Zaifu Zhan

zhan8023@umn.edu | Personal Website | Linkedin | GitHub

EDUCATION

University of Minnesota, Twin Cities	Ph.D	in	Electrical and Computer Engineering	GPA: 3.9/4.0	*2021 – 2027*
	Minor	in	Computer Science	GPA: 4.0/4.0	*2021 – 2027*
Tsinghua University	M.Eng.	in	Electrical Engineering	GPA: 3.6/4.0	*2018 – 2021*
Beijing Jiaotong University	BS	in	Electrical Engineering	GPA: 87/100	*2014 – 2018*

SKILLS

- **Programming**: Python(Proficient), MATLAB(Proficient), SQL, R, C/C++, Java, Kotlin
- **Frameworks**: PyTorch, TensorFlow, Scikit-learn, Hugging Face, vLLM
- **Math**: Linear Algebra, Probability & Statistics, Convex and integer optimization, Common ML Algorithms, DL basics
- **Data Processing**: Pandas, NumPy, Spark, Data Cleaning, Data Visualization (Matplotlib, Seaborn)
- **Others**: Linux Bash, Shell, CUDA C, Git, GitHub, Slurm, LaTeX, Processing, Jupyter

RELATED EXPERIENCES [More experiences on LinkedIn]

AI Research Intern, *Atlassian* *May. 2025 – Nov2025*

- Memory feature for Rovo ChatBot
  - Designed and implemented an Entity-Relationship Store table to store structured summaries of past conversation topic segments.
  - Developed a summary-based retrieval mechanism to efficiently fetch the most relevant historical conversation segments, enhancing chatbot personalization and contextual continuity.
  - Developed a dataset benchmark to evaluate LLM memory in company-specific scenarios by generating profile-based distractor conversations, inserting memory elements, and designing evaluation queries.

Graduate Research Assistant, *University of Minnesota* *Sep. 2021 – present*

- Multi-agent peer-review system for medical question answering
  - Designed a multi-agent peer-reviewed reasoning framework for medical question answering, where multiple LLM agents generate chain-of-thought solutions and then cross-evaluate each other's reasoning for a reliable reasoning chain
  - Achieved the best average accuracy of 0.820 across HeadQA, MedQA-USMLE, and PubMedQA, consistently outperforming single-model CoT and CoT-based majority voting ensembles
- CancerLLM: A Large Language Model in the Cancer Domain [Paper]
  - Trained a 7-billion-parameter Mistral-style model on 2.7M clinical notes and 515K pathology reports
  - Fine-tuned the CancerLLM on cancer phenotype extraction and diagnosis generation tasks
  - CancerLLM outperformed existing LLMs, with an average F1 score improvement of 9.23%
- Retrieval-augmented generation with LLM
  - Designed a retrieval-augmented multi-modal LLM framework that selects informative in-context examples across text and medical images, improving disease classification performance under few-shot and cross-modal settings. [paper]
  - Proposed the first retrieval-augmented multi-task LLM framework for dietary supplement information extraction, jointly solving NER, relation extraction, triple extraction, and usage classification across 8 state-of-the-art LLMs, achieving higher storage efficiency than single-task fine-tuning. [paper]
  - Conducted a systematic benchmark of retrieval-augmented LLMs (RALs) fine-tuned with LoRA across 5 biomedical tasks and 9 datasets, evaluating robustness under unlabeled, counterfactual, diverse, and negative-awareness. [paper]
  - Developed a multi-mode RAG framework (MMRAG) that dynamically integrates multiple retrieval strategies to improve in-context learning for biomedical QA and classification tasks. [paper]
- Early exiting for faster LLM inference and Rejection
  - EPEE: Proposed a hybrid entropy + patience early-exit strategy to mitigate "overthinking" and speed up inference across BERT/ALBERT/GPT-2, and extended to ViT for medical imaging. [paper]
  - Integrated a patience-based rejection (abstention) mechanism into early exiting to improve trustworthiness under uncertainty, validated across BERT/Llama-3.2/ViT on 11 medical decision-making datasets. [paper]

