XIAOYING SONG

xiaoyingsong1001@gmail.com Homepage LinkedIn Google Scholar

RESEARCH INTERESTS

Areas: LLM Agents · Personalized Alignment · LLM Post-Training · Human-Centered Evaluation

I develop LLM systems that collaborate effectively, adapt to user needs, and generate reliable responses. My research spans four directions: (i) **LLM Agents**: optimizing when to ask versus answer in multi-turn dialogue, and coordinating retrieval and response refinement through multi-agent RAG [EMNLP 2026; COLM 2025]; (ii) **Personalized Alignment**: adapting responses to users' literacy levels and preferences, and controlling safety behavior for authorized contexts [EMNLP 2025; NeurIPS 2026]; (iii) **LLM Post-Training**: using reinforcement learning and fine-tuning to optimize dialogue decisions and controllable generation [EMNLP 2026; EMNLP 2025; EMNLP 2024]; (iv) **Human-Centered Evaluation**: assessing response human likeness and consistency through automated metrics and human evaluation [COLING 2025; EMNLP 2025].

ACADEMIC EXPERIENCE

University of North Texas, TX, USA — *Aug 2023 – Present*
Ph.D. Candidate, Data Science; Research Assistant
Advisor: Prof. Lingzi Hong

Central China Normal University, China — *Aug 2021 – Jun 2023*
M.S., Information Science; Research Assistant

PUBLICATIONS

Conference Papers

1. **Song, X.**[*], Tan, Q.[*], et al. Palette: A modular, controllable, and efficient framework for on-demand authorized safety alignment relaxation in LLMs. In *Advances in Neural Information Processing Systems* (**NeurIPS**), 2026.
2. **Song, X.**, Anik, A. S., Liu, J., Tan, Q., et al. Ask or answer: A decision framework for multi-turn health misinformation intervention. In *Conference on Empirical Methods in Natural Language Processing* (**EMNLP**), 2026.
3. Tan, Q., **Song, X.**, Cheng, N., Li, G., et al. End-to-end on-device quantization-aware training for LLMs at inference cost. In *Conference on Empirical Methods in Natural Language Processing* (**EMNLP**), 2026.
4. Tan, Q., **Song, X.**, Cheng, N., Liu, N., et al. Q-Realign: Piggybacking realignment on quantization for safe and efficient LLM deployment. In *Conference on Empirical Methods in Natural Language Processing* (**EMNLP**), 2026.
5. **Song, X.**, Anik, A. S., Barua, D., Luo, P., et al. Speaking at the right level: Literacy-controlled counterspeech generation with RAG-RL. In *Conference on Empirical Methods in Natural Language Processing* (**EMNLP**), 2025.
6. **Song, X.**, Anik, A. S., Blanco, E., et al. A dynamic fusion model for consistent crisis response. In *Conference on Empirical Methods in Natural Language Processing* (**EMNLP**), 2025.
7. **Song, X.**, Perez, S. L., Yu, X., et al. Echoes of discord: Forecasting hater reactions to counterspeech. In *Conference of the North American Chapter of the Association for Computational Linguistics* (**NAACL**), 2025.
8. **Song, X.**, Mamidisetty, S., Blanco, E., & Hong, L. Assessing the human likeness of AI-generated counterspeech. In *International Conference on Computational Linguistics* (**COLING**), 2025.
9. **Song, X.**, Anik, A. S., Frías-Martínez, V., & Hong, L. Dynamic fusion of large language models for crisis communication. In *International Conference on Information Systems for Crisis Response and Management* (ISCRAM), 2025.
10. **Song, X.**[*], Anik, A. S.[*], Wang, E., Wang, B., et al. Multi-agent retrieval-augmented framework for evidence-based counterspeech against health misinformation. In *Conference on Language Modeling* (**COLM**), 2025.
11. Hong, L., Luo, P., Blanco, E., & **Song, X.** Outcome-constrained large language models for countering hate speech. In *Conference on Empirical Methods in Natural Language Processing* (**EMNLP**), 2024.
12. Liu, J., **Song, X.**, Zhang, D., Thomale, J., et al. A hybrid framework for subject analysis: Integrating embedding-based regression models with large language models. In *Annual Meeting of the Association for Information Science and Technology* (ASIS&T), 2025.
13. **Song, X.**, et al. ST-Crisis: A diagnostic benchmark for spatiotemporal grounding in crisis communication. Under review at **EACL** 2026.

