Lang Gao

E-mail: gaolang1643@outlook.com | Website: https://heartyhaven.github.io/
Address: No. 68, Zhangzhuang Road, Huaiyin District, Jinan, Shandong, China, 250000

Education Background

Huazhong University of Science and Technology (HUST) **Wuhan, China**
School of Computer Science & Technology Sept. 2021 – Jun. 2025
Bachelor of Engineering in Computer Science and Technology
GPA: 4.28/5.0

Publications

- **Gao, L.**, Geng, J., Zhang, X., Nakov, P., & Chen, X. (2024).*Shaping the Safety Boundaries: Understanding and Defending Against Jailbreaks in Large Language Models*. ARR 2025 submission.
- Liu, Y.*, **Gao, L*.**, Yang, M*., Xie, Y., Chen, P., Zhang, X., & Chen, W. (2024). *VulDetectBench: Evaluating the Deep Capability of Vulnerability Detection with Large Language Models*. arXiv preprint arXiv:2406.07595.
- Xie, Y*., Zhou, C*., **Gao, L*.**, Wu, J*., Li, X., Zhou, H.-Y., Liu, S., Xing, L., Zou, J., Xie, C., & Zhou, Y. (2025). *MedTrinity-25M: A large-scale multimodal dataset with multigranular annotations for medicine.* Proceedings of the International Conference on Learning Representations (ICLR 2025). (In press)
- Liu, W., Deng, Z., Niu, Z., **Gao, L.**, Wang, J., Wang, H., & Li, R. (2024). *Attacking for Inspection and Instruction: The Risk of Spurious Correlations in Even Clean Datasets*. ICML 2025 submission.

Research Experiences

Shaping the Safety Boundaries: Understanding and Defending Against Jailbreaks in LLMs

First author, Instructed by Prof. Xiuying Chen of MBZUAI Oct.2024–Now
"Try to interpret common mechanisms of diverse LLM jailbreak attacks in the activation space behaviors, and accordingly propose an efficient defense method." 【Link】

- **Problem Discovery**
- Jailbreak research proposed approaches to edit harmful prompts that deceive LLMs into generating harmful responses. By visualizing the dimensionality-reduced activation distribution in each layer, we found clear clustering effects of harmful prompt activations, and jailbreak activations were far from these regions..
- **Proposal and Evaluation of Hypothesis**
- We assumed that there existed a "safety boundary" in activation space, within which LLMs were more sensitive to harmful information. Jailbreak worked by driving activations out of the safety boundary. To verify this assumption, we proposed Random Activation Shift (RAS) to mitigate such effects and discovered that the model had significantly lower DSR under greater shifting distances. This verified the existence of safety boundaries.
- **Solutions**
- We proposed Activation Boundary Defense (ABD), which optimized penalty functions to penalize these outlier activations while minimally impacting activations without jailbreak attacks. ABD successfully constrained jailbreak activations within safety boundaries. The defense performance was competitive against various jailbreaks, and it preserved the general abilities of LLMs.

MedTrinity-25M: A Large-scale Multimodal Dataset with Multi-Granular Annotations for Medicine

First Co-Author, Instructed by Prof. Yuyin Zhou of the University of California, Santa Cruz Jan. 2024 – Jun. 2024
"A comprehensive, large-scale multimodal dataset for medicine, covering over 25M images across ten modalities, with multigranular annotations for more than 65 diseases." 【Link】

- **Image Data Collection and Pre-processing**
- Collected and filtered 104 image datasets from sources like Kaggle, TCIA, and Zenodo, retaining 85 balanced packages with 25M samples.
- Developed an image transformation toolkit using nibabel, SimpleITK, and OpenCV for lossless format conversion, normalization, and segmentation.
- **Data Management**
- Deployed LLaVA-Med++ annotation model across distributed servers to improve annotation speed
- Used ZSTD and SCP protocols for efficient data transmission, ensuring uniform image modes and physiological structures distribution.
- **Design of the Multi-granular Annotation Pipeline**
- Designed a pipeline integrating image annotation, global/local features, and external medical knowledge. Applied medical grounding and segmentation models for modality-specific annotations (e.g., dermoscopy, fundus, chest X-rays), and used a visual language model to generate multi-granular annotations.
- **Model Training and Optimization**
- Replaced Vicuna-7B into Llama3-8B in the original LLaVA-Med model to develop our independent updated version of LLaVA-Tri for improving the model's learning and processing abilities in basic medical information to generate high-quality captions.

