

ZIQIAN BI

AI engineer and researcher with experience across LLM systems, embedded hardware, sensing platforms, and end-to-end AI deployment. My background combines model development, system optimization, retrieval pipelines, embedded integration, hardware-aware experimentation, and strong bilingual communication for AI products, intelligent systems, manufacturing, and research.

bizi@iu.edu | [Website](#) | [GitHub](#) | [Google Scholar](#) | [LinkedIn](#)

RESEARCH INTERESTS

Large Language Models · Multimodal Large Language Models · Natural Language Processing · Generative Models · Diffusion Models · Computer Vision · Medical AI · Model Security & Robustness · Model Interpretability · mmWave Sensing · Sensors · Embedded Systems

TECHNICAL SKILLS

Programming Languages: Python/C/C++/CUDA, MATLAB, SQL

Frameworks: PyTorch/TensorFlow, DeepSpeed, Transformers/Diffusers, LangChain/LangGraph/Pydantic-AI, FastAPI/Django

Specialized Models: LLM/MLLM, DiT/Stable Diffusion, CLIP/ViT/ResNet, YOLO, RetinaFace, Point Transformer, SimCLR, 3D VQGAN

Tools & Platforms: Docker, Vertex AI, Ollama/vLLM, FAISS, RAG systems, PostgreSQL, Git

Manufacturing Systems: MES/ERP/SCADA, OEE analysis, Siemens PLC, production-line bottleneck analysis, rapid learning

Hardware: mmWave radar (TI IWR6843/1843, Cascade EVM), STM32, ESP8266/ESP32, I2C/SPI/UART/RS-485/Modbus, GPU/CUDA

Libraries: NumPy/Pandas/Scikit-learn, XGBoost, AutoGluon, OpenCV, VTK/Mayavi

INDUSTRY EXPERIENCE

China Post (Internship)

Jan 2026 – Mar 2026

AI Algorithm Engineer, Large Language Model Direction

- Built a China Post customer-service assistant for postal-domain dialogue and ticket-generation workflows, covering RAG retrieval, dialogue generation, ticket generation, and deployable Web/API behavior; implemented with Django, django-ninja, SSE streaming, Bootstrap 5, Tailwind CSS, PostgreSQL/pgvector, FAISS, and pluggable Ollama, vLLM, SGLang, OpenRouter, and OpenAI APIs
- Filtered 5,376 postal multi-turn dialogues with semantic embeddings and a gpt-oss-20b multi-class Agent pipeline, built 8,217 instruction-tuning samples from CSDS and postal data, compared Qwen, Llama, DeepSeek, and gpt-oss models, and selected Qwen2.5-7B-Instruct for 4-bit LoRA SFT, reaching 92.3% dialogue accuracy, 93.5% ticket-field extraction, and 2.8% hallucination rate

Xperfi (Internship)

Sep 2025 – Jan 2026

AI Research & AI/ML Optimization Engineer

- Developed a PSO-based closed-loop optimization framework for Marvell high-speed interconnect validation, supporting AEC parameter tuning across 400G/800G and 1.6T environments; improved BERT/bit-error-rate measurements from 10^{-6} to 10^{-12} and reduced manual engineering effort by 97%
- Built a multiprocessing PSO evaluation loop that connected BERT testing, parallel parameter tuning, hardware-feedback collection, and I2C EEPROM programming across OSFP-DD/QSFP AEC and optical-module workflows; applied Mask R-CNN to optical inspection of failed solder joints in DAC-related validation workflows, linking visual defects with hardware-test evidence for failure analysis

Infraeo (Internship)

May 2025 – Jan 2026

Machine Learning & Deep Learning Engineer

- Deployed, fine-tuned, and optimized LLMs up to 1000B parameters, including Qwen3 235B, Qwen3 Coder 480B, Kimi v2 1000B, Llama 3.3 70B, and GPT-OSS 120B; adapted domain QA with CoT prompting and RAG to reach 89% answer accuracy while delivering $3.2\times$ faster inference and 60% lower memory usage with LoRA/QLoRA, PyTorch, Unsloth, and Hugging Face Transformers
- Built semantic retrieval over 2M+ documents with FAISS-based embeddings that generated 768-dimensional vectors and a BERT reranker, achieving recall@10 of 0.94 and precision@5 of 0.87
- Built cloud AI services and ML pipelines with FastAPI, Django, Docker, Kubernetes, FAISS, Elasticsearch, PostgreSQL, and LangChain/LangGraph, supporting 500+ QPS and 100GB+ daily data

