Reliable Human-LM Collaboration under Uncertainty

Language Model (LM)-based AI systems have become ubiquitous tools in peoples' workflows across a variety of interaction modes: from assistants informing human decision-makers to autonomous agents functioning with minimal human oversight [8]. However, the reliability of these systems remains a concern. Unreliable LMs can mislead users and take undesirable actions that result in hard-to-detect silent failures and catastrophic downstream consequences. **My research improves the reliability of human-LM collaboration outcomes**, by quantifying and leveraging LM uncertainty, surfacing that uncertainty to users usefully, and actively modeling users' goals and preferences.

Quantifying and leveraging LM uncertainty: While uncertainty quantification (UQ) has traditionally been studied in classical (discriminative) ML models, attaining calibrated confidence estimates from generative LMs is less straightforward [3]. I have introduced methods for computing calibrated confidence scores for predictions [10] and explanations [2] from vision-language models (VLMs). I have also developed methods leveraging LM uncertainty to inform model reasoning and reflection. In a selective prediction setting, we mitigated unnecessary VLM abstention through ReCoVERR [10], an algorithm that verifies low-confidence VLM predictions by recovering high-confidence evidences in the image.

Surfacing uncertainty to users *usefully*: LM uncertainty can also signal when users should rely on model outputs; however, most UQ methods do not evaluate whether they actually foster appropriate reliance. In our position paper [1], we called for human uplift studies evaluating whether LM UQ methods improve user reliance in real-world tasks and whether standard calibration metrics are correlated with downstream utility to users. To this end, in He et al. [2] we demonstrated that our quality scores for VLM explanations improved user reliance when the user could not see the VLM's visual context.

Further, human factors also affect how users incorporate AI advice in decision-making. Our work [9] demonstrated that user trust is an important factor influencing whether users accept AI advice or not, with low and high levels of user trust resulting in increased under- and over-reliance, respectively. We further showed that adapting AI assistants' behavior in response to user trust levels can mitigate trust-induced inappropriate reliance. In two decision-making scenarios—laypeople answering science questions and doctors making medical diagnoses—we found that providing supporting and counter-explanations during moments of low and high trust, respectively, yields up to 38% reduction in inappropriate reliance and 20% improvement in decision accuracy.

Active user modeling: Effective human-LM collaboration requires LMs to model user characteristics, behaviors, and preferences to establish common ground. Some efforts have explored modeling users' personality traits [5] and mental states using Theory-of-Mind [6], yet many aspects of user modeling remain underexplored. In our work on studying user trust [9], we also trained models to predict user trust levels based on observable user-AI interaction features and found that they correlated *poorly* with actual user trust levels, highlighting the challenging nature of modeling users' cognitive states.

Although user preference modeling has been extensively studied in the dialog community (including in my SIGDIAL Best Paper work [7]), current AI systems trained for autonomous task completion still make implicit, and often incorrect, assumptions about user preferences. In my ongoing work, I am training LLM agents to model uncertainty about user preferences and appropriately insert positive friction [4] to resolve said uncertainties. As part of this work, I have conducted user studies to understand how user preferences evolve during collaboration, designed user-LM collaboration benchmarks that reflected realistic user goal dynamics, built LM user simulators reflecting observed user behaviors, and trained LMs to effectively collaborate with users using offline (SFT, DPO) and online (multi-turn RL) algorithms.

My experiences, experiments, and learnings during the Ph.D. have led me to multiple underexplored research directions: How can LMs learn in real time from user feedback and leverage noisy reward signals? Can we build better user simulators to reflect the diversity of real user behaviors rather than just the mode? How does the nature of collaboration change when going from task-centric interactions to open-ended ones (e.g. creative writing or mental health counseling)? I believe my research training during my Ph.D. has equipped me with the skills and the desire to tackle these challenging and important questions.

References

[1] Siddartha Devic, **Tejas Srinivasan**, Jesse Thomason, Willie Neiswanger, and Vatsal Sharan. From calibration to collaboration: Llm uncertainty quantification should be more human-centered. *arXiv preprint arXiv:2506.07461*, 2025.

[2] Keyu He, **Tejas Srinivasan**, Brihi Joshi, Xiang Ren, Jesse Thomason, and Swabha Swayamdipta. Believing without seeing: Quality scores for contextualizing vision-language model explanations. *arXiv preprint arXiv:2509.25844*, 2025.

[3] Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. Surface form competition: Why the highest probability answer isn't always right. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

[4] Mert Inan, Anthony Sicilia, Suvodip Dey, Vardhan Dongre, **Tejas Srinivasan**, Jesse Thomason, Gökhan Tür, Dilek Hakkani-Tür, and Malihe Alikhani. Better slow than sorry: Introducing positive friction for reliable dialogue systems. *Transactions of the Association of Computational Linguistics (TACL)*, 2025.

[5] Brihi Joshi, Xiang Ren, Swabha Swayamdipta, Rik Koncel-Kedziorski, and Tim Paek. Improving language model personas via rationalization with psychological scaffolds. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025.

[6] Michal Kosinski. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121, 2024.

[7] Shikib Mehri, **Tejas Srinivasan**, and Maxine Eskenazi. Structured fusion networks for dialog. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 165–177, 2019.

[8] Yijia Shao, Humishka Zope, Yucheng Jiang, Jiaxin Pei, David Nguyen, Erik Brynjolfsson, and Diyi Yang. Future of work with ai agents: Auditing automation and augmentation potential across the us workforce. *arXiv preprint arXiv:2506.06576*, 2025.

[9] **Tejas Srinivasan** and Jesse Thomason. Adjust for Trust: Mitigating Trust-Induced Inappropriate Reliance on AI Assistance. *ACM Conference on Intelligent User Interfaces (IUI)*, 2026.

[10] **Tejas Srinivasan**, Jack Hessel, Tanmay Gupta, Bill Yuchen Lin, Yejin Choi, Jesse Thomason, and Khyathi Raghavi Chandu. Selective "Selective Prediction": Reducing Unnecessary Abstention in Vision-Language Reasoning. In *Findings of the Association for Computational Linguistics (ACL)*, 2024.