(Ongoing) Large Language Model for Parameter Selection in Bayesian Optimization
First co-author, Instructed by Prof. Xiangliang Zhang of the University of Notre Dame July.2024 – Oct. 2024
"Large language models serve as agents in Bayesian Optimization frameworks for dynamic hyperparameter selection and optimization."

- Built a convenient and flexible codebase for integrating Bayesian Optimization (BO) under different frameworks, models and datasets, further supporting the intervention of LLM-based agents.
- Perfected the reformulation of BO from an agent-environment interaction perspective. Generally, the agent serves as a parameter selector and iteratively premises its selection with feedback from the evaluation part of the BO environment.
- Perfected the training-based agent capability enhancement paradigm. Implemented a greedy-based pipeline for constructing hyperparameter tuning dataset, forming offline DPO and SFT datasets to help the agent evolve.
- Conducted comprehensive ablation study on agent ability effects concerning information granularity, history record and training strategies.

VulDetectBench: Evaluating the Deep Capability of Vulnerability Detection with Large Language Models
First Co-Author, Instructed by Prof. Wei Chen of HUST and Dr.Yu Xie of FDU Jan. 2024 – Jun. 2024
"A novel benchmark assessing the code vulnerability detection capabilities of LLMs." 【**Link**】

- Collected and curated code vulnerability data from NIST, Devign, and Big-Vul, and constructed a tree-structured vulnerability relationship graph based on CWE to ensure balanced and representative CWE-type distribution in the benchmark.
- Designed a benchmark comprising five progressively challenging tasks to evaluate LLMs' capabilities in code vulnerability analysis from multiple dimensions.
- Developed novel metrics to assess LLM performance in vulnerability detection, being among the first to enable open evaluation of LLMs in this domain, and tested 17 LLMs, revealing their limitations in detailed vulnerability analysis despite success in simpler tasks like classification.

Attacking for Inspection and Instruction(A2I): Recognizing Model-added Spurious Correlations for Faithful Explanation
Group Research Member, Instructed by Prof. Ruixuan Li of IDC-Lab, HUST Oct. 2023 – Dec. 2023
"An interpretable causal model framework employing adversarial learning, aimed at correctly learning useful information from data with spurious correlations."

- Implemented common classic self-explanatory model architectures using RNN as the carrier: Recursive Neural Predictors (RNP) within the A2I code framework.
- Integrated Noise Injection (NI), which enhanced the robustness of the RNP architecture, into the RNP framework as a comparative experiment within A2I.
- Adjusted hyper-parameters to keep the result difference within a controllable range (2%) during experiments and conducted broad tests on interpretability benchmarks.
- Applied the RNP algorithm to graph neural networks for graph classification tasks.

Awards & Honors

- **Selected** for the 17th National College Students Innovation Conference (China), 2024
- **National Software Copyright Certificate:** Research on Intelligent Text Correction under Large Language Models
- **National First Price**, RAICOM Robotics Developer Contest - CAIR Engineering Competition，2024
- **National Second Price**,15th China College Students' Service Outsourcing Innovation and Entrepreneurship Competition, 2024
- **National Third Prize**, The 5th Global Campus Artificial Intelligence Algorithm Elite Competition,2023.
- **National Third Prize**, iFlytek Developer Competition, NLP Track, 2023
- **National Second Prize**, The 5th Integrated Circuit EDA Design Elite Challenge (Deep Learning Track), 2023
- **Optics Valley Morning Star Scholarship**, Hubei Province, China, 2023
- **Scholarship for Academic Excellence**, Huazhong University of Science and Technology, 2022

Skills

Deep Learning Framework: Proficient in **Pytorch, Tensorflow, Transformers**
Large Language Models:
Model editing: representation engineering and vector steering. Familiar with main-stream LLM structures.
Basic techniques (PEFT, full-parameter training; large-scale distributed training, evaluation and prompt engineering).
Strong Data Management and Processing Skills: deduplication, cleaning, formatting, and statical analysis.
Programming Languages: Proficient in **Python, Linux, C, C++**
Language: Chinese (native), English (advanced)