Hanergy Mobile Energy Holding Group (Internship)

2019

R&D Hardware Engineer

- Developed an environmental sensing and equipment-monitoring system integrating STM32, LCD2004, SPI OLED, air-pressure, temperature/humidity, GPS, WiFi, display, and anemometer modules, supporting end-to-end embedded sensing, communication, and system bring-up
- Adapted C++/OpenCV image-processing routines and early rule-based inspection utilities for manufacturing-line quality analysis, including solder-joint stability observation, camera-calibration support, visual evidence collection, and line-quality analysis that later informed Mask R-CNN-based optical inspection work for failed solder joints
- Worked with MES, ERP, SCADA, OEE, and Siemens PLC concepts in factory automation contexts while rapidly learning production-system workflows, equipment-side data flows, production-line efficiency analysis, bottleneck identification, station-level throughput review, and blocking-point diagnosis to support engineering analysis and process-improvement decisions

Chinese Academy of Sciences, Institute of Electrical Engineering (Internship)

Sep 2017 – Nov 2017

Smart Systems Solutions and Development Engineer

- Built a SCADA-adjacent industrial RS485/Modbus sensor-network monitoring system for concrete infrastructure monitoring and validation, preserving gateway architecture, data flow, and monitoring behavior while adapting C# integration code for Modbus frame parsing, CRC validation, gateway data aggregation, and a shared Redis backend contract for simulation and production collectors, supporting 2,048 keys/s across 128 gateways $\times$ 16 sensors with TTL=10s while keeping RS485 collectors on the same key format

Beijing University of Technology

Sep 2017 – May 2019

R&D Assistant

- Developed a six-degree-of-freedom robotic arm control system using STM32, stepper motors, sensors, cameras, and a custom four-layer PCB to integrate embedded control, sensing, vision input, and actuator control; built and configured an 8-node compute cluster with Intel Xeon E5-2696 v3 CPUs, NVIDIA Tesla K80 GPUs, RAID 50 storage, and VPN/SSH-enabled remote access to support GPU-based research workloads and remote experimentation

RESEARCH EXPERIENCE

Indiana University School of Medicine

Nov 2020 – Aug 2021; Aug 2025 – May 2026

Research Assistant

- Developed a two-stage generative framework combining 3D VQGAN with latent Diffusion Transformer (DiT) to synthesize volumetric sMRI scans for Alzheimer's research, leveraging DeepSpeed for distributed training
- Implemented a self-supervised Point Transformer pipeline for 4096-point facial point clouds with SimCLR-based contrastive learning for forensic anthropometry, producing pose- and scale-robust geometric representations
- Developed, maintained, and refactored DeepChrlInteract as advisor requirements changed, keeping the Python deep learning toolkit modular and maintainable for genome-wide chromatin interaction prediction; supported DNA/RNA analysis, enhancer-promoter prediction, chromatin interaction prediction, and input representations including one-hot, k-mer, and DNA-LLM
- Benchmarked 1D-CNN, 2D-CNN, LSTM, BLSTM, FCNN, flat VGG-like networks, ResNet, DenseNet, Random Forest, SVM, CNN, BiLSTM, mLSTM, xLSTM, Transformer, Linear Transformer, iTransformer, Mamba, RWKV, MAE, DNABERT, DNABERT-2, NT, and HyenaDNA with 6 fusion strategies, evaluating AUROC, AUPRC, F1, Accuracy, Precision, and Recall

Purdue University

Jan 2024 – Aug 2024

Research Assistant

- Conducted mmWave radar research spanning image and point-cloud processing, pattern recognition, image generation, and model security for attack and defense; performed signal analysis and window-function optimization on IWR6843, IWR1843, IWR1642, and Cascade EVM 2243x4 platforms using FFT-based range, Doppler, and angle processing to extract point clouds and heatmaps; developed and validated physical backdoor attacks against mmWave-based human activity recognition systems to advance empirical understanding of sensing-system security vulnerabilities

Purdue University

Aug 2024 – May 2025

Research Assistant

- Conducted multimodal learning research across vision, audio, and text; evaluated deep learning security through backdoor attack implementation, defense, and adversarial robustness; fine-tuned and deployed large language models for downstream tasks with emphasis on practical adaptation and robustness against adversarial attacks

