

Duy Tran Thanh

Staff AI Engineer

bradduy95@gmail.com | github.com/bradduy | linkedin.com/in/duytranthanh | +84-862-570-012

SUMMARY

Ph.D. in AI, **9+ years** across applied AI research, production GenAI, and enterprise systems. Hands-on across the LLM stack – **fine-tuning (SFT, GRPO, distillation)**, **RAG**, and **LLM evaluation** – and **multi-agent systems** with planning, tool use, memory, and human-in-the-loop. Shipped AI at scale (**200M+ tokens/day** LLM serving, banking agents under contractual KPIs); set technical direction across Engineering, Product, and Design, and lead research from paper to production.

EXPERIENCE

Staff AI Engineer & Program Manager

Jan. 2026 – Present

OneMount – AI Platform OneNexus

Hanoi, Vietnam

- **Technical Leadership:** Set AI technical direction with the CTO, Directors, and Heads of Engineering across **6 domains** – define scope, success metrics, and an evaluation plan for each AI initiative, with a clear **commit-or-decline decision** on what to build; choose **agentic / RAG / fine-tuning** per use case based on evals, latency, cost, and failure modes.
- **Program Management:** Plan OneMount's **AI project portfolio** – roadmap, milestones, and delivery tracking; own **resource planning and staffing**, matching engineers to projects by skill and priority, and set **product direction** with Product and business owners, sequencing work by business impact to keep delivery **efficient**.
- **RemittanceAI & Agentic AML – AI Review of Cross-Border Import Payments, Techcombank:**
 - * Built the core classification and AML-screening services (LangGraph agents, multimodal RAG over a regulatory store), choosing LLM agents over trained classifiers because labels were few. Processing fell from **17.5 to 3 min per transaction** (–83%) at up to **1,200 document sets/day**; both MVPs shipped **2 weeks early**.
 - * Built an eval harness that tunes prompts and code on banker-labeled ground truth, then passed **6/6 contractual AI KPIs** on ~8,000 validated transactions: **96.5% decision accuracy**, **99.99% key-field extraction**, **p95 4.9 min** vs. 5-min target. AML flags cite the breached regulation (**100,000+ entities** screened); staff check only key fields and flags.
 - * Architected the **LLM serving platform** shared by RemittanceAI, Dora, and OneHiring: a LiteLLM gateway routes **200M+ tokens/day** (**500K+ requests/week**, ~2s median) to self-hosted vLLM, SGLang, and Triton on 5×H200 and 6×AMD MI355X, plus Gemini. Rules-first cascades settle most decisions before any LLM call, cutting token spend **30%**.
- **Foundation Model for Personalized Recommendation – Techcombank, TechcomLife, Techcombank OneU:**
 - * Led research-to-production of a **VLM-backed recommendation ranker** over Techcombank's catalog – insurance, healthcare, vouchers, campaigns – with **two-phase post-training** (banking-domain adaptation, then ranking fine-tuning), a **catalog-aware ranking head** that scores only real products, and a **reward-weighted ranking loss** aligning recommendations with business goals.
 - * Replaced hand-engineered features with **context engineering**: verbalized transaction histories plus campaign text and creatives, pruned by a **feature-selection algorithm** that keeps high-signal events and compresses repeats within the token budget; **sliding-window training** learns long-term behavior at **no extra inference cost**, served **prefill-only on vLLM**.
- **Dora – Multi-Agent Reasoning Platform for End-to-End Document Verification – OneMount:**
 - * **Evidence-backed verification** (Techcombank pilot; insurance, credit, customs, tax): agents cite evidence, **deterministic rules decide pass/fail**. Set the technical direction and built the **Spec Agent**, replacing the per-customer Product Owner and developer loop: **hybrid RAG** over briefs, policies, samples, and scans finds **gaps and contradictions**, prefills only quote-backed answers, asks ≤5 plain-language questions; a Generator agent compiles it into a pipeline.
 - * **Multi-agent orchestration on LangGraph:** 5 roles (Spec, Generator, Metadata Engineer, **Judge**, Copilot) on one runtime – a **new role is configuration, not code**; **plan-and-execute** with plans checked by 8 deterministic laws, governed **MCP tools**, memory, **human-in-the-loop** interrupt/resume, tenant isolation.
 - * **Evaluation and fine-tuning:** benchmarked on **5 public human-labeled datasets** – **96.6% auto-fill precision**, **97.5% Hit@10** on Vietnamese legal retrieval, **100% of 34,824 spreadsheet rows read**; **SFT + GRPO distillation** into a **self-hosted model** behind LiteLLM. Shipped on **Kubernetes** with **56 invariants** and **3,300+ tests** in CI; authored 10 of 30 ADRs, aligning Product, Design, and the Java/Kotlin team on **schema-first contracts**.
- **OneHiring & Company AI Agent Platform – OneMount:**
 - * Owned the **reference architecture** for a no-code agentic studio on Dify (**build-vs-buy**: extend open source over in-house) – **4 agents across source-to-hire**, sourcing autonomous, hiring decisions **human-in-the-loop**; **RBAC/ACL by role**, PII redaction, versioned **MCP/A2A capability contract** so new domains onboard without runtime changes. Blueprint **6 engineers** build against.
 - * Designed the **RAG layer on Qdrant** – hybrid retrieval (dense + BM25) with cross-encoder reranking: **98% groundedness on a 180-query labeled set**, every answer cited; **per-domain collections**, **tenant scope pre-filtered inside vector search** – no cross-domain leakage, hard-delete propagation for candidate erasure.

