Han Bao

Email | Website | Github | Google Scholar

Education

- **University of Notre Dame** — *Sep 2026 – June 2031(Expected)*, South Bend, USA
  Ph.D in Computer Science, supervised by Prof. Fanny Ye.
- **Sichuan University** — *Sep 2021 – June 2026*, Chengdu, China
  Bachelor of Engineering in Cyber Science and Engineering
  - GPA: 3.56 / 4.00

Research Experience

- **National University of Singapore [🌐]** — *July 2024*, Singapore
  Summer Workshop Participant
  - **Advisor:** Prof. Tianbai Ma
  - **Research Area:** Cloud Computing, Network Traffic Analysis
- **MINE Lab, University of Notre Dame [🌐]** — *Aug 2024 – Present*, South Bend, USA
  Visiting Research Student
  - **Advisor:** Prof. Xiangliang Zhang
  - **Research Area:** Trustworthy Generative Models, LLM Alignment.

Publications

C=Conference, J=Journal, S=Submission/Preprint

[J.1] **CRISP: A Causal Relationships-Guided Deep Learning Framework for Advanced ICU Mortality Prediction**
Linna Wang, Xinyu Guo, **Han Bao**, Lihua Jiang, Li Zhao, Tao Zhu, Ziliang Feng
Accepted by *BMC Medical Informatics and Decision Making*.

[C.1] **LiD-FL: Towards List-Decodable Federated Learning**
Hong Liu, Liren Shan, **Han Bao**, Ronghui You, Yuhao Yi, Jiancheng Lv
Accepted by *AAAI Conference on Artificial Intelligence (AAAI 2025)*.

[C.2] **On the Trustworthiness of Generative Foundation Models: Guideline, Assessment, and Perspective**
Yue Huang, Chujie Gao, Siyuan Wu, ..., **Han Bao**, et al.
Accepted by *ICLR 2026*.
Project page: trustgen.github.io. ***Role: Leader of T2I Group*.**

[C.3] **IntraAI: Bridging Human-AI Understanding Through Intelligent Interface Design**
Han Bao, Yue Huang, Xiangliang Zhang, Yanfang Ye
Accepted by WWW 2026.
Project page: Intra_AI

[S.1] **AutoDavis: Automatic Benchmarking of Large Vision-Language Models on Visual Question-Answering**
Han Bao*, Yue Huang*, Yanbo Wang, Jiayi Ye, Xiangqi Wang, Xiuying Chen, Mohamed Elhoseiny, Xiangliang Zhang
Preprint on arXiv. (*Equal Contribution)

[S.2] ***PolicyLLM*: Towards Excellent Comprehension of Public Policy for Large Language Models**
Han Bao, Penghao Zhang, Yue Huang, Rui Su, Zhengqing Yuan, Yujun Zhou, Xiangqi Wang, Kehan Guo, Nitesh Chawla, Xiangliang Zhang
Manuscript in preparation.

[S.3] **Capability-Oriented Training Induced Alignment Risk**
Yujun Zhou*, **Han Bao***, Yue Huang, Kehan Guo, Zhenwen Liang, Pin-Yu Chen, Tian Gao, Werner Geyer, Nuno Moniz, Nitesh V Chawla, Xiangliang Zhang

[S.4] **Guardian-as-an-Advisor: Advancing Next-Generation Guardian Models for Trustworthy LLMs**
Yue Huang, Haomin Zhuang, Jiayi Ye, **Han Bao**, Yanbo Wang, Hang Hua, Siyuan Wu, Pin-Yu Chen, Xiangliang Zhang

[S.5] **Drift-Bench: Diagnosing CoopeRative Breakdowns in LLM Agents under Input Faults via Multi-Turn Interaction**
Han Bao, Zheyuan Zhang, Pengcheng Jing, Zhengqing Yuan, Kaiwen Shi, Yanfang Ye

[S.6] **AI Alignment Breaks at the Edge**
Han Bao, Yue Huang, Xiaoda Wang, Zheyuan Zhang, Yujun Zhou, Carl Yang, Xiangliang Zhang, Yanfang Ye

[S.7] **SkillGen: Verified Inference-Time Agent Skill Synthesis**
Yuchen Ma, Yue Huang, **Han Bao**, Haomin Zhuang, Swadheen Shukla, Michel Galley, Xiangliang Zhang, Stefan Feuerriegel

[S.8] **Confidence Laundering in Agent Systems: Why Uncertainty Needs a Latent Carrier**
Kaiwen Shi, Zheyuan Zhang, **Han Bao**, Colby Nelson, Yanfang Ye

[S.9] **UXBench: Measuring the Actionability of LLM-Generated UX Critiques**
Wenjie Wang, Yue Huang, Zipeng Ling, **Han Bao**, Hang Hua, Xiaonan Luo, Yu Jiang, Shiyi Du, Yuexing Hao, Xiaomin Li, Yuchen Ma, Dianzhuo Wang, Yanfang Ye, Xiangliang Zhang

[S.10] **SpecAlign: Efficient Specification-Grounded Alignment of Large Language Models via Synthetic Data**
Wenjie Wang, Yue Huang, Zhengqing Yuan, **Han Bao**, Shiyi Du, Yuchen Ma, Yue Zhao, Yanfang Ye, Xiangliang Zhang

[S.11] **Why Semantic Entropy Fails? Geometry-Aware and Calibrated Uncertainty for Policy Optimization**
Zheyuan Zhang, Kaiwen Shi, **Han Bao**, Zehong Wang, Tianyi Ma, Yanfang Ye

Selected Projects

- **Web-based Advanced Medical Image Processing Platform** *Oct 2023 – Sep 2024*
  Tools: Segment Anything Model (SAM), Glance, JavaScript, WebAssembly
  - Integrated the Segment Anything Model (SAM) for interactive and precise medical image segmentation.
  - Optimized the rendering pipeline by replacing the traditional OHIF viewer with the Glance viewer, significantly improving loading speeds.

Honors and Awards

- **Provincial Third Prize – China Undergraduate Mathematical Contest in Modeling** *Nov 2023*
  Chinese Society for Industrial and Applied Mathematics (CSIAM)
  - Recognized for excellence in mathematical modeling and collaborative problem-solving under tight deadlines.
- **Second-Class Scholarship (Top 10%)** *Sep 2023*
  School of Cyber Science and Engineering, Sichuan University
  - Awarded for outstanding academic achievement and research potential.

Technical Skills

- **AI & Machine Learning:** PyTorch, Transformers, LLaMA-Factory, LoRA/QLoRA, Reinforcement Learning (RL, VERL framework), Agentic Workflows
- **Development:** Python, C/C++, Docker, Git, MySQL, Redis, Wireshark, HTML/CSS
- **Research:** Experimental Design, Technical Writing, Data Visualization, Literature Review
- **Languages:** English (IELTS 7.0), Mandarin Chinese (Native)

Academic Service

- **Reviewer** *2027*
  International Conference on Learning Representations (ICLR 2027)
- **Reviewer** *2027*
  The ACM Web Conference (WWW 2027)
- **Reviewer** *2026*
  AAAI Conference on Artificial Intelligence (AAAI 2026)
- **Program Committee / Reviewer** *Nov 2025*
  NeurIPS 2025 Workshop on Socially Responsible and Trustworthy Foundation Models