YANG ZHE

TEL: +65 8264 4077 | **EMAIL:** ZHE012@e.ntu.edu.sg

D.O.B: 16/06/2003 | **NATIONALITY:** Chinese | **GENDER:** Female

EDUCATION BACKGROUND

- **09/2021-07/2025** **The Chinese University of Hong Kong |** Hong Kong SAR
  Major: B.Sc. in Computer Science
  CGPA: 3.785/ 4.0
  Major GPA: 3.802 / 4.0
- **07/2023** **Summer School of Peking University |** Beijing
  Course: Economic Principal for Management (Grade: 94)

FINAL YEAR PROJECT

- **09/2024-05/2025** **The Entropy-Memorization Law: Characterizing Data Memorization in LLMs**

  Supervisor: Prof. Michael R. Lyu

  Core Contribution: Served as a co-first author (equal contribution) on a research paper investigating the fundamental principles of data memorization, planned for submission to ICLR.

  Research Abstract: This work addresses the fundamental question: How can we characterize the memorization difficulty of data in LLMs? We propose the Entropy-Memorization Law, a novel principle establishing a strong linear correlation between the entropy of training data and its memorization difficulty. This law not only provides a new theoretical framework for assessing privacy risks but also enables an effective, novel method for Dataset Inference (DI) to detect issues like test set contamination.

  Key Responsibilities & Methodology:

  • Novel Method Development: Devised and implemented a level-set-based entropy estimation method to overcome the sample space limitations of instance-wise analysis, enabling a robust measurement of data entropy across large datasets.

  • Large-Scale Empirical Validation: Conducted extensive experiments on the fully open OLMo model family (OLMo-1B, OLMo-2-7B), sampling over 500,000 prompt-answer pairs to empirically verify and establish the Entropy-Memorization Law.

  • In-depth Case Study: Investigated the "gibberish paradox"—why seemingly random, high-entropy strings are easily memorized—by analyzing them at both the character and token levels. This revealed that BPE tokenization is the key mechanism that reduces their effective entropy for the LLM.

  • Application to Dataset Inference: Designed and implemented EMBEDI (Entropy-Memorization law-BasEd Dataset Inference), a simple, reference-free method that uses the regression parameters (slope and intercept) of the EM-Law to accurately distinguish between member and non-member data.

  Key Findings & Contributions:

  • Established the Entropy-Memorization Law: Discovered and empirically proved a strong, positive linear correlation (Pearson r > 0.90) between data entropy and memorization score, providing a predictable model for memorization behavior.

  • Explained the "Gibberish" Memorization Paradox: Revealed that the discrepancy between human perception and model behavior is due to tokenization. High char-level entropy "gibberish" is often converted into low token-level entropy sequences, making them surprisingly easy for LLMs to memorize, which has significant implications for secret leakage.

  • Developed a Novel Method for Dataset Inference: Created a new, effective baseline for Dataset Inference (DI) that successfully differentiates training data from unseen data by observing their distinct patterns under the EM-Law, aiding in the detection of data contamination and IP misuse.

- **09/2024-07/2025** **Understanding Secret Leakage Risks in Code LLMs: A Tokenization Perspective**

  Supervisor: Prof. Michael R. Lyu

  Core Contribution: Served as a core author, deeply involved in the entire research process, and co-authored a research paper planned for submission to ICLR.

  Research Abstract: This project conducts an in-depth investigation into an overlooked attack surface in Code Large Language Models (CLLMs): Tokenization. Our research reveals that the prevalent Byte-Pair Encoding (BPE) algorithm systematically reduces the post-tokenization entropy of high-entropy secrets (e.g., API keys), creating what we term "Gibberish Bias." This phenomenon significantly increases the risk of secret memorization and leakage.

  Key Responsibilities & Methodology:

  • Hypothesis Formulation & Validation: Systematically demonstrated the inherent flaws of BPE tokenizers in processing random strings (i.e., secrets) through visualizations and information-theoretic analysis.

  • Quantitative Analysis of Bias Origins (RQ1): Experimentally verified a significant distributional shift between tokenized secrets and general source code using tokenizers from models like Deepseek Coder and Qwen2.5-Coder. We quantified this shift with a KL Divergence of 2.668, identifying it as the root cause of "Gibberish Bias."

