Tae-Hoon Yong

yongtaehoon@gmail.com | LinkedIn | GitHub

SUMMARY

AI Engineer specializing in large-scale LLM systems and agentic AI, with hands-on experience leading full-lifecycle development of production-grade AI agents in constrained enterprise environments. Designed and deployed large-scale multi-agent systems with hierarchical tool orchestration, improving latency and response reliability through iterative system optimization. Experienced in building scalable, secure AI services under real-world constraints and driving measurable improvements in system performance and user impact.

EXPERIENCE

KT — AI Engineer Seoul, South Korea | Apr 2025 – Present

- **Large-scale AI Agent for Financial Services (Full Lifecycle):** Led end-to-end design and deployment of a production LLM-based assistant in a 5-engineer team over 4 months, replacing manual document search and rule-based workflows for insurance agents (300+ CCU, 1,000 peak users).
- **End-to-End Optimization & Measurable Impact:** Improved retrieval quality and token efficiency by redesigning chunking strategies and optimizing agent context usage, resulting in user satisfaction improvement from 2.8 to 4.7 (5-point scale) and latency reduction from ~120s to ~50s.
- **Large-scale Agent Orchestration & LLM System Design:** Designed hierarchical multi-agent systems (ReAct, Plan-and-Execute) with master agent + 6 sub-agents (6–8 tools each), and built full-lifecycle LLM pipelines including intent classification, query refinement, planning, tool execution, and validation.
- **Retrieval & Hallucination Optimization:** Identified failure cases in naive chunking and improved retrieval precision using parent-child chunking and table-aware parsing; introduced multi-hop querying and validation workflows to improve answer reliability.
- **System Scalability & Production Stability:** Resolved ingestion bottlenecks caused by throttling via buffered transfer and adaptive scaling, and improved system robustness through data-driven iteration based on production feedback.

OSSTEM IMPLANT Co., Ltd. — AI Research Engineer / Team Leader Seoul, South Korea | Aug 2021 – Mar 2025

- **AI Team Leadership & End-to-End Ownership:** Led an AI engineering team of 8+ members, owning full lifecycle from research to production deployment across multiple commercial medical AI systems.
- **Production AI Systems in Real-world Environments:** Developed and deployed AI modules integrated into commercial dental software, optimized for on-device environments with strict performance constraints.

Selected Contributions

- Designed 3D morphable models (3DMM) and deformation pipelines for anatomically accurate dental modeling
- Developed deep learning models for detection, segmentation, and classification on 2D/3D medical imaging
- Applied GAN-based augmentation to improve robustness under real-world noise and artifacts
- Optimized models via pruning, quantization, and knowledge distillation, enabling real-time inference with TensorRT
- Built internal RAG-based knowledge assistant (~70K Q&A) to automate engineering support workflows
- Delivered AI features into multiple commercial products (OneGuide, One3, OneOrtho, OneClick, V-Ceph)

EDUCATION

Seoul National University — M.S. in Applied Bio-Engineering (AI / Computer Vision) Seoul, South Korea | 2019 – 2021

- Thesis: GAN-based Quantitative Cone-Beam CT for Bone Mineral Density

Hongik University — B.S. in Computer Engineering Seoul, South Korea | 2014 – 2019

PUBLICATIONS (SELECTED)

Journal Papers (SCI/SCIE)

- **T.H. Yong*** et al., QCBCT-NET for Bone Mineral Density Measurement from Quantitative Cone-beam CT, Scientific Reports, 2021.
- O. Kwon*, **T.H. Yong*** et al., Automatic diagnosis for cysts and tumors of both jaws on panoramic radiographs, DMFR, 2020.

International Conferences

- D. Lee*, **T.H. Yong** et al., Revolutionizing Dental Image Segmentation (Filter Pruning & KD), IEEE ICASSP, 2024.
- **T.H. Yong*** et al., Automatic Detection of Inferior Alveolar Nerve on CBCT, IPIU, 2023. (Oral presentation & Best paper award)
- H.G. Ahn*, **T.H. Yong*** et al., Deep high-resolution landmark detection, MICCAI Workshop, 2022.
- **T.H. Yong*** et al., Inferior Alveolar Nerve Detection on CBCT, IEEE EMBC, 2022.
- **T.H. Yong*** et al., Bone Age Assessment in Hand-wrist Radiography, ICDMFR, 2021. (Oral presentation & Best paper award)
- **T.H. Yong*** et al., Periodontitis detection and classification in panoramic radiographs, CARS, 2019.

CHALLENGES & AWARDS (SELECTED)

- MICCAI Grand Challenge (2023): 8th Prize — CBCT Segmentation Challenge
- Best Paper Award, IPIU 2023 — Inferior Alveolar Nerve Detection (Oral)
- MICCAI Grand Challenge (2022): 6th Prize — 3D Teeth Scan Segmentation and Labeling

TECHNICAL STACK (SELECTED)

- **Languages/DL:** Python, C++, PyTorch, TensorFlow, ONNX, TensorRT
- **AI/LLM:** LangChain, LangGraph, Azure OpenAI, Vector DB (Chroma, Azure AI Search)
- **Cloud/DevOps:** Azure, Docker, GitLab CI/CD