

Henry Kang

Research Scientist, LLM Agents, Memory & Retrieval

Los Angeles, CA, United States | [LinkedIn](#)

AI researcher with 8+ years taking models from research to production. Currently CTO of a character-chat startup with US\$1.5M trailing-12-month revenue, after leading a 10-person AI team at Ericsson.

EXPERIENCE

Chief Technology Officer

Jul 2026 — Present

Dololabs (character / companion AI)

Remote

Lead research and engineering for a 4-person character-chat startup, covering LLM selection and fine-tuning, conversational memory and persona, and real-time inference. Built short- and long-term memory benchmarks that now decide which model and memory changes ship. Raised KV and prompt cache hit rate from under 10% to 30% in the custom routing architecture, and secured about US\$200K in signed contracts.

AI Lecturer

Mar 2025 — Present

Kookmin University Global Program at Westcliff University

Irvine, California, United States

Lead applied research on on-device small language models, cutting mobile inference latency from 14s to 3s (a 79% reduction) through quantization, layer pruning and runtime tuning. This work led to a first-author EMNLP 2026 main-conference paper on RAG with frozen SLMs, plus 3 papers submitted as corresponding author. Teach LLMs and applied AI to international students and mentor student teams building AI products.

Chief Technology Officer

Jan 2025 — Jul 2026

UStechlab

Publishing independent AI research paper on mobile edge AI under constrained compute environments. Secured Google IT project support; authored and published AI e-books for the Korean market. Build the AI based Game.

Technical Advisor (Contract)

Jul 2024 — Sep 2024

GetPickle (YC W25)

San Francisco, CA

Designed the AI architecture for a real-time conversational avatar product, including the LLM reasoning loop, fine-tuning and evaluation. Guided 4 junior scientists through data curation, fine-tuning, eval design and error analysis. The team was accepted to Y Combinator W25 and raised a \$700K seed.

Applied Research Scientist (L7)

Feb 2023 — Jul 2024

Ericsson

Toronto, Ontario, Canada

Led a 10-person AI team moving generative-model research into product, covering LLM fine-tuning, deepfake generation and detection, and super-resolution. Built evaluation frameworks combining automatic metrics, LLM-as-judge and human review. Curated a 20TB proprietary multimodal dataset with an auto-labeling pipeline, which improved generalization to out-of-distribution data.

University Ambassdor & Instructor

May 2022 — Feb 2023

NVIDIA

Toronto, ON

Delivered specialized workshops on AI and machine learning techniques in Hanyang girl's university, York university. Covered topics such as fundmental deeplearning and computer vision and practical applications in various domains. Engaged with participants through hands-on sessions, case studies, and interactive discussions. Contributed to advancing knowledge and skills in the AI community.

Applied Machine Learning Scientist

Jan 2022 — Jan 2023

Qii.AI

Toronto, Ontario, Canada

Led AI delivery on 6+ government and industry projects, including auto-labeling, semi-supervised learning and defect segmentation, from client requirements through deployed models. Won the Royal Canadian Navy Fleet Inspection Sandbox with a live-deployed inspection system. Shipped vision models from cloud to embedded boards.

Artificial Intelligence Researcher

Jan 2022 — Jan 2023

Unknown

Spearheaded 6+ government and industry AI projects covering auto-labeling, segmentation, and cross-domain applications. Won Canada Navy Fleet Inspection Sandbox competition out of numerous competing teams. Delivered AI solutions integrating IoT, robotics, and civil engineering domains.

Applied Machine Learning Scientist

Oct 2021 — Jan 2022

T2D2.ai

New York, United States

- Fast-paced, dynamic innovator and experience with machine learning research platforms including Opencv and Pytorch deep learning platform leading to the development of machine learning based structure management system.
- Ability to identifying process pain points with an ability to create automated data analysis process resulting as evidenced by the innovation of high-throughput experimental procedures in structure management software field.
- Ability to manage project timelines with an ability to produce time-dependent deliverables for structure management project deliverables resulting in successful collaborations from in public sector or industry sector field.

Research Assistant

Sep 2018 — Apr 2019

WorldQuant

Canada

Worldquant provide me very good education every weekend about risk management and financial management by machine learning algorithm.

It is remote work and I learn about many different aspective about pro and con about remote work.

After this I got some confidence about control my self, time management as soft skill.

As the technical skill, I realize that database management is very important for machine learning in real world scenario.

Research Assistant (Ph.D.)

Aug 2016 — Nov 2021

University of Manitoba

Winnipeg, Canada Area

Researched deep learning and autonomous UAVs for infrastructure inspection. Produced 5 SCI(E) journal papers and 1

U.S. patent application, and this body of work has 1,600+ citations.

Squad Leader/ Sergeant

Apr 2009 — Feb 2011

Republic of Korea Army

Seoul, South Korea

I do my mandatory army service in Korea.

After first year, I also manage my squad as squad leader.

During that time, I raise my project management and team management skill.

EDUCATION

University of Manitoba

Sep 2016 — Jun 2022

Doctor of Philosophy - PhD in Civil engineering

Thesis: AI and Robotics for Infrastructure Management

Graduate school of Soongsil University

Jan 2014 — Jun 2016

Master's degree in Software

Top Student of the Year

Soongsil University

Mar 2007 — Jun 2014

*Bachelor's degree in Business Administration and Management, General***Georgia State University - J. Mack Robinson College of Business***IT Project Business Management Certificate***PROJECTS**

Small AI memory and research

Research the AI research for ondevice or efficient AI technique.