  • Prospective Trend Analysis (RQ2): Investigated the impact of the industry trend towards "larger vocabularies." By training and evaluating three new tokenizers with "optimal" vocabulary sizes, we discovered that larger vocabularies tend to exacerbate the "Gibberish Bias," further widening the entropy gap between secrets and non-secret data.

  • Mitigation Strategy Proposal: Based on our findings, we proposed forcing character-level tokenization for secrets as a promising mitigation strategy to eliminate the entropy disparity and reduce leakage risks.

  Key Findings & Contributions:

  • Finding 1: First to identify and define the "Gibberish Bias" phenomenon, establishing that BPE tokenization constitutes a systemic security vulnerability in Code LLMs.

- Finding 2: Revealed that the trend of increasing vocabulary size paradoxically worsens secret leakage risks, providing critical insights for designing more secure future models. This work opens a new perspective for understanding and mitigating privacy risks in LLMs.

RESEARCH EXPERIENCE

● **06/2024-09/2024** **UG Summer Research Internship 2024**

Supervisor: Prof. Michael R. Lyu

Research Topic: Evaluate the Memorization Difficulty of Data in Large Language Models

My Duties:

- Spearheaded the comprehensive preprocessing of the open-source Dolma dataset, selecting and structuring data to underpin the inference analysis of the OLMo model;
- Conducted advanced inference analysis with the OLMo model, methodically evaluating its capacity for memorization across various contexts;
- Established and quantified essential metrics for gauging memory challenges, such as Perplexity, Entropy, and Memorization Rate, leveraging deep data analytics to uncover significant patterns and trends;
- Designed and executed experiments to investigate the effects of prompt length and data type on memorization efficacy, culminating in the authorship of a detailed research report.

● **02/2024-03/2024** **Introduction to Neuroengineering and Brain-Computer Interfaces Online Research Seminar**

Supervisor: Prof. Dejan Marković

Research Topic: Prediction of EEG Signal Disease Onset Based on Independent Ear-EEG Device

My Duties:

- Used the Standard Scaler for data preprocessing to ensure dataset standardization;
- Responsible for designing prediction models and programming for logistic regression model, decision tree classifier, and random forest algorithm;
- Wrote confusion matrix code to evaluate the accuracy and bias of the predictive models;
- As the group leader, organized team tasks, reported project meetings, and facilitated technical exchanges among team members;
- Responsible for writing the research report.

● **06/2023-09/2023** **UG Summer Research Internship 2023**

Supervisor: Prof. Tsung-Yi Ho

Research Topic: Exploring Adversarial Attacks on Gender Recognition Deepface API: A Comprehensive Study

My Duties:

- Conducted attack experiments on the DeepFace API, analyzed results, and evaluated the API's robustness.

Research Findings:

- Discovered that while the DeepFace API effectively identified gender information with clean data, its robustness significantly decreased when facing various adversarial attack strategies;
- Noted that Simple Black-box Adversarial Attack had a relatively minor impact on DeepFace, whereas Projected Gradient Descent and spatial transformation attacks had a pronounced effect on model robustness;
- Revealed the DeepFace API's reliance on hair-related features in gender classification and confidence variations across different racial groups, emphasizing the importance of diversity and fairness in training datasets for building robust facial recognition models.

PUBLICATION

● **12/2023-02/2024** **RAUIE: A Relation-Augmented Document-level Event Extraction Model Based on UIE, Co-author**

Publication: Presented at the 5th International Conference on Artificial Intelligence, Networks, and Information Technology (AINIT 2024) in Nanjing, China on February 17, 2024

My Duties:

- Learned information extraction framework UIE and researched on information extraction algorithm based on UIE frame;
- Studied the process of designing a document extraction model and its modelling;
- Responsible for proofreading and overall translation of the paper.

PROJECT

● **C Language**

2048 Game: Utilized Pygame library for game development.

RV32I Assembler & RISC-V LC Simulator: Converts and tests RV32I code.

