Tai Dang

(413)-425-3568 | taitdang11@gmail.com | linkedin.com/in/tai-dang11 | github.com/tai-dang11

Education

University of Massachusetts Amherst, B.S., Computer Science — May 2024
- *Selected for Chancellor's Award for academic excellence performance* — GPA: 3.8
- Coursework: TensorFlow for Artificial Intelligence, Machine Learning, Natural Language Processing, Computer Vision, Data Structures & Algorithms, Object Oriented Techniques, Software Engineering, Advanced Algorithms

Publications

"Drug Discovery with Expert Preferences"
- Chemist-guided preferential screening, recovering **16/37** EGFR and **37/58** DRD2 drugs from 100K ligand library || arXiv

"Gathering Context that Supports Decisions via Entropy Search with Language Models"
- Information-seeking with LLMs under uncertainty, closing **85%** of the performance gap to a fully-informed agent || arXiv

"Enriching Biomedical Knowledge for Low-resource Language Through Translation"
- State-of-the-Art in Vietnamese biomedical benchmark and the high-quality Vietnamese MedNLI dataset || **EACL'23**

"MTet: Multi-domain Translation for English and Vietnamese"
- State-of-the-Art in English-Vietnamese translation and high-quality multi-domain bilingual corpus || arXiv

"AURORA-M: The First Open Source Multilingual Language Model Red-teamed according to the U.S. Executive Order"
- Open-source model aligned with U.S. safety standards, achieving state-of-the-art in multilingual and code || arXiv

Experience

AI Researcher || *Stanford University (Advisors: Thang Luong, Jeff Glenn)* — Jan 2024 – Present
- Led the development of a multi-objective Bayesian optimization for virtual screening on **100K ligand libraries**, identifying **16/37** drugs for EGFR and **37/58** drugs for DRD2 with chemist feedback, achieving SOTA-level accuracy at **4x** faster speeds
- Developed a diffusion model for molecular docking, leveraging a **11M compound-protein pair** dataset with advanced data augmentation techniques, achieving **55% RMSD < 2 Å** accuracy on Posebusters with a 4M-parameter model
- Built an LLM agent for under-specified tasks (GSM8K, ARC-1D) that identifies missing info and asks clarifying questions via entropy-based search, achieving **80%** and **60%** accuracy—on par with full-context baseline

AI Research Engineer || *VietAI (Advisors: Thang Luong)* — April 2022 – May 2024
- Implemented and built a state-of-the-art English-Vietnamese Machine Translation Model that outperformed Google Translate by more than **2%** in BLEU Score by training 3 TPUv3-8 in parallel via Google Cloud Platform
- Published the largest high-quality Engish - Vietnamese Translation MTet corpus (**4.2M** sentence-pairs) that gained attention from both academia and industry with **10K+** downloads per month
- Drove **6%** enhancement in BLEU score for Vietnamese Biomedical Neural Machine Translation via Self-Training, additionally machine-translating a premium Vi-MedNLI dataset with minimal human intervention

AI Research Intern || *Ontocord* — May – July 2023
- Distilled a 7B Llama model with sparsification + quantization, achieving a 5x reduction in model size and a 2x improvement in inference speed, while preserving 87% performance compared to the teacher model's
- Crawled, cleaned and quality-classified **1TB** text-dataset and utilized transfer learning and vocab-extending techniques to develop an open-source LLM for Vietnamese using LLaMA and reduced the GPU usage by 50%

Software Engineer Intern || *EOG Resources* — Aug – Dec 2023
- Migrated 15 repositories from Jenkinsfiles to GitHub Actions, improving workflows by removing hard-coded secrets and easing testing and deployment to improve the reliability and scalability of CI/CD pipelines
- Implemented OIDC authentication procedure for 4 APIs resulted in a **15%** reduction in unauthorized access attempts
- Developed a Neo4j graph database to visualize and manage oil & gas equipment data in 5 facilities

Research Assistant || *FPT Software* — May – Aug 2021
- Built interactive frontend for AI model website using REST APIs, serving **50,000+** users
- Analyzed vision state-of-the-art papers and developed new ideas from the previous algorithm's strength and robustness

Technical Skills

Frameworks & Libraries: PyTorch, Flax, JAX, TensorFlow, Hugging Face, Flask, Node.js, Neo4j, MongoDB
Languages: Python, Java, JavaScript, C/C++, SQL
Developer Tools: Git, Linux, GCP, Slurm, Docker, GitHub, Kubernetes