Tiansheng Huang

PhD candidate at Georgia Institute of Technology
Email: thuang374@gatech.edu
Homepage: https://huangtiansheng.github.io/

Education

Georgia Institute of Technology, Atlanta, USA Aug 2022 – Present

- Fourth-year PhD student, School of Computer Science
- Program Advisor: Prof. Ling Liu

South China University of Technology, Guangzhou, China Sept 2015 – June 2022

- 4-Year B.S + 3-Year Master study with School of Computer Science
- Program Advisor: Prof. Weiwei Lin
- Master Thesis: Application of Multi-arm Bandit in Client Selection of Federated Learning

Research Interest

Current interest

- My current research interest lies in trustworthy and efficient machine learning, distributed machine learning, and parallel and distributed computing.
- In the long term, I am interested in solving real-world problems concerning the **societal impact & efficiency** of machine learning models.

Previously studied

- Resource scheduling on cloud/edge computing.

Industrial Experience

Google DeepMind, Mountain View, USA May 2025 – August 2025
Research Intern

- Safety issues related to multi-modal large language model.
- Program Advisor: Virat Shejwalkar

Dolby Advanced Technology Group, Atlanta, USA May 2024 - August 2024
Research Intern

- Refine service delivery pipeline for LLMs/VLMs against harmful fine-tuning attack.
- Program Advisor: Gautam Bhattacharya, Pratik Joshi, Josh Kimball

JD explore academy, Beijing, China March 2022 - June 2022
Research Intern

- Develop Personalized FL algorithms with factorization and sparse compression.
- Program Advisor: Li Shen

JD explore academy, Beijing, China June, 2021 - Sept 2021
Research Intern

- Develop high efficiency sparse training algorithms for personalized FL.
- Program Advisor: Li Shen

Publications

Conference

[C1] **T. Huang,** G. Bhattacharya, P. Joshi, J. Kimball, L. Liu, “Antidote: Post-fine-tuning Safety Alignment for Large Language Models against Harmful Fine-tuning,” **ICML2025** [pdf]

[C2] **T. Huang**, S. Hu, F. Ilhan, S. Tekin, L. Liu, “Booster: Tackling Harmful Fine-tuning for Large Language Models via Attenuating Harmful Perturbation,” **ICLR2025 (Oral)** [pdf]

[C3] **T. Huang**, S. Hu, L. Liu, “Vaccine: Perturbation-aware Alignment for Large Language Model against Harmful Fine-tuning,” **NeurIPS2024** [pdf]

[C4] **T. Huang**, S. Hu, F. Ilhan, S. Tekin, L. Liu, “Lisa: Lazy Safety Alignment for Large Language Models against Harmful Fine-tuning Attack,” **NeurIPS2024** [pdf]

[C5] S. Tekin, F. Ilhan, T. Huang, S. Hu, L. Liu, "LLM-TOPLA: Efficient LLM Ensemble by Maximising Diversity," **EMNLP 2024 (Findings)** [pdf]

[C6] K. Chow, S. Hu, **T. Huang**, L. Liu, "Personalized Privacy Protection Mask Against Unauthorized Facial Recognition"**, ECCV2024.** [pdf]

[C7] K. Chow, S. Hu, **T. Huang**, F. Ilhan, W. Wei, L. Liu, "Diversity-driven Privacy Protection Masks Against Unauthorized Face Recognition", **PET2024.** [pdf]

[C8] F. Ilhan, G. Su, S. Tekin, **T. Huang**, S. Hu, L. Liu, "Resource-Efficient Transformer Pruning for Finetuning of Large Models", **CVPR2024.** [pdf]

[C9] S.Hu, **T. Huang**, KH. Chow, W. Wei, Y. Wu, L. Liu. "ZipZap: Efficient Training of Language Models for Ethereum Fraud Detection", **WWW2024.** [pdf]

[C10] F. Ilhan, KH. Chow, S. Hu, **T. Huang**, S. Tekin, W. Wei, Y. Wu, M. Lee, R.Kompella, H. Latapie, G. Liu, L. Liu, "Adaptive Deep Neural Network Inference Optimization with EENet," **WACV2024.** [pdf]

[C11] **T. Huang**, S. Hu, KH. Chow, F. Ilhan, S. Tekin, L. Liu, "Lockdown: Backdoor Defense for Federated Learning with Isolated Subspace Training," **NeurIPS2023.** [pdf]

[C12] Y. Sun, L. Shen, **T. Huang**, and D. Tao, "FedSpeed: Larger Local Interval, Less Communication Round, and Higher Generalization Accuracy," **ICLR2023.** [pdf]

