques for AWS Inferentia by Nota AI](#), AWS Technical Blog, 2026.

- [Two MoE Quantization Papers Accepted to an ICML 2026 Workshop](#), Nota AI, 2026.
- [NotaMoEQuantization: An MoE-Specific Quantization Method for Solar-Open-100B](#), Nota AI, 2026.
- [Unseen Language-Aware Detection System for Multilingual Machine-Generated Text](#), Nota AI, 2025.
- [Where Do LLMs Encode the Knowledge to Assess the Ambiguity?](#), Nota AI, 2025.
- [Deploying an Efficient Vision-Language Model on Mobile Devices](#), Nota AI, 2024.

Model Releases & Software

- [Solar-Open-100B-NotaMoEQuant-Int4](#), public quantized release of Upstage Solar-Open-100B with Int4 weight-only MoE-specific quantization, Hugging Face, 2026.
- [Solar-Open-100B-NotaMoEQuant-NVFP4](#), public quantized release of Upstage Solar-Open-100B with NVFP4 MoE-specific quantization, Hugging Face, 2026.