

Hanzhi Zhang (she/her)

Ph.D. Candidate in Computer Science | University of North Texas, Denton, TX, US

hanzhizhang@my.unt.edu | [Personal Website](#) | [Google Scholar](#)

RESEARCH INTERESTS

Efficient AI Systems: Hardware-aware LLM inference, sparse and mixed-precision computation, post-training quantization, GPU optimization, and HPC.

Reliable and Responsible AI: Multilingual fact verification, hallucination detection, benchmark development, safety-critical perception for autonomous systems, and AI for science.

EDUCATION

University of North Texas (UNT) | Aug. 2023 - present

Ph.D. in Computer Science (candidate; **expected May 2027**)

Advisor: Dr. Yunhe Feng

University of Birmingham | Sept. 2021 - Dec. 2022

M.Sc. in Data Science, awarded with **Distinction**

Xiamen University | Sept. 2017 - Aug. 2021

B.Sc. in Computer Science and Technology, with **Honours**

PEER-REVIEWED PUBLICATIONS

Seven peer-reviewed publications, including four as first or co-first author.

Bold marks my name; * denotes equal contribution. [F] first/co-first author; [C] co-authored.

[F1] Bing Fan*, Minghao Li*, **Hanzhi Zhang***, Shaohua Dong, Naga Prudhvi Mareedu, Weishi Shi, Yunhe Feng, Yan Huang, and Heng Fan. Towards Visual Query Segmentation in the Wild. *Advances in Neural Information Processing Systems (NeurIPS)*. 2026. [\[Paper\]](#)

[F2] **Hanzhi Zhang**, Sumera Anjum, Heng Fan, Weijian Zheng, Yan Huang, and Yunhe Feng. Poly-FEVER: A Multilingual Fact Verification Benchmark for Hallucination Detection in Large Language Models. *Findings of the Association for Computational Linguistics: EMNLP 2026*. 2026. [\[Paper\]](#)

[F3] **Hanzhi Zhang**, Heng Fan, Kewei Sha, Yan Huang, and Yunhe Feng. DAM: Dynamic Attention Mask for Long-Context Large Language Model Inference Acceleration. *Findings of the Association for Computational Linguistics: ACL 2025*. 2025. [\[Paper\]](#)

[F4] **Hanzhi Zhang**, Harold Lucero, Kewei Sha, Heng Fan, Song Fu, and Yunhe Feng. D2SO: Detecting Distant and Small Objects for Vision-Based Vehicle Autonomous Systems. *IEEE International Conference on Mobility: Operations, Services and Technologies (MOST)*. 2025. [\[Paper\]](#)

[C1] Mingchen Li, **Hanzhi Zhang**, Heng Fan, Junhua Ding, and Yunhe Feng. Harmful Factuality: LLMs Correcting What They Shouldn't. *Findings of the Association for Computational Linguistics: EACL 2026*. 2026. [\[Paper\]](#)

[C2] Sumera Anjum, **Hanzhi Zhang**, Wenjun Zhou, Eun Jin Paek, Xiaopeng Zhao, and Yunhe Feng. HALO: Hallucination Analysis and Learning Optimization to Empower LLMs with Retrieval-Augmented Context for Guided Clinical Decision Making. *IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*. 2025. [\[Paper\]](#)

[C3] Tong Xie, **Hanzhi Zhang**, Shaozhou Wang, Yuwei Wan, Imran Razzak, Chunyu Kit, Wenjie Zhang, and Bram Hoex. ByteScience: Bridging Unstructured Scientific Literature and Structured Data with Auto Fine-Tuned Large Language Model in Token Granularity. *IEEE International Conference on Data Mining Workshops (ICDMW)*. 2024. [\[Paper\]](#)

TEACHING EXPERIENCE

Teaching Assistant, CSCE 4201: Introduction to Artificial Intelligence | Fall 2026

Undergraduate course, **74 students**, UNT.

Independently designed and delivered **three 3-hour lectures on CNNs, NLP, and Transformers**. Student feedback highlighted interactive whiteboard examples and explanations that helped students better understand the material.

Teaching Assistant, CSCE 5300: Introduction to Big Data and Data Science

Fall 2024 and Spring 2025

Graduate course, **79 students (Fall 2024)** and **64 students (Spring 2025)**, UNT.

Received an overall teaching evaluation of **Excellent in both semesters**. Managed grading and office hours; coordinated randomized-question exams and analyzed score distributions.

Guest Lecturer: Long-Context Evaluation of Large Language Models | Summer 2026

NSF Research Experiences for Undergraduates (REU) program, UNT.

MENTORING AND ADVISING

REU Academic Coordinator and Research Mentor | Summer 2026

NSF Research Experiences for Undergraduates (REU), UNT.

Coordinated research training for **nine undergraduates**, including a lecture on long-context LLM evaluation, weekly research talks, and ongoing mentoring.

Student Research Mentoring and Academic Guidance | 2023 - present

Responsible AI Lab, UNT.