[C13] F. Ilhan, SF Tekin, S Hu, **T. Huang**, KH Chow and L Liu, "Hierarchical Deep Neural Network Inference for Device-Edge-Cloud Systems[C]" **WWW2023.** [pdf]

[C14] S. Hu, **T. Huang**, F. İlhan, SF. Tekin, L. Liu, "Large Language Model-Powered Smart Contract Vulnerability Detection: New Perspectives" **IEEE TPS2023.** [pdf]

Journal

[J1] **T. Huang**, L. Shen, Y. Sun, W. Lin, and D. Tao, "Fusion of Global and Local Knowledge for Personalized Federated Learning," 2022, Transactions on Machine Learning Research (**TMLR**). [pdf]

[J2] **T. Huang**, W. Lin, L. Shen, K. Li and A. Y. Zomaya, "Stochastic Client Selection for Federated Learning with Volatile Clients," 2022, IEEE Internet of Things Journals (**IoT-J**). [pdf]

[J3] **T. Huang**, W. Lin, X. Hong , X. Wang, Q. Wu, R. Li, CH. Hsu, AY. Zomaya, "Adaptive Processor Frequency Adjustment for Mobile Edge Computing with Intermittent Energy Supply", 2021, IEEE Internet of Things Journals (**IoT-J**). [pdf]

[J4] **T. Huang**, W. Lin, W. Wu, L. He, K. Li and AY. Zomaya, "An Efficiency-boosting Client Selection Scheme for Federated Learning with Fairness Guarantee," 2020, IEEE Transactions on Parallel and Distributed Systems (**TPDS**). [pdf]

[J5] **T. Huang**, W. Lin, C. Xiong, R. Pan and J. Huang, "An Ant Colony Optimization Based Multi-objective Service Replicas Placement Strategy for Fog Computing," 2020, IEEE Transactions on Cybernetics (**TCYB**). [pdf]

Under Submission

[U1] **T. Huang**, S. Hu, W. Wei, L. Liu, "Silencer: pruning-aware backdoor defense for decentralized federated learning," Under Submission.

[U2]**T. Huang**, S. Hu, F. Ilhan, S. Tekin, L. Liu, "Harmful Fine-tuning Attacks and Defenses for Large Language Models: A Survey," Under Submission. [pdf]

[U3]**T. Huang**, S. Hu, F. Ilhan, S. Tekin, L. Liu, “Virus: Harmful Fine-tuning Attack for Large Language Models bypassing Guardrail Moderation,” Under Submission. [pdf]
[U4]**T. Huang**, S. Hu, F. Ilhan, S. Tekin, Z. Yahn, Y. Xu, L. Liu, “Safety Tax: Safety alignment makes your large reasoning models less reasonable,” Under Submission. [pdf]

Projects with Detailed Information

Area 1: Large language/reasoning models and their safety alignment (Current focus)

We develop a line of research associated with the safety issue of Large language models (LLMs)& Large reasoning models (LRMs), including:

- Attacks: **[Virus, U3]**
- Defenses: **[Antidote (ICML25), C1], [Booster (ICLR25 Oral), C2] , [Vaccine (NeurIPS24), C3], [Lisa (NeurIPS24), C4].**
- A survey: **[Harmful Fine-tuning Survey, U2]**
- A finding: **[Safety Tax, U4]**

All projects are open-sourced with runnable scripts providing good reproducibility.

1.1 **[Virus, U3]:** a harmful fine-tuning **attack** bypassing guardrail moderation

- Confirm a safety vulnerability (named harmful fine-tuning attack) of the fine-tuning service provided by mainstream LLM service provider (e.g., OpenAI).
- Demonstrate that attackers can manipulate the fine-tuning data to break the safety alignment and elicit harmful/untruthful behaviors of the target LLMs.
- Propose an improved data manipulation solution that simultaneously evades the data filtration detection and elicits harmful behavior of the target LLMs.

1.2 **[Vaccine (NeurIPS24), C3]:** an alignment-Stage **defense**

-Uncover the reason of failure of safety alignment after fine-tuning an LLM on harmful data. We name such phenomenon “harmful embedding drift”.
-Based on the finding of “harmful embedding drift”, we develop an alignment stage defense solution, which “vaccinate” the model to be immune of harmful finetuning.
- The proposed method, named Vaccine is empirically validated to strengthen the safety alignment performance towards the attack.

