Xueren Ge

GitHub | Website | Google Scholar | LinkedIn | zar8jw@virginia.edu

Education

University of Virginia Jun. 2022 – Present
Ph.D. in Computer Engineering ***GPA: 3.88/4.0*** *Charlottesville, Virginia*

Georgia Institute of Technology Aug. 2020 – May. 2022
M.S. in Electrical and Computer Engineering ***GPA: 3.91/4.0*** *Atlanta, Georgia*

Chongqing University Sep. 2016 – Jun. 2020
B.E. in Electrical Engineering and Automation ***GPA: 3.58/4.0*** *Chongqing, China*

Coursework

Courses: Statistical Machine Learning, Natural Language Processing, Deep Learning, Computer Vision, Reinforcement Learning, Convex Optimization, Random Processes, Computer Communication Networks, Intro to Database System, Object-Oriented Programming, Data Structures & Algorithms, Embedded Systems, Learning for Interactve Robots, Discrete Math, Linear Algebra, Calculus, Probability & Statistics

Scholarships & Awards:

- Georgia Tech Shenzhen Campus *Merit-Based Scholarship (Level A)*, 2020 (top 5%).
- Yangtze Power Scholarship, 2018 (2/324, Top 0.01%).
- Excellent Second Undergraduate Comprehensive Scholarship, 2016/2018/2019 (top 6%).
- Chongqing University Excellent Undergraduate, 2019 (Top 10%)
- Chongqing University Excellent Student, 2018 (Top 5%)
- Chongqing University Science and Technology Innovation Advanced Individual, 2018 (Top 5%)
- Mathematical Contest in Modeling (MCM) — *Meritorious Winner*, 2018 (Top 10%)

Research Interests

Natural Language Processing
Knowledge-Augmented LLMs
Multimodal Learning
Large Language Models
Uncertainty Estimation
Computer Vision
Conversational Recommendation Systems
Synthetic Data Generation
Activity Recognition

Skills

Languages: C/C++, Python, MySQL, Java, JavaScript/TypeScript, HTML/CSS, MATLAB, LaTeX
Tools: Slurm, Git, Linux, Docker, Microsoft Azure, AWS, HPC Systems, GitHub, Unix Shell, VSCode
Machine Learning Frameworks: PyTorch, Tensorflow, Wandb, Huggingface, DeepSpeed, FSDP, DDP, vLLM
Embedded System: STM32, Keil, Multisim, Proteus, AutoCAD

Publications

- **Ge, X.**, Murtaza, S., Cortez, A., & Alemzadeh, H. (2025). *Expert-Guided Prompting and Retrieval-Augmented Generation for Emergency Medical Service Question Answering*. Submitted to **AAAI 2026**.
- Weerasinghe, K., **Ge, X.**, Heick, T., Wijayasingha, L. N., Cortez, A., Satpathy, A., Stankovic, J. A., & Alemzadeh, H. (2025). *EgoEMS: A High-Fidelity Multimodal Egocentric Dataset for Cognitive Assistance in Emergency Medical Services*. Submitted to **AAAI 2026**.
- **Ge, X.**, Satpathy, A., Williams, R. D., Stankovic, J. A., & Alemzadeh, H. (2024). *DKEC: Domain Knowledge Enhanced Multi-Label Classification for Diagnosis Prediction*. In **EMNLP 2024**. (Acceptance rate: 18.4%.)
- Weerasinghe, K., Janapati, S., **Ge, X.**, Kim, S., Iyer, S., Stankovic, J. A., & Alemzadeh, H. (2024). *Real-Time Multimodal Cognitive Assistant for Emergency Medical Services*. In **IoTDI 2024**. (Acceptance rate: 36.7%.)
- Fang, X., Wang, B., Kong, H., **Ge, X.**, Yang, Z., Yu, J., & Li, W. (2023). *Human posture feature recognition method for neuropsychological comprehension test*. In **Journal of Chongqing University**.

- Ma, Y., & **Ge, X.** (2019). *An Effective Method for Defect Detection of Copper Coated Iron Wire Based on Machine Vision.* In **IOP Conference Series: Materials Science and Engineering**.

Patents

- Yu, H., Wang, B., Kong, H., He, W., **Ge, X.**, Yang, W., Li, W., Yang, Z., Lü, Y. (2025). *An action recognition method based on TensorFlow target detection.* China Patent **CN111860103B** (granted Jul. 15, 2025; assignee: Chongqing Zhiyixing Technology Development Co., Ltd.).

- Yu, H., Kong, H., **Ge, X.**, Wang, B., Wang, Z., Li, W., Yang, Z., Yu, W. (2024). *Automatic gesture recognition system for AD (analog-to-digital) meter understanding capability test.* China Patent **CN111652076B** (granted; assignee: Chongqing Zhiyixing Technology Development Co., Ltd.).