Multi-Level Feedback Queue Scheduler: Simulated an operating system's scheduling mechanism.

Paging Technique Simulator: Demonstrated the conversion of a process's physical address to a physical address.

● **Java Language**

Tic-Tac-Toe Game: Created with AWT for the graphical user interface.

Computer Sales System: A standalone application for managing computer sales, incorporating interactive queries and transaction recording.

Gobang Battle Game: Developed using JavaFX, Socket, JDBC, and MySQL, featuring user registration, opponent selection, move notifications, time tracking, game records, undo, chat, concede, new game, save/load game, and a built-in AI assistant based on a scoring system, playable both locally and online.

- **Python Language**

  HDR Tone Mapping and Image Stitching: Enhances and merges images for visual realism.

  Photomosaic Generator: Constructs detailed large images from arrays of small, distinct photos.

  Lempel-Ziv-Welch (LZW) Compression Algorithm: Implemented data compression technique.

  Peer-to-Peer Voice Chat System: Facilitated real-time voice and video communication between users.

INTERNSHIP EXPERIENCE

- **07/2024-08/2024** **Algorithm Intern | Product R&D Department | Geovis Technology Co., Ltd**
  - Deployed network proxies and encryption technologies, including Shadowsocks and ProxyChains, and solved the problem for access restriction and security of data transmission;
  - Developed the framework for an automated data acquisition to efficiently and accurately extract and structure raw information from multiple data sources;
  - Optimized a transformer model, OneKE, and improved the accuracy and depth of knowledge graph;
  - Enhanced the analytical ability of the model trough Fine-tuning technology and gave technical support for complex analysis and decision making;
  - Utilize Neo4j graph database technology to visualize the knowledge graph, optimizing query performance and data integration processes.

HONOR & CERTIFICATE

- **06/2025** **HKSAR Government Scholarship**, Top 1% in the major of Computer Science, HKSAR Government
- **07/2024** **Dean's List 2023-2024,** Top 10% in the Faculty of Engineering, CUHK
- **05/2024** **CSE Scholarship 2023-24 -Silver Award**, Top 1% in the Faculty of Engineering, CUHK
- **09/2023** **Certification for Student Helper for the College Assembly**, Shaw College, CUHK
- **07/2023** **Dean's List 2022-2023**, Top 10% in the Faculty of Engineering, CUHK
- **02/2023** **Certification for Student Helper for the College Assembly**, Shaw College, CUHK
- **07/2022** **Dean's List 2021-2022**, Top 10% in the Faculty of Engineering, CUHK
- **11/2021** **The Chinese University of Hong Kong Student Envoy**, CUHK

EXTRACURRICULAR ACTIVITIE

- **Student Organization:**

  Shaw College Mainland Student Union: Propagandist
  - Boosted engagement by designing promotional materials with over 500 downloads and 1000+ social shares, and by organizing activities that increased participation by 30% to over 2000 students.

  Independence Times Electronic Magazine: Art Editor
  - Enhanced magazine visibility and engagement by strategically planning content, leading social media campaigns, designing 30+ promotional posters, and managing the production of 10+ articles for the class's new media platform.

  Three Heart Club: Promotion Officer
  - Boosted public welfare project visibility, reaching a record 2,200 views on a key post and gaining 300+ new followers.
  - Spearheaded multiple public welfare activities, enhancing organizational reach and engagement.
- **Hobby & Interest:** Latin dance (Intermediate), Piano (Intermediate), Skiing(amateur), Photography (amateur).

SKILLS

- **Computer Skills:**

  Programming Languages:
  - Python: Experienced (Primary language used for all final year project research, implementation, and data analysis).
  - C/C++: Experienced (Developed a small-scale game in C; completed academic projects using C++).
  - Familiar with: Java, R, MATLAB, SAS, RISC-V Assembly.
  - Frameworks & Technologies: PyTorch, MySQL, Oracle, Linux, Git, GitHub

  Office Software:
  - LaTex, Microsoft Office, Excel,WPS Office, Adobe (PS,LR,PR).
- **Language:** Mandarin (Native), English (GRE:322, Fluent), Cantonese (Intermediate).