- **Marketing Brand Brain – Masterise Group (MAG):**

- * Decomposed the CMO’s marketing workflow into **agent skills** with typed I/O contracts, **eval sets**, and a **human-oversight boundary** per step; made the **build-vs-buy** call on evidence – benchmarked **self-hosted vLLM against vendor LLM/VLM APIs** on accuracy, **p95 latency**, and **TCO per validated submission**: metered APIs over GPU capex below break-even.
- * Designed the **reference architecture**: a **governed policy knowledge base** compiled into versioned executable rules, **per-rule orchestration** over **multimodal RAG** (policy text with design-system images for **VLM** brand conformance), and an approval workflow on full **audit logging**; one **model gateway** owns routing, schema-validated output and provider fallback, with **prompt caching** for cost control. **Findings cite the rule version**; named approver, **no autonomous publishing**.

Research Scientist

Jan. 2025 – Jan. 2026

Physical AI Lab, Korea Evaluation Institute of Industrial Technology

Seoul, South Korea

- **Advisory for Industry Partners:** Advised hospital and healthcare partners on **privacy-preserving federated learning** (Substra) and embodied-AI adoption – robot-platform feasibility, perception and VLA stack selection, sim-to-real risk on ROS 2 / Isaac Sim – with a build-or-decline recommendation per use case.
- **Production Model Optimization:** Cut inference cost $\sim 2\times$ training-free via sparse MoE expert routing; quantized PyTorch vision models to INT8/INT4 and served them with TensorRT on Jetson Orin NX (**8 $\times$ smaller, <1% accuracy loss**), with custom CUDA kernels on the hot path. Delivered **3 $\times$ classification accuracy over CLIP** on real-world agricultural imagery and state-of-the-art open-vocabulary 3D scene understanding (+3% over the 2025 ICML baseline).

Research Scientist

Sep. 2020 – Jan. 2025

Machine Vision Lab, National Research Foundation of Korea

Seoul, South Korea

- **Open-vocabulary Segmentation:** First-author research with vision-language foundation models – automated knowledge-distillation pipeline on MMSegmentation: frozen VLM teachers generate multimodal pseudo-labels as weak supervision for zero-shot recognition of unseen classes with **zero manual annotation, +22.1 mAP over prior state-of-the-art**. Trained multi-GPU (DDP, mixed precision) with W&B tracking; served real-time inference monitored for latency and drift.
- **CC-Person re-ID:** Designed **PAH-Net** (*Neurocomputing*, Q1) for clothes-changing person re-ID – state-of-the-art on three benchmarks (**+7.3 rank-1** on PRCC cross-clothes); **architecture alone reached 107 FPS at 32.6M parameters**, up to $8\times$ fewer than competing generative models. Deployed on NVIDIA DeepStream with multi-camera RTSP ingest and Kafka streaming, matching identities by embedding similarity against a gallery.
- **ROI Object Detection:** Adapted an **MMDetection** detector into a real-time soccer ROI detector for sports broadcasting, **exceeding contractual KPIs** for a government-industry consortium; engaged the industry client directly to align architecture with broadcasting requirements.

Senior Software Engineer

Sep. 2018 – Aug. 2020

Robert Bosch GmbH

Ho Chi Minh City, Vietnam

- **Enterprise B2B Architecture:** Designed systems from scratch to production – service decomposition, REST/OpenAPI contract design, data modeling across Oracle/MongoDB/event-driven Kafka messaging – as Java Spring Boot / Spring Cloud microservices; owned authentication (JWT), load balancing, and Jenkins CI/CD for a global HR portal serving **10,000+ users** at $\sim 1,000$ concurrent requests peak.
- **Distributed Back End & Stakeholder Engagement:** Architected Bosch’s company-wide internal social network for a $\sim 400,000$ -employee organization – horizontal scaling on Docker/Kubernetes, Redis caching, asynchronous Kafka. Presented technical proposals cross-region to German and Vietnamese stakeholders; core code reviewer; awarded **Best Performance 3 times**.

Software Engineer

Jun. 2017 – Aug. 2018

TMA Solutions

Ho Chi Minh City, Vietnam

- Full-cycle delivery (Angular + Java) for production client applications.

