DANIEL (LEYUAN) TANG

(903) 969-5236 ⋄ leyuan@tamu.edu
L.F. Peterson Building, Room 330, 435 Nagle St, College Station, TX 77843
https://danielustang.github.io

RESEARCH INTEREST

Natural Language Processing, Large Language Models, LLM Evaluation & Alignment Robustness, Post-Training Bias & Sycophancy Mitigation, Multi-Agent Simulation, Computational Social Science

EDUCATION

Texas A&M University — College Station, TX
B.S. in Computer Science — Expected May 2027
Corps of Cadets — Fall 2025 – Present

- Professional Development Officer (S-5), Squadron 16 (Fall 2026 – Present): lead career-readiness workshops for a 60-cadet outfit; brief and mentor cadets on internships, resumes, and career planning.

Tyler Junior College — Tyler, TX
A.S. in Computer Science — May 2025

PUBLICATIONS & PREPRINTS

- **Leyuan Tang**, Kangda Wei, Tianyu Jiang, Ruihong Huang. *SPINE: Measuring LLM Sycophancy under Sustained Multi-Turn Pressure.* Under ACL Rolling Review. arXiv:2609.09090. [Read] [Media]

RESEARCH EXPERIENCE

Undergraduate Researcher, Natural Language Processing — Spring 2026 – Present
Texas A&M University — Research Advisor: Prof. Ruihong Huang — College Station, TX

Measuring LLM Sycophancy under Sustained Multi-Turn Pressure via Multi-Agent Systems

- Built a multi-turn sycophancy benchmark in which an LLM user-proxy adaptively challenges a target model for up to 25 turns; found that collapse rates rise with conversation length, from 0–57% within 5 turns to 20–97% within 25 turns.
- Compared model responses across proxy tactics and identified emotional appeals as the most effective at inducing sycophancy (position weakened in 44.3% of turns).
- Showed through reasoning-trace analysis that the correct position is still present in 50–89% of collapse turns, suggesting models choose to please the user rather than lack the knowledge.
- *Status:* Preprint on arXiv; under review at ACL Rolling Review.

Mitigating Sycophancy and Improving Reasoning in LLMs via Critical Thinking and RL

- Proposed a data generation framework that builds critical thinking in LLMs by having them argue through multi-turn debates over false presuppositions and unethical queries.
- Fine-tuned open-weight LLMs with SFT and RL on the generated data, substantially reducing sycophancy collapse rates on the SPINE benchmark.
- *Status:* Ongoing; working toward an ACL 2027 submission (January 2027 ARR cycle).

Multi-Agent Systems with LLMs for Cyberbullying Intervention Training and Safe Social Simulation

- Designed and built a multi-agent LLM simulator to help adolescents practice responding to cyberbullying alongside AI agents.
- Implemented safeguards against harmful agent output through centralized, message-level safety review and severity-capped agent behavior.
- Contributed to an NIH R21 grant proposal by drafting the technical/system-design section (2 of 6 Research Strategy pages), with the simulator serving as preliminary work.
- Contributing to the IRB application for formative user testing of the simulator with college student participants.
- *Ongoing:* NIH R21 proposal submitted Oct 2026; aiming for an ACL 2027 System Demonstrations track submission.

TALKS

LLM Sycophancy under Sustained Multi-Turn Pressure Oct 2026
Gave a talk on Clubhouse about our SPINE paper to an audience of over 7,000 listeners. [Listen] [Poster]

PODCAST

Rubato
A tech podcast where guests share their lives, career choices, recent anecdotes, and the detours they've taken, and reflect on how they view rubato in life. [Website]
Status: Ongoing; Episode 1 in production.

AWARDS

- Sul Ross Corps of Cadets Scholarship, Texas A&M University (Fall 2026 – Spring 2027)
- Maxine Foster & John McDuffie Jr. Scholarship, Texas A&M University (Spring 2026; 2026–27)
- Helaine & Gerald Ray '54 Corps Scholarship, Texas A&M University (Fall 2025)
- JKCF Undergraduate Transfer Scholarship Semifinalist
- George Tefteller Mathematics Annual Scholarship, Tyler Junior College

SKILLS

Programming: C, C++, Python, Java, JavaScript, Haskell

ML & LLM: PyTorch, Hugging Face Transformers, LLM APIs, LLM-as-judge evaluation, multi-agent simulation

LLM Training & Post-training: SFT, RLHF, prompt engineering, synthetic data generation

Software & Tools: Linux, Git, LaTeX, Jupyter, Slurm/HPC, Node.js, MySQL, Redis, Nginx, Spring Boot

Cloud & Infrastructure: AWS, Microsoft Azure

Languages: English (native), Mandarin (native), Japanese (conversational)