Purdue University

Jan 2021 – Jan 2024

Research Assistant

- Conducted Alzheimer's research on relationships between age, education, MMSE, BMI, and brain fMRI connectivity across HCI, MCI, and AD stages; developed visualization software for high-dimensional medical data with arbitrary 2D projections and PCA traversal
- Created animations for addictive-behavior analysis using nearest-neighbor, bilinear, bicubic, quintic, and affine Shepard interpolation to illustrate neurological changes
- Extracted and fine-tuned features with ResNet, EfficientNet, MobileNet, Xception, DenseNet, and CLIP (ViT) for medical imaging tasks
- Used NIH ABCD data to predict adolescent suicide risk and contributing factors, and applied machine learning to analyze the impact of COVID-19
- Applied deep learning to discover iris-gene pattern links, used diffusion and Stable Diffusion models to generate irises for specific genotypes, and combined signal processing with deep learning for fine-grained mmWave action distinction and intention recognition under high similarity

EDUCATION

Ph.D., Computer Science (CS)

Aug 2025 – May 2027 (expected)

Indiana University Luddy School of Informatics, Computing, and Engineering

Ph.D., Computer and Information Technology

Aug 2021 – May 2025

Purdue University; transferred with advisor to Indiana University

M.S., Computer Science (CS)

Aug 2019 – May 2021

Purdue University

B.Eng., Electrical and Computer Engineering (ECE)

Sep 2015 – May 2019

Beijing University of Technology

SELECTED PUBLICATIONS

LLM, Reasoning, and Multimodal AI

- Q. Niu, J. Liu, **Z. Bi**, et al., "Large Language Models and Cognitive Science: A Comprehensive Review," *BIO Integration*, 2026.
- C. Liang, P. Tian, et al., **Z. Bi**, et al., "A Comprehensive Survey and Guide to Multimodal Large Language Models in Vision-Language Tasks," *Computation*, 2026.
- Z. Bi**, H. Duan, J. Xu, et al., "GeneralBench: A Comprehensive Benchmark Suite and Evaluation Platform for Large Language Models," 2025.
- Z. Bi**, L. Chen, J. Song, et al., "Exploring Efficiency Frontiers of Thinking Budget in Medical Reasoning," *arXiv:2508.12140*, 2025.
- Z. Bi**, K. Chen, C. Tseng, et al., "Is GPT-OSS Good? A Comprehensive Evaluation of OpenAI's Latest Open Source Models," *arXiv:2508.12461*, 2025.
- Z. Bi**, K. Chen, T. Wang, J. Hao, X. Song, "CoT-X: An Adaptive Framework for Cross-Model Chain-of-Thought Transfer and Optimization," *arXiv:2511.05747*, 2025.
- S. Wu, C. Liang, **Z. Bi**, et al., "AutoSurvey2: Empowering Researchers with Next Level Automated Literature Surveys," *arXiv:2510.26012*, 2025.

mmWave Sensing, Security, and ML Systems

- Z. Bi**, A. Singha, H. Xue, et al., "Physical Backdoor Attacks against mmWave-based Human Activity Recognition," *IEEE ICDCS*, 2025.
- Z. Bi**, R. Gao, S. Fang, "A Subspace Projection Based Technique for Visualizing Machine Learning Models," *VISIGRAPP*, 2025.
- J. Xu, **Z. Bi**, A. Singha, et al., "mmlock: User Leaving Detection against Data Theft via High-Quality mmWave Radar Imaging," *IEEE ICCCN*, 2023.
- A. Singha, **Z. Bi**, et al., "Securing Contrastive mmWave-based Human Activity Recognition against Adversarial Label Flipping," *ACM WiSec*, 2024.

Medical AI Collaborative Research (ML/DL Algorithms and Tooling)

- A. Korada, Y. Chen, **Z. Bi**, et al., "Temporal and Spatial Patterns of Secondary Motor Cortex Calcium Activity in Cocaine Self-Administration," *Neuropharmacology*, Elsevier, 2025.
- P. Pan, S. Li, **Z. Bi**, et al., "CD9-positive Tumor-Associated Macrophages Promote Renal Cell Carcinoma Progression," *Cancer Cell International*, 2025.
- J. Huang, L. Chen, **Z. Bi**, et al., "Use of Machine Learning Models to Predict Mortality in Dialysis Patients," *Frontiers in Public Health*, 2025.
- H. Wang, G. Lao, **Z. Bi**, et al., "Predicting ICU in-hospital mortality from text-encoded structured EHR data using adaptive transformer layer fusion," *iScience*, 29(6), 2026.