Dung (June) Thai, Ph.D.

San Jose, CA · +1-669-233-3660 · thainjune@gmail.com

linkedin.com/in/dung-thai · github.com/dungtn · Google Scholar

Summary

Research Scientist with 8+ years building NLP and generative AI systems, from foundational research to production deployment in healthcare. PhD in Computer Science (UMass Amherst, advised by Andrew McCallum). Deep expertise in large language models (LLMs), retrieval-augmented generation (RAG), agentic AI systems, hallucination detection & mitigation, and clinical AI. Particular focus on **trustworthy and safe AI** — building systems that are factually grounded, clinically reliable, and evaluable. 15+ publications at EMNLP, NAACL, KDD, ICML; 2 US patents. Proven record of taking research from formulation to production — including two deployed clinical AI systems (virtual utilization review and insurance appeal generation) at Ensemble Health Partners and Mendel AI. Experienced reviewer at NeurIPS, ICLR, SIGIR, and ARR.

Technical Skills

- **ML / AI:** Large Language Models, Agentic AI Systems, RAG, Hallucination Detection & Mitigation, RLHF, DPO, LoRA / QLoRA / PEFT, LLM Pre-training & Fine-tuning, Representation Learning, Sequence Labeling, Knowledge Graphs
- **Frameworks & Tools:** PyTorch, HuggingFace Transformers, vLLM, Unsloth, TRL, LangChain, Weights & Biases, Ray, Docker, Git, Cursor, GitHub Copilot
- **Clinical AI:** Clinical NLP, Medical Text Summarization, Hallucination Mitigation in Healthcare, Clinical Trial Matching, Revenue Cycle Management (RCM), MIMIC-IV, HealthBench
- **Languages:** Python (primary), Java, C++, R

Experience

- **Ensemble Health Partners** *July 2024 – Present*
  Staff Research Scientist San Jose, CA
  - **Virtual Utilization Review (VUR):** Designed and built a novel clinical AI system from scratch for real-time utilization review — a task with no public benchmark; defined task formulation, evaluation criteria, and modeling approach end-to-end. Deployed to production via Azure ML and Function App (eng team handled service bus and job queue). Iterated through error analysis to improve pipeline F1 by **9.54%** over internal baseline.
  - **Insurance Appeal Generation:** Built LLM pipelines combining RAG, self-refinement, and DPO to generate appeal letters grounded in clinical evidence, ICD/CPT coding, and payer-specific justification requirements; applied hallucination mitigation to reduce factual errors in generated clinical content. Distinct from VUR in operating over longitudinal patient records with a focus on cost efficiency per letter rather than real-time responsiveness.
- **Mendel AI** *Nov 2023 – July 2024*
  Senior Research Scientist → Staff Research Scientist San Jose, CA
  - **Hallucination Detection & Mitigation:** Built a hallucination detection framework outperforming prior state of the art by 2.3–4.8%; used detection signals to drive LLM self-refinement and preference learning (DPO), achieving end-to-end mitigation of factual errors in clinical summarization.
  - **Clinical Text Summarization:** Developed a semi-parametric memory mechanism allowing LLMs to reason across longitudinal patient records beyond context-window limits, specifically targeting reconciliation of conflicting medical events across time.
  - **Clinical Trial Matching (ACR benchmark, BioKDD 2024):** Co-developed a neuro-symbolic hybrid pipeline for large-scale cohort retrieval; benchmarked LLM baselines (including GPT-4) against the hybrid system on 1,400 patients across 113 complex oncology queries. Hybrid approach outperformed pure LLM baselines by **10.1–26.7% F1** (e.g., 62.9 vs. GPT-4's 20.8 F1 on one benchmark), demonstrating superior precision and scalability for real-world clinical trial matching.
  - Conducted continued pre-training of Llama 3 (8B and 70B) on proprietary clinical corpora; fine-tuned for downstream medical summarization and clinical event extraction tasks.
- **University of Massachusetts Amherst — IESL Group** *Aug 2016 – Nov 2023*
  Research Assistant (Ph.D.) Amherst, MA
  - **Case-Based Reasoning for NLP:** Introduced CBR-iKB, the first non-parametric CBR framework for knowledge-base QA — surpassed prior SOTA by **22.3% on WebQSP**; extended to unstructured text (CBR-MRC, EMNLP 2023) for interpretable machine reading comprehension, outperforming baselines by **+11.5 EM on NaturalQuestions** and **+8.4 EM on NewsQA**.
  - **Tabular Representation Learning (TABBIE, NAACL 2021):** Co-developed a dual-Transformer model for tabular data using an ELECTRA-inspired corrupt-cell detection objective; achieved state-of-the-art on column population (MAP 37.9 vs. TaBERT's 33.1), row population, and column type prediction, while requiring **10× less compute** than TaBERT (8 vs. 128 V100 GPUs).
  - **Internships:** Adobe Research (2017, 2018) — tabular QA and document understanding; IBM Research (2020, 2021) — knowledge-base QA and semantic parsing.

Independent Projects

- **AppealGen: Agentic Workflow for Grounded Clinical Appeal Generation** *2025*
  Independent project | *Python, vLLM, Google MedGemma, MIMIC-IV, HealthBench* [][]
  - Built during maternity leave as a submission to the **Kaggle MedGemma Impact Challenge**; open-source **agentic** end-to-end toolkit for generating clinically grounded insurance appeal letters using Google MedGemma — implementing a multi-step agentic workflow for evidence retrieval, clinical reasoning, and structured letter generation.
  - Designed a HealthBench-compatible rubric evaluation framework for systematic benchmarking of appeal letter quality across accuracy, grounding, and safety axes on MIMIC-IV claims data.

Education

- **University of Massachusetts Amherst** *2016 – 2024*
  Ph.D. in Computer Science (Advisor: Prof. Andrew McCallum) | *GPA: 3.83/4.00* Amherst, MA
- **Vietnam National University** *2014*
  M.S. / B.E. in Computer Science and Engineering (Thesis: 10/10) Ho Chi Minh City, Vietnam

Selected Publications & Patents

FULL LIST: GOOGLE SCHOLAR

[1] **[First author]** D.N. Thai et al. **Faithfulness Hallucination Detection in Healthcare AI.** *KDD Workshop on AI & Data Science for Healthcare (KDD-AIDSH)*, 2024.

[2] **[First author]** D.N. Thai et al. **ACR: A Benchmark for Automatic Cohort Retrieval.** *BioKDD @ KDD*, 2024.

[3] **[First author]** D. Thai et al. **Machine Reading Comprehension Using Case-Based Reasoning.** *Findings of ACL: EMNLP*, 2023.

[4] **[Co-first author]** R. Das, M. Zaheer, D. Thai et al. **Case-Based Reasoning for Natural Language Queries over Knowledge Bases.** *EMNLP*, 2021.

[5] **[Co-first author]** H. Iida, D. Thai et al. **TABBIE: Pretrained Representations of Tabular Data.** *NAACL-HLT*, 2021.

[P] D. Thai et al. **Semantic Reasoning for Tabular Question Answering.** US Patent 17/317,052 (2022). **+1 additional US patent** (17/930,288, 2024).

Recognition & Service

- **Vietnam Education Foundation (VEF) Fellowship** (2015) — U.S. Congress-funded; one of 34 fellows selected from a competitive, merit-based program (~2–3% acceptance rate) supporting outstanding Vietnamese scholars in STEM.
- **Reviewer:** NeurIPS, ICLR, SIGIR, ACL Rolling Review (ARR)
- **Organizing / Program Committee:** Spa-NLP Workshop @ ACL 2022; SUKI Workshop @ NAACL 2022