14. **Song, X.**, et al. Towards human-centered natural language processing: A survey of cognitive signal integration. Under review at **EACL** 2026.
15. Anik, A. S., **Song, X.**, Kanuganti, S., Yuan, J., et al. Granularity-aligned health misinformation refutation with knowledge graphs. Under review at **NAACL** 2026.

Journal Papers

1. **Song, X.**, Luo, P., Thomale, J., et al. Comparative analysis of large language models' performance in book classification tasks using Library of Congress Classification system. *Journal of Information Science*, 2026.
2. Perez, S. L., **Song, X.**, & Hong, L. Analyzing the language of rejection: A study of user flagging responses to hate speech on Reddit. *Information Research*, 2025.
3. Tan, Q., **Song, X.**, Ye, G., et al. An effective negative sampling approach for contrastive learning of sentence embedding. *Machine Learning*, 2023.

PROJECT EXPERIENCE

Reinforcement Learning for Multi-Turn Dialogue *Project Leader · Jan 2026 – Present*

Led the design of **RO-PnR**, an RL framework balancing information gain and interaction cost to decide when to ask versus answer. Achieved the highest cost-adjusted utility among evaluated baselines with ∼30% fewer turns than always-probe policies. **Outcome:** EMNLP 2026.

Multi-Agent RAG and Literacy-Controlled Generation *Project Leader · Aug 2024 – Present*

Led multi-agent RAG research for evidence retrieval and response refinement, and developed literacy-controlled generation using **LoRA SFT and GRPO**. Validated LLM-based preference ratings against human judgments. **Outcomes:** EMNLP 2025 and COLM 2025.

Personalized Safety Alignment for LLMs *Project Collaborator · Aug 2025 – Present*

Co-developed **Palette**, combining refusal-direction search, lightweight LoRA adaptation, and parameter merging for authorized, domain-specific safety control. Enabled multi-domain composition without retraining, evaluated across four safety benchmarks. **Outcome:** NeurIPS 2026.

HONORS AND AWARDS

Second Place, CLEF 2026 CheckThat! Lab Competition	*2026*
Honorable Mention, ASIS&T SIG-SM Student Competition	*2026*
First Place, UNT Phil and Lis Turner Outstanding Paper Award	*2025*
UNT Porter-Evans Scholarship	*2025*
UNT COI/IS Leon M. Liddell Scholarship	*2024*
UNT Graduate School Travel Award	*2024*

INVITED TALKS, TEACHING & MENTORSHIP

Invited Tutorials *IDEA Institute on AI (ASIS&T), with Prof. Lingzi Hong*

- *AI-assisted Communications* — *2025, 2026*
- *Data Collection, Classification, and Transformation* — *2024*
- *Comprehending the Semantic Essence of Text: Leveraging SEANCE for Textual Analysis* — *2024*

Teaching Assistant *UNT, Aug 2023 – Jul 2024*

Introduction to Data Science; Data Visualization; Information Architecture

Mentoring & Advising *Jan 2024 – Present*

- Mentor, UNT AI Summer Program – guided master's and undergraduate students through AI research projects; mentees co-authored papers at COLM and EMNLP — *2024, 2025, 2026*
- Advised Sujana Mamidisetty, Diana Zhang, and Elliott Wang – admitted to computer science programs at UMD, UT, and Purdue — *2023 – 2026*

TECHNICAL SKILLS & SERVICE

LLM Training & Alignment: Supervised fine-tuning and RL-based alignment (PPO, GRPO) with PEFT/LoRA.

LLM Systems & Evaluation: Retrieval-augmented generation, multi-agent frameworks, synthetic dataset and benchmark construction, LLM-as-judge evaluation validated against human annotation.

Academic Service: Reviewer for ACL, NAACL, EMNLP, COLING, COLM, NeurIPS, JCDL, and ASIS&T.