BANGZHAO SHU

Medford, MA | +1 734-210-3367 | shu.b@northeastern.edu | Google Scholar | GitHub | LinkedIn | Website

Third-year Computer Science Ph.D. student at Northeastern University specializing in natural language processing (NLP), machine learning, and generative AI, with a focus on large language models (LLMs). Develops benchmarks, causal analyses, and mechanistic interpretability methods to evaluate and improve LLM reliability, safety, and social and emotional intelligence. First/co-first author at EMNLP, ICWSM, and NAACL.

Research Focus: LLM Evaluation | AI Safety | Mechanistic Interpretability | Affective NLP | Computational Social Science

EDUCATION

Ph.D. in Computer Science, Northeastern University — Expected 2029
Advisor: Mai ElSherief

M.S. in Information, University of Michigan — 2022 - 2024

M.S. in Environment and Sustainability, University of Michigan — 2021 - 2024

B.E. in Geomatics, Nanjing Normal University — 2016 - 2020

TECHNICAL SKILLS

Programming	Python, R, SQL
ML & NLP	PyTorch, Hugging Face Transformers, scikit-learn, PEFT, SentenceTransformers, LangChain, deep learning, generative AI
Research Methods	LLM evaluation and benchmarking, mechanistic interpretability, sparse autoencoders, causal inference, supervised fine-tuning, prompt-based evaluation, statistical analysis
Data	Large-scale data processing, dataset development and annotation, experimental design

RESEARCH EXPERIENCE

Computing for Social Good Lab, Northeastern University — 2024 - Present
Ph.D. Researcher | Advisor: Mai ElSherief

- Lead research on how LLMs represent and infer emotion, spanning fine-grained benchmark development, cross-context bias analysis, and mechanistic interpretability.
- Identified and causally validated emotion-specific features with sparse autoencoders; used targeted feature steering to mitigate failures and improve emotion recognition.
- Co-created EXPRESS, a 251-label emotion benchmark grounded in psychological theory; evaluated 14 language models across prompting strategies and released data and code.

Blablablab, University of Michigan — 2023 - 2025
Researcher | Advisor: David Jurgens

- Co-developed MODEL-PERSONAS: 693 questions spanning 39 psychometric instruments and 115 persona axes; evaluated 17 LLMs under prompt perturbations, revealing instability and low negation consistency.
- Co-developed SIZZLER, a dataset of 1,800 annotated emails; benchmarked classical and LLM methods, then applied causal analysis to 11.3M emails to identify linguistic and social factors associated with response behavior.

GISense Lab, The University of Texas at Austin — 2023 - 2024
Researcher | Advisor: Yuhao Kang

- Fine-tuned LLMs to generate and enrich semantic information from geospatial data.

SELECTED PUBLICATIONS

[1] Why Are Some Emotions Harder for LLMs? Uncovering the Causal Mechanisms of Emotion Inference via Sparse Autoencoders
Bangzhao Shu, Arinjay Singh, Mai ElSherief
EMNLP 2026 (accepted) | First author
Identified causal sparse features underlying emotion inference and improved recognition with targeted feature interventions and steering-vector optimization.

[2] Fluent but Unfeeling: The Emotional Blind Spots of Language Models
Bangzhao Shu*, Isha Joshi*, Melissa Karnaze, Anh C. Pham, Ishita Kakkar, Sindhu Kothe, Arpine Hovasapian, Mai ElSherief
ICWSM 2026 | Co-first author
Introduced EXPRESS, a fine-grained benchmark of 251 self-disclosed emotion labels, and evaluated 14 language models across prompting strategies.

SELECTED PUBLICATIONS (CONTINUED)

[3] You don't need a personality test to know these models are unreliable: Assessing the Reliability of Large Language Models on Psychometric Instruments
Bangzhao Shu*, Lechen Zhang*, Minje Choi, Lavinia Dunagan, Lajanugen Logeswaran, Moontae Lee, Dallas Card, David Jurgens
NAACL 2024 (Oral) | Co-first author
Built MODEL-PERSONAS and showed that minor prompt perturbations substantially reduce the consistency and reliability of LLM responses.

[4] Causally Modeling the Linguistic and Social Factors that Predict Email Response
Yinuo Xu*, Hong Chen*, Sushrita Rakshit*, Aparna Ananthasubramaniam*, Omkar Yadav*, Mingqian Zheng*, Michael Jiang*, Lechen Zhang*, Bowen Yi*, Kenan Alkiek*, Abraham Israeli*, Bangzhao Shu*, Hua Shen*, Jiaxin Pei*, Haotian Zhang*, Miriam Schirmer*, David Jurgens
NAACL 2025 | Equal contribution
Released SIZZLER and combined predictive modeling with causal analysis of 11.3M emails to study response dynamics.

[5] "He gets to be the fun parent": Understanding and Supporting Burnt-Out Mothers in Online Communities
Nazanin Sabri, Ananya Malik, Bangzhao Shu, Jason Snyder, Laurie Kramer, Mai ElSherief
ACM Transactions on Social Computing (accepted) | Co-author
Trained models to identify maternal-burnout discussions and analyzed support and co-parenting dynamics in online communities.

** Equal contribution*

MANUSCRIPT UNDER REVIEW

[6] HEAT framework for evaluating systematic biases in LLM emotion recognition
Bangzhao Shu et al.
Under review at Nature Mental Health | First author
Developed an evaluation framework revealing consistent LLM biases toward negative and higher-intensity emotions across topics and models.

ADDITIONAL EXPERIENCE

Cooperative Institute for Great Lakes Research 2021 - 2023
Graduate Student Intern
- Developed generalized linear models to identify factors associated with chlorophyll and microcystin concentrations and measured the effect of sediment resuspension on predictive performance.