Guangyao Dou

guangyaodou.github.io
gydou@seas.upenn.edu

Research Interest

Natural Language Processing, Safety and Privacy, Large Language Models, Planning and Reasoning, etc

Education

University of Pennsylvania Aug 2023 – Present
Master of Science in Engineering in Data Science GPA: 4.00/4.00

Brandeis University Sep 2019 – May 2023
Bachelor of Science in Computer Science GPA: 3.977/4.00

Research and Industrial Experience

University of Pennsylvania Philadelphia, USA
NLP Researcher Feb 2024 – Present

- Conducted comprehensive evaluations of model editing and machine unlearning strategies in LLMs to address privacy concerns, eliminate unwanted behaviors, and expunge detrimental information.
- Proposed a novel unlearning framework that avoids copyright infringement for LLMs.

Wharton Business School Philadelphia, USA
NLP Research Assistant Nov 2023 – Apr 2024

- Examined the manifestation of helping and leadership behaviors within the gig economy using LLMs.

Amazon Seattle, USA
Software Development Engineering Intern May 2021 – Aug 2021

- Employed dependency injection techniques to enable dynamic control of logging metrics in a multi-threaded tool.
- Deployed the Dynamic Configuration system that enabled real-time control of logging metrics across global hosts.

Scholarships and Awards

- Best Thesis Award 2024
- Professional Student Individual Grant ($1,000) 2024
- Safest AI Award at the Generative AI Hackathon ($500) 2024
- Phi Beta Kappa (Top 10 %) 2023
- Dean Scholarship ($11,000 dollars per year) 2019-2023
- Dean's List (Every semester) 2019-2023

Publications

Preprints:

[7] **Dou, G**, Liu, Z, Lyu, Q, Ding, K, & Wong, E. Avoiding Copyright Infringement via Large Language Model Unlearning. Under Review.

[6] Liu, Z, **Dou, G**, Jia, M, Tan, Z, Zeng, Q, Yuan, Y, & Jiang, M. Protecting Privacy in Multimodal Large Language Models with MLLMU-Bench. Under Review.

[5] Liu, Z, **Dou, G**, Tan, Z, Tian, Y, & Jiang, M. (2024). Machine Unlearning in Generative AI: A Survey. Under Review.

Peer-reviewed Papers:

[4] Liu, Z, **Dou, G**, Tan, Z., Tian, Y., & Jiang, M. (2024). Towards safer large language models through machine unlearning. In ACL Findings (2024).

[3] Liu, Z*, **Dou, G***, Chien, E, Zhang, C, Tian, Y, & Zhu, Z. Breaking the trilemma of privacy, utility, and efficiency via controllable machine unlearning. In Proceedings of the ACM on Web Conference 2024.

[2] **Dou, G**, Zhou, Z, & Qu, X. Time majority voting, a PC-based EEG classifier for non-expert users. In International Conference on Human-Computer Interaction 2022.

[1] Zhou, Z, **Dou, G**, & Qu, X. BrainActivity1: a framework of EEG data collection and machine learning analysis for college students. In International Conference on Human-Computer Interaction 2022.

Professional Services

- Reviewer for NAACL 2025.
- Reviewer for EMNLP 2024.
- Reviewer for NeurIPS 2022 Datasets and Benchmarks Track.

TEACHING EXPERIENCE

- **Teaching Assistant** at University of Pennsylvania — Fall 2024
  Applied Machine Learning (CIS 5190)
- **Teaching Assistant** at Brandeis University — Fall 2021
  Data Structures and the Fundamentals of Computing (COSI 21A)

Skills

- **Programming Skills:** Python, Java, JavaScript, SQL, Matlab
- **Language Skills:** English (native), Mandarin (native), Cantonese (conversational)