Yi Zhang

Email | Website | Google Scholar | GitHub | LinkedIn

Research interests

I investigate how foundation models reason across languages, cultures, and human cognition, focusing on representation analysis, multimodal evaluation, and the reliability of AI systems in human-centered settings.

Education

Friedrich-Alexander-Universität Erlangen–Nürnberg (FAU), Germany Oct. 2024 – Present
M.Sc. in Data Science Main area: Machine Learning and Artificial Intelligence
Relevant coursework: Artificial Intelligence, Deep Learning, Pattern Recognition

Beijing Normal University, Zhuhai (BNUZ), China Sep. 2020 – Jun. 2024
B.Eng. in Computer Science and Technology Final grade: 87/100, Top 5%
Focus: Computer Vision, emotion recognition, deep learning

Publications and Manuscripts

ChinaHeritaQA: A Culturally-Grounded Visual Question Answering Dataset for World Heritage Sites in China [huggingface] [preprint]
ACL ARR May Under review.

Do Large Language Models Think Like the Brain? Sentence-Level Evidence from fMRI and Hierarchical Embeddings [code] [paper]
AAAI 2026.

M-ABSA: A Multilingual Dataset for Aspect-Based Sentiment Analysis [huggingface] [paper]
EMNLP 2025 Main.

MTFNet: Multi-Scale Transformer Framework for Robust Emotion Monitoring in Group Learning Settings [code] [paper]
APSIPA ASC 2024.

Research Experience

Research Assistant, SODA Lab, LMU Munich Feb. 2025 – Present
NLP, LLM fine-tuning, social and cognitive language analysis

- Contributed to multiple human-centered AI research projects, including the multilingual M-ABSA benchmark construction and the LLM4Brain study, while leading the development of the ChinaHeritaQA benchmark.
- Managed lab-level HPC access permissions and approval workflows, and contributed to user guidelines that improved the efficient use of server storage and computational resources.
- Served as teaching assistant for Seminar: Natural Language Processing Meets Computational Social Science, providing technical support and building survey-agent templates for course projects.

ChinaHeritaQA Benchmark May. 2025- Present
Cross-Cultural VLM evaluation, factuality, hallucination analysis

- Developed a bilingual VQA benchmark with 2,279 in-the-wild images and 14,133 Chinese/English QA pairs to evaluate VLM limitations in cultural, historical, functional, and architectural reasoning on Chinese UNESCO World Heritage sites.
- Evaluated six open-weight VLMs against a human baseline, revealing that strong visual recognition does not reliably transfer to historical, functional, and architectural reasoning.
- Analyzed VLM failure modes including same-type site confusion, weak image-to-period grounding, image-to-function grounding, and dynasty- or region-level cultural grounding bias.

LLM–Brain Alignment in Narrative Comprehension Feb. 2025 – Dec. 2025
LLM embedding extraction, CKA similarity analysis

- Extracted layer-wise sentence embeddings from multiple LLMs on bilingual *The Little Prince* stimuli to support analysis of hierarchical representations in narrative comprehension.
- Implemented CKA-based similarity computation to compare inter-layer and inter-model representation structures from LLM hidden states.

Project Experience

Video-based Emotion Monitoring in Collaborative Learning Oct. 2022 – Sep. 2024
Video processing, affective computing, long-tail data

- Collected, synchronized, and annotated classroom video/audio data for studying student participation and affective states in collaborative learning environments.
- Built a multimodal perception pipeline with face detection, identity recognition, and temporal emotion modeling for real-world classroom interaction analysis.
- Proposed MTFNet, a multi-scale transformer framework for robust emotion monitoring, achieving 67.5% accuracy on DFEW and leading to a first-author APSIPA ASC 2024 paper.

Audio-Based Depression Detection from Real Counseling Conversations Nov. 2025 – Apr. 2026
Speech processing, Wav2Vec, clinical audio analysis

- Developed a preprocessing pipeline for noisy two-person counseling audio using Wiener denoising, faster-Whisper transcription, speaker diarization, and rule-based patient speech extraction.
- Compared MFCC-BiLSTM and Wav2Vec models for classifying depression-related labels, including depression, agitation, retardation, and HRSD, with subject-level split, class-balanced loss, and nested cross-validation.
- Used SpeechBrain-based emotion profiling to analyze sad, neutral, happy, and angry speech distributions across clinical symptom groups.
- Identified task boundaries of audio-only modeling through failure analysis on emotion-change regression and EI-stage classification, where low PCC and app-log/audio misalignment limited reliability.

Industry Experience

Algorithm Engineer Intern, ATA Assessment Technology, Beijing Jun. 2023 – Aug. 2023
Computer vision, human detection, pose estimation, deployment-oriented model evaluation

- Processed large-scale real-world dance-assessment data, including 100K human detection samples and 1K human keypoint samples, to support model evaluation and deployment.
- Benchmarked YOLOv8, PP-Detection, MediaPipe, and TinyPose for human detection and pose estimation under practical assessment scenarios.
- Improved human detection accuracy by 3% and keypoint estimation accuracy by 10%, while reducing redundant detections in the evaluation pipeline.

Skills

Machine Learning: Python, C++, PyTorch, Hugging Face Transformers/Datasets, vllm;

Tools: Git, Docker, HPC, LaTeX

Languages: Chinese Mandarin (native); English (C1);