Ha Vo

thuhavothi2001@gmail.com | linkedin.com/in/havo2001 | github.com/havo2001 | havo2001.github.io | (413) 658-4000

Summary

- MS Statistics student with hands-on experience in data science, statistical modeling, and natural language processing.
- Proficient in developing predictive models with Python, R, and SQL, and in NLP with PyTorch, Transformers, and LangChain.
- Industry experience from a data science internship in pricing modeling, backed by strong academic records.

Education

University of Massachusetts Amherst **Amherst, MA**
Master in Statistics *Sep 2024 – May 2026*
***GPA:** 4.0 - Full graduate assistantship (tuition + stipend)*

Moscow Institute of Physics and Technology **Dolgoprudny, Russia**
Bachelor in Applied Mathematics and Computer Science *Sep 2020 – June 2024*
***GPA:** 3.94 - Top 2% of department*

- Courses: Advanced NLP, Machine Learning, Deep learning, Regression Modeling, Hypothesis Testing, Data Visualization, Database, Statistics, Probability Theory, Linear Algebra, Multivariate Calculus, Optimization, Data Structure & Algorithm, OOP

Experience

Data Science Intern **Jun 2025 - Aug 2025**
The Travelers Companies, Inc. (S&P 500) *Hartford, CT*

- Developed a risk segmentation pure premium model with Elastic Net GLM and LightGBM for **over 4M** policies to reflect true risk across customer groups, boosting **model lift by 50%** over the production model.
- Built an automated training pipeline that **reduced the time to rerun experiments by 70%**. Implemented the pipeline on AWS EC2 with data from S3, using Optuna for hyperparameter tuning and generating SHAP summaries to interpret model behavior.
- Delivered results through clear reports and presentations to actuarial teams and non-technical stakeholders, helping actuaries refine rating plans and align pricing models with business objectives.

Graduate Teaching Assistant **Sep 2024 - Present**
University of Massachusetts Amherst *Amherst, MA*

- Graded exams and homework for about **100+** students in an introductory statistics class; led weekly calculus tutoring sessions that provided clear feedback, review materials, and practice questions to help students prepare for exams.

Graduate Teaching Assistant **Mar 2024 - Jun 2024**
Computer Vision Laboratory, Moscow Institute of Physics and Technology *Dolgoprudny, Russia*

- Implemented a Python and OpenCV pipeline with a pretrained YOLO model to detect floor line markers, fuse dual camera feeds into a top down view, and generate precise pick and place coordinates for depalletizing robot operations.
- Achieved **93% accuracy** in estimating robot speed by developing a top view camera analytics module that converted video frames into world space trajectories. The system is in production at **1K+** supermarkets across Russia.

Projects

Graph-Based RAG Summarization | *Python, PyTorch, Transformers, LangChain, OpenAI API, FAISS, NetworkX*

- Built a retrieval augmented generation (RAG) pipeline for long meeting summarization on QMSum, comparing sparse BM25, dense Contriever, and Graph of Records (GoR) retrievers on FAISS indexes.
- Evaluated summary quality with ROUGE and an LLM-based judge, analyzing retrieved chunk relevance to iteratively refine chunking rules, retrieval configurations, and prompts for more accurate, faithful summaries.

Real vs Fake Text Detection | *Python, PyTorch, Transformers, PEFT (LoRA), Hugging Face Accelerate, scikit-learn*

- Fine-tuned a Longformer with LoRA for paired text classification to detect real vs. fake text, boosting **accuracy to 91.13%** using LLM-generated synthetic data and augmentation; placed **65/994 (Top 7%)** in Kaggle's Fake or Real: The Impostor Hunt in Texts.

Sequence Modeling with Transformers for Letter Prediction | *Python, PyTorch, Transformers*

- Synthesized a **6M sample training** set from a 250K word dictionary by randomly masking letters and converting each partly hidden word into a fixed length sequence, framing Hangman as a next letter classification problem.
- Trained a transformer model that raised the Hangman solver's win rate on unseen words from 18% with frequency based baseline to **53%**.

Skill Extraction for Biostatistician Roles | *Python, R, PyTorch, Transformers, NLTK, Pandas*

- Led a team of four to extract and standardize **1K+** technical and domain skills from **27K** biostatistician job postings using BERT NER model, Sentence Transformers, and embedding driven clustering.
- Eliminated manual tagging and **uncovered 500+ new meaningful skills** beyond traditional keyword search. Delivered ranked skill reports, and the proposed solution was adopted into production at Biogen Inc.

Skills

- **Languages**: Python, R, SQL, C/C++, Java, JavaScript, HTML, CSS
- **Frameworks**: PyTorch, scikit-learn, Transformers, Spark, LangChain, NumPy, Pandas, Matplotlib, Seaborn, OpenCV, Optuna, Plotly
- **Developer Tools**: AWS (EC2, S3), Docker, GitHub Copilot, Cursor, Jupyter Notebook, Visual Studio, PyCharm, React, OpenAI
- **Data Science**: A/B testing, Experimental Design, Statistical Modeling, Feature Engineering, Model Evaluation, SQL Optimization