- Li, R., **Ge, X.**, Li, Q., Zhao, M., Bao, M., Ma, J. (2021–2024). *Respiratory rate detection method and device, electronic equipment, and storage medium.* China Patent **CN113887474B** (granted; assignee: Shenzhen Sensetime Technology Co., Ltd.).

- Li, R., **Ge, X.**, Li, Q., Chen, C., Ma, J. (2023). *Heart rate measurement method and apparatus, and electronic device and storage medium.* WIPO Patent **WO2023061042A1** (published Apr. 20, 2023; assignee: Shanghai Sensetime Intelligent Technology Co., Ltd.).

- **Ge, X.** (2019). *Highway electronic information direction board.* China Utility Model Patent **CN209118652U** (granted Jul. 16, 2019; assignee: Individual).

Experience

University of Virginia Jun. 2022 – Present
Graduate Research Assistant *Charlottesville, VA*

- **Expert-Guided Prompting and Retrieval-Augmented Generation for EMS Question Answering**
  PyTorch, RAG, FAISS, vLLM, LoRA, DeepSpeed, Wandb, Selenium, BeautifulSoup, Linux, CUDA
  - Created **EMSQA**, the first EMS MCQA dataset of 24.3K questions, curated based on public and private sources, covering 10 subject areas and 4 certification levels, and accompanied by a structured, subject area aligned EMS knowledge base (KB) with 40K documents and 4M real-world patient care reports.
  - Developed two techniques to inject domain expertise into Large Language Models: 1) **an expertise-guided prompting (Expert-CoT)** that encourages step-by-step reasoning from a domain-specific perspective. 2) **an expertise-guided RAG (ExpertRAG)** that retrieves expertise-aligned knowledge from curated EMS KBs and patient records.
  - Benchmarked multiple LLMs on EMSQA, evaluating performance across certification levels and subject areas, and compare our framework against SOTA RAG methods. Experimental results show that combining Expert-CoT and ExpertRAG yields up to a **4.67% improvement in accuracy**. Notably, the **32B expertise-augmented models pass all the EMS certification** simulation exams.
- **DKEC: Domain Knowledge Enhanced Multi-Label Classification for Diagnosis Prediction**
  PyTorch, UMLS, BeautifulSoup, Prompting, BERT, GNN, GPT, FSDP, Wandb, Linux, NER
  - Automated **heterogeneous knowledge graph** construction from 3,000+ medical webpages (Wikipedia, MayoClinic, MedlinePlus) using BeautifulSoup, Web APIs, and LLM prompting (chain-of-thought), extracting 5,000+ normalized medical entities via UMLS API.
  - Designed a **label-wise attention** mechanism that incorporates heterogeneous knowledge graphs to train language models, to improve multi-label classification by **5% in micro f1** compared with SOTAs on MIMIC-III datasets.
- **Synthetic Multi-person EMS Conversation Generation from Medic Notes via LLM Agents**
  PyTorch, AI Agents, vLLM, LoRA
  - Designed a **topic-flow** planner that sequences EMS dialogues into guideline-aligned stages, auto-generating per-turn blueprints (speaker, intent, topic, evidence) from real-world EHRs.
  - Integrated an AI **self-critic** to iteratively enforce medical rules, improving evidence alignment by **25%** against a ground-truth clinical checklist.
  - Built an end-to-end data generation pipeline by multiple agents (**self-critic planning → self-critic role-play → refinement**) that produced **4,000+** high-fidelity synthetic EMS conversations.
- **EgoEMS: A High-Fidelity Multimodal Egocentric Dataset for Cognitive Assistance in EMS**
  C++, Python, Android, Vision-Language Model, ASR models, Action Recognition

- ∗ Designed and executed benchmarking pipeline for **zero-shot audio models** (Whisper, Whisper-Timestamped, Google Speech, Gemini-2.5-Pro), evaluating ASR accuracy, latency, and word-level timestamp precision in self-collected real-world EgoEMS dataset.
- ∗ Conducted comprehensive comparative analysis of **zero-shot vision-language models** (Qwen-2.5-VLM, VideoLLaMA-3, Gemini-2.5-Pro) for **video understanding**, measuring **action classification** accuracy, **temporal segmentation** quality, and inference efficiency to guide model selection for downstream applications.

- **Real-Time Multimodal Cognitive Assistant for Emergency Medical Services**
  Python, PyTorch, (EMSTiny)BERT, GNN, CUDA, NVIDIA Jetson
  - ∗ Cleaned EMS electronic patient care reports (ePCRs) with regex-based normalization, schema mapping, and quality filters to create train/val/test splits.
  - ∗ Developed **EMSTinyBERT** (15M params), fine-tuned on 4K ePCRs plus curated guideline text; proposed *group-wise* training to address rare classes and boost minority-label recall.
  - ∗ Deployed the model on an NVIDIA Jetson (CUDA) as an on-device inference pipeline; achieved **80.0%** accuracy with **0.31 s** end-to-end latency.