Digital human project

- Design and implement an AI algorithm that will serve as the foundation for the Digital Human Project.
- Utilize cutting-edge AI techniques to achieve a high degree of realism and believability in the digital human.
- Gather and analyze extensive data on human behavior, emotions, and cognitive processes to inform algorithm development.

Fleet inspection using by AI and robotics

- Develop algorithms and models to enable the drone to autonomously identify and classify rust patterns on ship surfaces.
- Collaborate with a multidisciplinary team to define requirements and specifications for the AI system.
- Implement real-time monitoring capabilities to provide instant feedback on the rust condition.

Automatic structure inspection using by A.I.

Do the project management and machine learning engineer role.

Accomplish the two projects about two client companies from Newyork and British.

Provide the all of deliveralbe on time.

UAV with Ultrasonic Beacon and laser sensor for structural health monitoring

We are using the Ultrasonic beacon and laser sensor for autonomous navigation about UAV.

This is very complex and small place but very accruate walk. We almost finish the work.

Tire inspection using by deep learning

Data handling, network modeling.

Work with supervisor Dr.Cha.

Interesting project but also face lots of limitation.

Study helper

This project is the 2020 pytorch hackathon result. The speech to text algorithm is applied using by Google API. The text summary algorithm is applied with BERT based pretraiend model.

We try to make the web User interface for recording and get the result.

Alrington Bridge SHM

The steel structural damage is monitoring in real-time with a manually built solar camera system, LTE modem. Thanks for the City of Winnipeg to help and support this project.

UAV with Ultrasonic Beacon for structural health monitoring

I developed the whole system.

CERTIFICATIONS

Certificate of Completion: AI Fluency Framework & Foundations , Anthropic	Jun 2026
Advanced Idea start up competition , Artificial intelligence industry cluster agency	Jun 2023
Computer Vision for Industrial Inspection , NVIDIA	Sep 2022
Instructor Materials - Computer Vision for Industrial Inspection , NVIDIA	Sep 2022
AI UNIVERSITY AMBASSADOR , NVIDIA	Apr 2022
Fundamentals of Deep Learning (Instructor Materials) , NVIDIA	Apr 2022
NVIDIA DLI Certificate – Fundamentals of Deep Learning , NVIDIA	Mar 2022
Nvidia DLI certificate , NVIDIA	Mar 2022
Data Manipulation with pandas , DataCamp	Aug 2021
Joining Data in SQL , DataCamp	Aug 2021
Augmented Reality 101 by PyImageSearch University Certification , PyImageSearch	Jul 2021
Autonomous Vehicle Fundamentals , PyImageSearch	Jul 2021
OpenCV 101 — OpenCV Basics , PyImageSearch	Jul 2021
Triplebyte Certified Machine Learning Engineer , Triplebyte	Jul 2021
Introduction to SQL , DataCamp	May 2021
ISS , University of Manitoba	Sep 2017
UAV knowledge test , Transport Canada - Transports Canada	Aug 2017
IGSS , University of Manitoba	Sep 2016
Embedded software contest , Qualcomm	Dec 2014
Office executive program , Georgia State University - J. Mack Robinson College of Business	Aug 2014
Deeplearning nanodegree , Udacity	
Flying Car nanodegree , Udacity	
Pocket RAG: On-Device RAG for First Aid Guidance in Offline Mobile Environment ,	
Deep learning-based obstacle-avoiding autonomous UAVs with fiducial marker-based localization for structural health monitoring ,	
Efficient attention-based deep encoder and decoder for crack segmentation ,	
Hybrid Concrete Crack Segmentation and Quantification Across Complex Backgrounds Without a Large Training Dataset ,	
Real-time multiple damage mapping using autonomous UAV and deep faster region-based neural networks for GPS-denied structures ,	
Hybrid pixel-level concrete crack segmentation and quantification across complex backgrounds using deep learning ,	
Autonomous UAVs for Structural Health Monitoring Using Deep Learning and an Ultrasonic Beacon System with Geo-Tagging ,	
Personalization and mobile characteristics affecting smart learning ,	
Emergency Landing Zone Selecting Model for UAV ,	

SKILLS

Artificial Intelligence (AI) · IT Project+ · Deep Packet Inspection · Team Mentoring · 2D Animation · Model Making · Cross-Functional Team Management · Google API · Large Scale Deployments · Hardware Deployments · 20/20 Technologies · Data Curation · Digital Publishing · Patent Preparation · Scientific Reports · DeepLearning4j · Fleet Maintenance · Team Building · Generalized Linear Models · Manufacturing Modeling · Computational Intelligence · Post Processing · Large Language Models (LLM) · Cross-platform Development · Image Synthesis · Super-resolution · Conversational AI · Computer Science · Embedded Software Programming · Expression Cloning · Robotics · Deep Learning · Artificial Intelligence · Computer Vision · Critical Thinking · Lean Startup · Startup Development · Applied Technology · Machine Learning · Generative AI · Python (Programming Language) · Civil Engineering · Project Management · Statistics · Image Processing · Data Analysis · Data Science · Research · Scientific Writing · conversational memory · RAG · retrieval and grounding · tool use · multi-step agent pipelines · context design · benchmark design · golden sets · LLM-as-judge · human review · error analysis · regression tracking · fine-tuning (SFT, LoRA) · SLM compression · quantization · inference optimization · caching · PyTorch · Hugging Face Transformers · vLLM · OpenAI API · Anthropic API · Docker · AWS · Git