Provided research and academic guidance to **8 graduate students, 2 undergraduates, and 1 TAMS student**. Mentored two students on research projects resulting in **Poly-FEVER (EMNLP Findings 2026)** and **D2SO (IEEE MOST 2025)**.

Developed [HPC](#) and [Docker](#) tutorials to enable independent research experiments.

HONORS AND AWARDS

Top Reviewer, NeurIPS 2026 | October 2026

Student Travel Grant Award, IEEE MOST 2025 | April 2025

Selected Participant, ATPESC 2024, Argonne National Laboratory | 2024

First Place, Poster Presentation, UNT Engineering Showcase | March 2024

PROFESSIONAL SERVICE

Reviewer

Machine learning and NLP: ICLR 2027; NeurIPS 2026 (Top Reviewer); ICML 2026; ACL Rolling Review 2026 (May, August, and October cycles).

Data mining and systems: CIKM 2025; ICDM 2024; IEEE ICPADS 2023.

External Reviewer

ACM SEC 2026; ICDM 2025; WWW 2024; AAAI 2024; WACV 2024; ICNC 2024; EISA 2023; Inscrypt 2023.

Conference and Workshop Volunteering

Student Volunteer, NSF CISE MSI Aspiring PI Workshop, UNT | December 2024

Student Volunteer, IEEE MOST 2024 | May 2024

Professional Memberships and Community

Member, Women in Data Science (WinDS), UNT chapter | 2025 - present

Participant, Women-in-Computing session, IEEE MOST 2025 | April 2025

Outreach speaker on Responsible AI for K-12 educators and the regional AI community through AIxHEART and the Presidential AI Challenge | 2025 - 2026

RESEARCH AND PROFESSIONAL EXPERIENCE

Responsible AI Lab, University of North Texas | Aug. 2023 - present

Graduate Research Assistant | Denton, TX

Lead hardware-aware LLM inference research spanning sparse attention (DAM), mixed-precision attention, and post-training quantization. Led research that established the technical foundation for the lab's Amazon Research Awards grant (TileRoute-Trainium) and contributed to writing the proposal. Developed two benchmarks: **Poly-FEVER (77,973 claims, 11 languages)** and **VQS-4K (4,111 videos, more than 1.3 million frames; NeurIPS 2026)**. Conduct large-scale GPU experiments on ALCF Polaris and TACC Lonestar6.

Oak Ridge National Laboratory (ORNL) | 2026 - present

Research Collaborator, Analytics and AI Methods at Scale (AAIMS) Group | Remote

Selected for ORNL's Summer 2026 research internship (offer received May 2026); the work continues as an ongoing remote research collaboration. Within Dr. Feiyi Wang's AAIMS Group, collaborate with Dr. Emily Herron and Dr. Aditya Kashi on AI-based 3D boundary-to-domain operator learning for partial differential equations (PDEs).

Amazon Research Awards (ARA) | 2026 - present

Research Collaborator, TileRoute-Trainium Project | Remote

Collaborate with Qian Qu (Amazon scientist) on efficient, mixed-precision LLM inference using AWS Trainium accelerators.

Argonne National Laboratory, ATPESC 2024 | Summer 2024

Selected Participant | Lemont, IL

Competitively selected for an intensive two-week program on extreme-scale computing: GPU computing, parallel programming models, and performance optimization on leadership-class supercomputers.

GRANTS AND COMPUTING ALLOCATIONS

Amazon Research Awards: TileRoute-Trainium | 2026

PI: Dr. Yunhe Feng (UNT). **Role:** Lead student researcher and proposal co-writer.

The proposal builds directly on Hanzhi's tile-centric mixed-precision research.

Computing Resources: ALCF Polaris | 2024 - present

Role: PI; 5,000 node-hours per allocation.

Project: Large Language Models' Hallucination Detection and Mitigation Ability across Multi Languages.

INVITED TALKS AND PRESENTATIONS

May 2026: Invited talk, *Responsible AI*, Presidential AI Challenge (Denton ISD and UNT), for educators and industry professionals.

October 2025: Invited talk, *Responsible AI: The Hidden Curriculum*, AIXHEART, professional development for middle-school teachers.

July 2025: Conference presentation, *DAM: Dynamic Attention Mask for Long-Context LLM Inference Acceleration*, ACL 2025 (Findings).

May 2025: Conference presentation, *D2SO: Detecting Distant and Small Objects for Vision-Based Vehicle Autonomous Systems*, IEEE MOST 2025.

November 2024: Proposal presentation, *Efficient Small Language Models for Edge Devices via Task-Aware Sparse Attention*, NSF IUCRC eCAT meeting.

Research presentations: LLaVi Lab AI Seminar, UNT (2024 - 2026); Data Science and AI Talk Series, UNT (2023 - 2024), to faculty and Ph.D. students.

Poster presentations: UNT CSE Open House (2023 - 2026); UNT University Research Day (2024, 2025); UNT Engineering Showcase (2024, First Place).