1.3 **[Lisa (NeurIPS24), C4]:** a fine-tuning-Stage **defense**

- Develop a fine-tuning stage prototype solution for preserving safety alignment after harmful finetuning.
- Observed that *excess drift* towards the switching point might be the performance bottleneck for the prototype solution.
- A refined solution named Lisa is proposed to control the excess drift phenomenon.

1.4 **[Antidote (ICML2025), C1]:** a post-Fine-tuning-Stage **defense**

- Observed that alignment stage and fine-tuning stage defenses are vulnerable when specific hyper-parameters (e.g., learning rate in the fine-tuning stage) are chosen.
- Develop a post-fine-tuning stage prototype solution named Antidote to restore the model's safety alignment after harmful fine-tuning attack.
- Experimental results show that proposed post-fine-tuning defense Antidote can be applied even though undesirable hyper-parameters are chosen.

1.5 **[Booster (ICLR25 Oral), C2]:** a strengthened alignment stage **defense**

- Observed from three statistics (model harmfulness, training/testing loss)that harmful perturbation over the aligned model could be a cause of safety degradation.
- Propose a solution (named Booster) to simulate and attenuate the impact of harmful perturbation in the alignment stage.
- The proposed solution can be combined with Vaccine and Lisa to achieve better defense performance (where the name "Booster" comes from).

1.6 **[Harmful Fine-tuning Survey, U2]:** a **survey** covering existing attacks and defenses methods.

- Systematically survey the existing attacks/defenses/mechanical study of harmful fine-tuning issue.
- Provide guidance of evaluations/benchmarking methods for researching.
- Provide future vision on the research development.
- Continuously update the Github Repo (https://github.com/git-disl/awesome_LLM-harmful-fine-tuning-papers) to include recent advancement.

1.7 **[Safety Tax, U4]:** a **Finding** demonstrates a tradeoff between safety alignment and reasoning ability

- Identify that safety alignment done after the RL/SFT-based reasoning training can **compromise** the learned reasoning ability of the LRMs, which however, is **necessary** to enforce safety of the model.
- A lightweight benchmark is provided to re-produce the observed phenomenon.

Area 2: Safety aspect of Federated Learning & Distributed learning (Previously study)

2.1 **[Lockdown, C11]** Backdoor Defense for Federated Learning with isolated subspace training

- First to identify poison coupling effect in federated learning.
- Invent isolated subspace training technique to decouple and filter the poisoned parameters.
- Source code available at https://github.com/LockdownAuthor/Lockdown.

2.2 **[Silencer, U1]** Pruning-aware Backdoor Defense for Decentralized Federated Learning

- Theoretically identify empirical Fisher information as a reliable indicator of poisoned parameters.

- Empirically study the Fisher-guided pruning technique to purify the poisoned model .
- Invent a defense to boost pruning-awareness in the training phase.

Area 3: Efficient Federated Learning/Personalized Federated Learning (Previously study)

3.1 **[RBCS-F, J4]** Efficient client selection in FL with multi-arm bandit

- Identify system heterogeneity/selection fairness/ cumulative participation as main factors for federated learning system performance.
- balance system heterogeneity/selection fairness/cumulative participation with UCB/stochastic multi-arm bandit algorithms.

3.2 **[FedSLR, J1]** Efficient PFL with low-rank+sparse

- Low-rank+sparse joint compression for personalized federated learning.
- Design a proximal algorithms.to solve the problem with theoretical guarantee.

Area 4: Resource scheduling in cloud/edge environment (Previously study)

4.1 **[NAFA, J3]** Resource scheduling for renewable-energy supply edge devices

- Study computation offloading in a scenario that the edge devices are powered by renewable-energy supply.
- Formulate the problem as an event driven semi Markov decision process
- Problem solving with a deep reinforcement learning technique.

4.2 **[MRPACO, J5]** Service replicas placement in Fog computing

- Study replicas placement problem in Fog computing
- Formulate the problem as a Mixed Integer Linear Problem
- Problem solving with an ant colony algorithm.

Honor and Awards

-	Outstanding reviewer of ICLR'24	2024
-	Top reviewer of NeurIPS'23	2023
-	IEEE TPS 2023 student travel grant	2023
-	National Scholarship (Top scholarship for graduate student in China)	2021
-	National Scholarship	2020
-	First-Class School Scholarship	2019

Academy Service

- **Conference Reviewer:** NeurIPS (2023,2024,2025), ICLR (2024,2025), ICML (2024,2025), AAAI2024, AAAI2024-AIA, CVPR2025, ACL-ARR
- **Journal Reviewer:** IEEE TMC, IEEE TCOM, IEEE TP, ACM TOIT, TMLR, IEEE TIFS