- Optimized dataset combinations for multi-task learning via reinforcement learning [paper]
  - Generated random combinations from 12 datasets and fine-tuned the Llama3 model to collect combination-F1 score pairs across four tasks: named entity recognition, relation extraction, event extraction, and classification.
  - Used a multi-layer neural network to predict the best combination, fine-tuned the LLM, and iteratively optimized dataset combinations.
- Computing stabilizing linear controllers via policy iteration [Github]
  - Proposed an iterative Q-learning reinforcement learning algorithm to find the optimal controller for noisy LTI systems using the action-value Bellman equation, demonstrating effectiveness under high noise conditions.
  - Transformed the core learning step into a least-squares problem, achieving high computational efficiency.
  - Proved the convergence of the Q-learning algorithm using conditional probability.
- Adversarial Learning Project: Out of Distribution Detectors vs. Attackers [Github]
  - Generated adversarial images from the CIFAR-10 dataset using over 30 attack methods.
  - Evaluated the performance of 18 detectors under all attacks using metrics like AUROC.
  - Automated the generation of shell scripts to run experiments on multiple GPUs for parallel computing.
  - Developed a novel two-sided threshold method, improving AUROC scores.

Research Assistant, Xu's Research Group, *UWM* *Mar. 2020 – Dec. 2020*

- Developed a control-barrier-function-based, velocity-constrained algorithm for trajectory planning and controller design.
- Implemented a visual servoing trajectory planning algorithm for quadrotor navigation.
- Achieved stable results in experiments, demonstrating the robustness and effectiveness of the proposed algorithms.

Software Testing Engineer, *Beijing Internet-Based Engineering Company* *Jul. 2019 – June. 2021*

- Tested Simdroid software through simulations, identifying numerous bugs and collaborating with development teams to improve software quality.

SELECTED PUBLICATIONS [Google Scholar]

[1] **Zhan, Z.**, Zhou, S., & Zhang, R. (2026). PEER: Towards reliable and efficient inference via Patience-Based Early Exiting with Rejection. Journal of Biomedical Informatics, 104988.

[2] **Zhan, Z.**, Wang, J., Zhou, S., Deng, J., & Zhang, R. (2025). Mmrag: Multi-mode retrieval-augmented generation with large language models for biomedical in-context learning. Journal of the American Medical Informatics Association, 32(10), 1505-1516.

[3] **Zhan, Z.**, & Zhang, R. (2025). Towards Better Multi-task Learning: A Framework for Optimizing Dataset Combinations in Large Language Models. In Findings of the Association for Computational Linguistics: NAACL 2025.

[4] **Zhan, Z.**, Zhou, S., Li, M., & Zhang, R. (2025). RAMIE: retrieval-augmented multi-task information extraction with large language models on dietary supplements. Journal of the American Medical Informatics Association, ocaf002.

[5] **Zhan, Z.**, Zhou, S., Zhou, H., Liu, Z., & Zhang, R. (2025). EPEE: Towards efficient and effective foundation models in biomedicine. arXiv preprint arXiv:2503.02053.

[6] **Zhan, Z.**, Zhou, S., Zhou, H., Deng, J., Hou, Y., Yeung, J., & Zhang, R. (2025). An evaluation of deepseek models in biomedical natural language processing. arXiv preprint arXiv:2503.00624.

[7] Zhou, S., Xie, W., Li, J., **Zhan, Z.**, Song, M., Yang, H., ... & Zhang, R. (2025). Automating expert-level medical reasoning evaluation of large language models. npj Digital Medicine.

[8] Hou, Y., **Zhan, Z.**, Zeng, M., Wu, Y., Zhou, S., & Zhang, R. (2025). Benchmarking GPT-5 for biomedical natural language processing. arXiv preprint arXiv:2509.04462.

[9] Li, M., **Zhan, Z.**, Yang, H., Xiao, Y., Zhou, H., Huang, J., & Zhang, R. (2025). Benchmarking retrieval-augmented large language models in biomedical nlp: Application, robustness, and self-awareness. Science Advances, 11(47), eadr1443.

[10] Zhou, S., Xu, Z., Zhang, M., Xu, C., Guo, Y., **Zhan, Z.**, ... & Zhang, R. (2025). Large language models for disease diagnosis: A scoping review. npj Artificial Intelligence, 1(1), 1-17.

[11] Zhou, H., Gu, H., **Zhan, Z.**, … (2025). The Efficiency vs. Accuracy Trade-off: Optimizing RAG-Enhanced LLM Recommender Systems Using Multi-Head Early Exit. ACL 2025

[12] Hou, Y., Patel, J., Dai, L., Zhang, E., Liu, Y., **Zhan, Z.**, ... & Zhang, R. (2025). Benchmarking of Large Language Models for the Dental Admission Test. Health Data Science, 5, 0250.