SenseTime Incorporated Dec. 2020 – Jun. 2021
Algorithm Developer Intern *Shenzhen, China*

- **Contactless Vital Signs Estimation from Thermal Imaging using Computer Vision**
  Python, Signal Processing, Image processing, Computer Vision
  - ∗ Designed and implemented a contactless physiological monitoring algorithm for the **SenseThunder Air product**, integrating **landmark detection** and **homography transformations** to extract Regions of Interest from thermal video streams for real-time heart rate and body temperature estimation.
  - ∗ Applied advanced signal processing methods—including **FFT**, smoothing, and **bandpass filtering**—to recover weak respiratory and cardiac signals from continuous thermal frames.
  - ∗ Developed a complete **end-to-end Python framework** for data preprocessing, algorithm implementation, performance evaluation, and error analysis, enabling robust real-world deployment.

Chongqing University Jul. 2019 – Nov. 2019
Undergraduate Research Assistant *Chongqing, China*

- **Intelligent Diagnosis of Alzheimer's disease Based on Computer Vision**
  Python, CNN, Activity Recognition, Object Detection
  - ∗ Designed and implemented an AI-based system to automatically assess the severity of Alzheimer's disease using video and image data, enabling objective and scalable cognitive evaluation.
  - ∗ Developed a human eye state recognition GUI using **landmark detection Dlib** for automated cognitive assessment.
  - ∗ Designed a video analysis pipeline integrating **OpenPose-based posture estimation**, **image morphology processing**, and **Fast R-CNN** detection to detect patient's action recognition for Alzheimer's severity evaluation.
  - ∗ Trained a **CNN with a ResNet-50 backbone** in PyTorch to classify patient hand-drawn geometric figures, improving diagnostic accuracy and enabling early detection of cognitive impairment.

Service

- **Teaching Assistant:** ECE Statistical Machine Learning (Fall 2022); Dependable Computing System (Fall 2025).
- **Reviewer:** ICRA 2024; NAACL 2025; ACL 2025; IJCAI 2025; AAAI 2026; NeurIPS 2025 Efficient Reasoning Workshop; ACM Transactions on Computing for Healthcare.
- **Volunteer:** EMNLP 2024; LinkLab Open House 2024
- **Mentoring:** Saahith Janapati; Shruti Bala; Sion Kim; Abhishek Satpathy; Sahil Murtaza

Course Projects

LLMs for Diagnosis Prediction | *Python, PyTorch, LoRA, Slurm, FSDP* Aug. 2024 – Dec. 2024

- Explored prompting and finetuning strategies for diagnosis prediction using Electronic Health Records (EHRs).
- Designed a *chain-of-diagnosis* prompt template and finetuned Llama-3.1-8B with LoRA, improving diagnostic accuracy.
- Benchmarked the finetuned model against state-of-the-art medical LLMs using chain-of-thought prompting.

N-Version Programming on LLMs | *GPT-3.5, PaLM-1, Llama-2, Ensembling, vLLM* Aug. 2023 – Dec. 2023

- Improved multiple-choice question answering accuracy by merging responses from multiple LLMs.
- Designed majority and weighted-majority voters over GPT-3.5/PaLM-1/Llama-2 outputs, yielding **+5%** accuracy.

BERT Visualization and Interpretation | *PyTorch, BERT, SNLI* Aug. 2022 – Dec. 2022

- Analyzed whether BERT produces reasonable layer-wise embeddings.
- Visualized anisotropy by sampling words and computing cosine similarities across 12 layers.
- Studied bias by removing "female"-related sentences from SNLI and measuring effects.
- Investigated redundancy by freezing individual layers on classification tasks.

Phishing Website Detection | *Python, SVM, RF, GBDT, Java, PhishTank API* Jan. 2021 – May 2021

- Designed phishing detection algorithms and a phishing query web application.
- Engineered URL features (domain, IP, DNS, redirects, protocol, etc.) for model inputs. And trained SVM/Random Forest/GBDT and built an ensemble to boost performance; Build an en-to-end pipeline by integrating trained ensembled model with PhishTank API.

Wireless Bus-Stop Crowd Counter | *STM32, C, PCB, MATLAB* Jun. 2017 – Jul. 2018

- Developed a wireless terminal to support campus sightseeing-bus dispatch at Chongqing University.
- Implemented STM32 firmware in C with wireless comms, LCD UI, matrix keypad, and power management; designed/soldered the PCB with ZigBee and peripherals.
- Formulated dispatching as a TSP and solved via an ant-colony algorithm in MATLAB.