

Khanh-Tung Tran

📍 Cork, Ireland ✉ tktung1805@gmail.com 🌐 khanhtungtran.github.io

Summary

PhD Researcher in Computer Science specializing in large language models, data-efficient language adaptation and alignment, cross-linguality and multi-linguality, multi-agent systems.

Research interests include efficient LLMs, adaptation, mechanistic interpretability, alignment, agentic AI, and RAG.

5 years of experience in AI research and engineering across LLMs, multi-agent systems, reinforcement learning, GPU/HPC infrastructure, Docker-based deployment, and production AI systems.

Education

- | | | |
|------------|---|---------------------------|
| PhD | University College Cork , Computer Science | Cork, Ireland |
| | <ul style="list-style-type: none">Supervisor: Assoc. Prof. Harry Nguyen & Prof. Barry O’Sullivan.Thesis: Data-Efficient Cross-Lingual Understanding, Generation, and Alignment in Large Language Models. | Jan 2024 – Jan 2028 |
| BSc | Ho Chi Minh City University of Technology , Computer Science | Ho Chi Minh City, Vietnam |
| | <ul style="list-style-type: none">Major: Computer Science. Honors Program. GPA: 9.33/10, top 5/400.Thesis: Development of an Expressive Neural Text-to-Speech System. | Oct 2018 – Nov 2022 |

Work Experience

- | | |
|---|-----------------------|
| CloudCIX Ltd , AI Architect Consultant | Ireland |
| <ul style="list-style-type: none">Developed and deployed multi-tenant multi-agent chatbot systems for LLM-based RAG and document processing pipelines.Built development pipelines for continuous integration of LLMs to new domain and knowledge.Implemented CI/CD pipelines for model deployment and monitoring using Docker, Kubernetes, FastAPI, PostgreSQL, and OpenTelemetry on multi-node multi-gpu cluster with 2x8 B200 GPUs. | June 2024 – June 2026 |
| Umeå University , Research Engineer | Sweden |
| <ul style="list-style-type: none">Built and administered Slurm-based HPC infrastructure, Linux modules, Docker, and enroot containers.Deployed GPU workload environments for model training, inference, job scheduling, monitoring, logging, and alerting.Engineered a 3D Digital Avatar with NVIDIA Omniverse and NVIDIA Riva, integrating facial animation, ASR, TTS, and conversational AI. | Aug 2023 – Jan 2024 |
| FPT Software , AI Researcher | Vietnam |
| <ul style="list-style-type: none">Researched deep learning models for speech, audio, and time-series tasks using PyTorch, Transformers, torchaudio, librosa, Lightning, Hydra.Built experiments for neural TTS, representation learning, hyperparameter tuning, and inference optimization on HPC/GPU clusters. | May 2021 – July 2023 |

Projects

UCCIX

Large language model trained on Irish language data, achieving state-of-the-art performance on Irish NLP benchmarks.

- Developed data collection and preprocessing pipelines for Irish language corpora.
- Trained and fine-tuned LLMs using PyTorch and Transformers on HPC infrastructure.
- Evaluated model performance on a suite of Irish NLP tasks, demonstrating significant improvements over previous models.

Guiden

Multi-tenanted chatbot system deployed on private cloud infrastructure.

- Authentication and SSO using LDAP.
- LLM and embedding model deployment and serving using vLLM, Kubernetes, FastAPI, nginx, and BentoML.
- Tenant knowledge integration through RAG system based on PostgreSQL.
- Built CI/CD pipelines for deployment automation.
- Used py-spy to debug latency issues and reduced inference latency by 20%.

ML-Workbench

RAG-based document processing and retrieval system.

- Built ingestion workflows (scraping, chunking, embedding generation)
- Used PostgreSQL (tsvector/pgvector) for vector similarity search.
- Orchestrated workflows using Hatchet for durable pipeline execution.

Publications

Disentangling Continued Pre-training: Attention-Driven Routing and Semantic Hub Preservation in Language Adaptation	ACL 2026
<i>Tran, K. T., Tran, V. K., O'Sullivan, B., Nguyen, H. D.</i>	
LaCoMSA: Language-Consistency Multilingual Self-Alignment with Latent Representation Rewarding	EACL 2026
<i>Tran, K. T., O'Sullivan, B., Nguyen, H. D.</i>	
IRLBench: A Multi-modal, Culturally Grounded, Parallel Irish-English Benchmark for Open-Ended LLM Reasoning Evaluation	KDD 2026
<i>Tran, K. T., Nguyen, D. H., O'Sullivan, B., Nguyen, H. D.</i>	
Reasoning Transfer for an Extremely Low-Resource and Endangered Language: Bridging Languages Through Sample-Efficient Language Understanding	AAAI-26
<i>Tran, K. T., O'Sullivan, B., Nguyen, H. D.</i>	
Disentangling Language Understanding and Reasoning Structures in Cross-lingual Chain-of-Thought Prompting	EMNLP 2025
<i>Tran, K. T., Vu, H., O'Sullivan, B., Nguyen, H. D.</i>	
UCCIX: Irish-eXcellence Large Language Model	ECAI 2024
<i>Tran, K. T., O'Sullivan, B., Nguyen, H. D.</i>	

Achievements

- 30 Under 30 in Ireland 2026, The Business Post.
- Best Paper Award, LoResLM Workshop, EACL 2026.

Technical Skills

AI / Data: LLM adaptation (continued pre-training, mechanistic interpretability, neuron intervention) and alignment (post-training, reinforcement learning, preference optimization, latent representation steering), agentic AI (multi-agent, agentic memory, RAG, MCP)

Deep Learning Framework: PyTorch, Transformers, TRL, vLLM, librosa, torchaudio, Kaldi/ESPnet/SpeechBrain, TensorFlow, LangGraph, smolagents, ReAct

Software Engineering: Python, C/C++, Bash, REST APIs, distributed systems concepts, asynchronous systems (RabbitMQ, Celery)

API Serving & Web Deployment: Django, FastAPI, NetworkX, Streamlit, PostgreSQL (tsvector/pgvector), Prometheus, Grafana, Elasticsearch, Sentry, OpenTelemetry, Docker, Docker Swarm