Hongtao Hao

Madison, WI | hongtaoh@cs.wisc.edu | hongtaoh.com | Google Scholar

PROFESSIONAL SUMMARY

Ph.D. Candidate in Computer Science with deep expertise in Machine Learning, Generative AI, and scalable software systems. Expertise in large-scale HPC pipelines (9k+ distributed jobs), time-series modeling, LLMOps, and Retrieval-Augmented Generation (RAG). Strong record of transforming research into production-ready AI tools, including three open-source ML packages for multimodal clinical datasets, and full-stack RAG application. Passionate about designing scalable, high-performance AI systems that connect research innovation with real-world impact.

RELEVANT WORK EXPERIENCE

University of Wisconsin-Madison Madison, WI

PhD Researcher (Machine Learning) *Aug. 2021 — May 2026*

- Designed and implemented novel Bayesian probabilistic models for healthcare disease progression and subtype discovery that are validated against real-world clinical datasets (ADNI, NACC), achieving 21-46% ordering accuracy improvements and 89% disease staging accuracy improvements over state-of-the-art methods across diverse synthetic datasets.
- Designed and executed 9,000+ distributed computing jobs on CHTC HPC clusters for large-scale experiments, demonstrating strong experience with scalable ML/AI systems.
- Built and maintained three production-ready open-source Python packages (pysaebm, pyjpm, bebms) for probabilistic disease progression modeling, demonstrating ability to transform complex research algorithms into well-documented, pip-installable tools adopted by the broader research community.

Robert Bosch LLC Sunnyvale, CA

Research Intern, LLM Applications *Jun 2023 — Aug 2023*

- Built full-stack LLM powered data analysis application using Python, MongoDB, and Streamlit, enabling domain-specific querying and automated insight generation from corporate datasets.
- Implemented scalable evaluation pipeline using Microsoft Azure for metrics such as response quality and relevance, contributing to 2 peer-reviewed publications on LLM-human interaction patterns and LLM capabilities of domain-specific data analysis (CHI EA '24, IUI '24 Companion).

YY Lab, Indiana University Bloomington Bloomington, IN

Research Assistant (Full-time), Data Visualization *Aug 2020 — May 2021*

- Maintained COVID-19 trend visualization dashboard (covid19-dashboard.pages.dev) in D3.js under Prof. Yong-Yeol Ahn, achieving Top 10 Most Liked notebooks on Observable platform (2020).
- Automated data workflow with CI/CD pipeline, ensuring daily synchronization with Our World in Data (OWID) COVID-19 Pandemic dataset covering 200+ countries and regions.
- Implemented interactive visualization features including temporal controls (delay parameter, play speed, time-reversal), data range filtering, and user interface.

TECHNICAL SKILLS

- **Generative AI & LLMs:** RLHF (Reinforcement Learning from Human Feedback), DPO (Direct Preference Optimization), GRPO, LLM Fine-tuning (PeFT/LoRA), RAG (Retrieval-Augmented Generation), LangChain, Prompt Engineering.
- **Machine Learning:** Bayesian Modeling, Probabilistic Programming, Time-Series Analysis, PyTorch, TensorFlow, scikit-learn, XGBoost, Hugging Face Transformers.
- **Software Engineering:** Python (Package Development), JavaScript/TypeScript, SQL, Docker, CI/CD, Git, Linux, Streamlit, FastAPI, Next.js, React.
- **Infrastructure & Data:** High-Performance Computing (HPC/Slurm), Google Cloud Platform, Microsoft Azure, MongoDB, Vector Databases.

EDUCATION

University of Wisconsin-Madison Madison, WI

PhD in Computer Sciences *Sept 2021 — May 2026*

SELECT PROJECTS

Creator • **TriFetch AI: RLHF Control Room** • github.com/hongtaoh/trifetch

- Built an interactive RLHF workbench for aligning medical AI, featuring real-time simulation of expert ranking and calculation of optimization metrics (DPO Loss, GRPO Advantages).
- Implemented robust generation pipelines using rejection sampling with fallback mechanisms to ensure effective alignment experiments even with smaller language models (e.g., Qwen-0.5B).
- Designed a modular, configuration-driven architecture to support rapid prototyping and dynamic model switching.

Creator • **CHTC GPU Lab: LLM Inference Template** • github.com/hongtaoh/chtc_llm_demo

- Built an end-to-end LLM inference pipeline on HPC infrastructure using HTCondor job scheduling, GPU compute nodes (CUDA 9.0+), and HuggingFace Transformers.
- Optimized GPU job workflows by pre-packaging conda environments and model weights, reducing cold-start latency.

Maintainer • **LLM-Powered Movie Recommendation System** • htmovies.vercel.app

- Engineered a retrieval-augmented generation (RAG) system using LangChain retrievers, vector stores (e.g., embeddings for similarity search), and LLMs for personalized movie recommendations.
- Deployed as a production web app on Vercel, enabling scalable real-time queries from users efficiently.

Maintainer • **Deep Learning for American Time Use Survey (ATUS)** • atus.hongtaoh.com

- Trained TensorFlow-based deep learning models to predict demographic time-use patterns from survey data.
- Deployed as an interactive web interface, demonstrating skills in low-latency interactive predictions.

Maintainer • **pysaebm: Open-Source Python Package** • pypi.org/project/pysaebm

- Developed and maintained an open-source Python package for event-based modeling, demonstrating strong documentation & written communication skills, and showcasing experiences with transforming research into production ready tools.

PUBLICATIONS

† These authors contributed equally to this work.

1. Hao, H., & Austerweil, J. L. (2025). **Bayesian Event-Based Model for Disease Subtype and Stage Inference.** Machine Learning for Health (ML4H) Symposium, & NeurIPS 2025 Workshop (Learning from Time Series for Health). PDF

2. Hao, H., & Austerweil, J. L. (2025). **Joint Progression Modeling (JPM): A Probabilistic Framework for Mixed-Pathology Progression.** Machine Learning for Health (ML4H) Symposium, & NeurIPS 2025 Workshop (Learning from Time Series for Health). PDF

3. Hao, H., Prabhakaran, V., Nair, V. A., Adluru, N., & Austerweil, J. L. (2025). **Stage-Aware Event-Based Modeling (SA-EBM) for Disease Progression.** Machine Learning for Healthcare (MLHC) Conference. Link

4. Guo, J., Mohanty, V., Piazentin Ono, J. H., Hao, H., Gou, L., & Ren, L. (2024, May). **Investigating interaction modes and user agency in human-llm collaboration for domain-specific data analysis.** Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (pp. 1-9). DOI

5. Guo, J., Mohanty, V., Hao, H., Gou, L., & Ren, L. (2024, March). **Can LLMS infer domain knowledge from code exemplars? A preliminary study.** Companion Proceedings of the 29th International Conference on Intelligent User Interfaces (pp. 95-100). DOI

6. Hatfield, H. R.†, Hao, H.†, Klein, M., Zhang, J., Fu, Y., Kim, J., Lee, J., & Ahn, S. J. (2024). **Addressing Whiteness in communication scholar composition and collaboration across seven decades of ICA journals (1951-2022).** Journal of Communication, 74(6), 451-465. DOI

7. Hao, H. (2023). **Selfie-editing among young Chinese women may have little to do with self-objectification.** Current Psychology, 1-18. DOI

8. Hao, H., Cui, Y., Wang, Z., & Kim, Y.S. (2022). **Thirty-two Years of IEEE VIS: Authors, Fields of Study and Citations.** IEEE Transactions on Visualization and Computer Graphics. DOI