COMMUNITY LEADERSHIP & MENTORSHIP

AI Coach – AI20K in Action Program - QS Reimagine Education Awards 2026

Jul. 2026 – Present

Vingroup – VinUniversity

Hanoi, Vietnam

- **Cohort 3 – Security:** Co-mentored with the **VinSOC Head of AI Solutions** on AI for cybersecurity – **agents and guardrails for SOC and SIEM** and **anti-money-laundering (AML)**.
- **Cohort 4 – Banking & Fintech:** **Main coach** of the track, guiding teams building **multi-agent systems** for tax, AI secretary, CAD drawing extraction and cost estimation for civil engineering, and **fraud detection for ride-hailing transactions**.
- Across a national program of **22,000 students, personally coach 80 mentees** with an architecture-first curriculum – requirement discovery, build-vs-buy and agentic-vs-deterministic trade-offs, non-functional design, and reviews that take prototypes to production with evals and quality gates. **Appointed Committee Member** of AI in Action, Vingroup’s AI community, setting the program’s technical review bar.

AI Expert Trainer

Jan. 2026 – Present

OneMount

Hanoi, Vietnam

- Appointed **one of 4 company-wide AI Experts** – train engineering and business teams on prompt and context engineering and agent evals, and coach applied AI engineers. Raised engineering delivery **3× at constant headcount across 7 disciplines** via agent harnesses, MCP integrations, and preset Claude/Gemini/Codex configurations, with per-use-case token-cost models.

EDUCATION

Chung-Ang University

Seoul, South Korea

Visitor – Physical AI

2025

Seoul National University of Science and Technology

Seoul, South Korea

Ph.D. in Artificial Intelligence

2020 – 2025

HCMC University of Technology and Education

Ho Chi Minh City, Vietnam

B.S. in Engineering (Honors)

2016

PUBLICATIONS

AGENTIC AND MULTI-AGENT SYSTEMS

- (*IEEE Access*, 2026) **Adjudicated Captioning: Multi-Agent Alignment Scoring and Consensus-Distilled Beam Arbitration for Strict Zero-Shot Image Captioning**: multi-agent system verifying generated text against the image in real time to eliminate decoder drift – **+8% caption accuracy**, no labeled data.
- (*Preprint*, 2026) **The AI-First SDLC: Harnessing Autonomous Agents for Enterprise-Grade Quality Control**: a playbook for enterprise leaders adopting AI agents across the SDLC, with a 30/60/90-day roadmap.

MULTIMODAL AND PHYSICAL AI PERCEPTION

- (*Neurocomputing*, 2026) **Learning from Multimodal Pseudo-Labels for Robust Open-Vocabulary Instance and Panoptic Segmentation**: distills vision-language foundation models into automated pseudo-labels at training time – open-vocabulary segmentation at **zero runtime overhead**.
- (*Neurocomputing*, 2024) **Enhancing Long-term Person Re-identification using Global, Local Body Part, and Head Streams**: cross-camera person tracking robust to clothing change – state of the art on three benchmarks with **no manual labeling**.

AI COST GOVERNANCE (FINOPS)

- (*Preprint*, 2026) **Edges Before Embeddings: A Confidence-Aware Blur Gate for Vision-Language Pipelines**: token-cost control blocking GPU and API spend on unreadable documents before any heavy model runs.

AWARDS AND CERTIFICATIONS

SDLC Agents – OpenAI Codex Contributor

June 2026

A Codex plugin harness for Jira and Confluence workflows through Atlassian MCP

Generative AI Leader – Google Cloud

May 2026

Business-leadership credential for driving responsible gen-AI transformation across enterprise functions

Wiki-Knowledge-Compiler – Anthropic Claude Plugin Publisher

Apr. 2026

Published a document-organization architecture for LLM Wiki and Agentic systems

OPEN SOURCE AND INDUSTRY CONTRIBUTIONS

MistralCoreML Contribution

Apple

Open-source PR merged into Apple’s MistralCoreML pipeline for Mistral models on Apple Silicon.

Banhmi-Gemma4-E4B

Google DeepMind Hackathon

Live on-device translated subtitles in 16 languages for the 300M+ people worldwide living outside their native language.

TECHNICAL SKILLS

Languages and Frameworks: Python, CUDA, Java (Spring Boot), SQL, PyTorch, Hugging Face, LangGraph, FastAPI, Pydantic
Machine Learning, Agentic and GenAI: Multi-agent orchestration, function calling, MCP, A2A, agent harnesses, agent evals, LLM-as-judge, guardrails, multimodal RAG, hybrid retrieval (BM25 + dense), reranking, RAG evals, prompt and context engineering, prompt management (Langfuse), human-in-the-loop design, AI-assisted SDLC, fine-tuning (SFT, GRPO, LoRA, QLoRA, distillation), quantization, document AI (OCR, KIE), VLM, VLA models, recommender systems, Dify
Model Serving and Infrastructure: vLLM, SGLang, LiteLLM, Triton, TensorRT, NVIDIA H200, AMD MI355X, MLOps, token-cost optimization, GCP (GKE, Cloud SQL, GCS)
Architecture and Integration: Microservices, event-driven design, Temporal, Kafka, REST/gRPC, SaaS multi-tenancy, PostgreSQL, Qdrant, Kubernetes, Docker, CI/CD, observability
Security and Governance: OWASP LLM Top 10, prompt-injection defense, access control (RBAC/ACL), agent sandboxing, federated learning (Substra), data privacy, AI quality KPIs, banking compliance