Abhijnan Nath web: *https://www.abhijnannath.com/* abhijnan.nath@colostate.edu

Research Statement

Research Motivation: My research explores autoregressive language generation in Large Language Models (LLMs) [21], focusing on mitigating hallucinations that misalign with human preferences [26]. Despite their widespread adoption, LLMs continue to exhibit tendencies toward generating erroneous or irrelevant outputs—outputs that are often not "aligned" with human preferences or values. *On a high-level, my research attempts to make LLMs more user-friendly such that its outputs, on average, are more preferable to the end user*. Specifically, I aim to develop compute-efficient and robust preference-alignment (PA) algorithms[1]that capture the diversity of true human preferences. Additionally, given that human preferences in collaborative tasks are dynamic [5] and often intransitive [27], I am also interested in designing belief-driven "friction" agents that guide collaborators in task completion by prompting reflective thinking in participants. These efforts enhance the reliability and trustworthiness of LLMs, particularly in mission-critical applications like educative and learning environments. My work has also been sponsored by multiple DARPA programs and demonstrated relevance to mission-critical natural language understanding and human-AI interaction systems. I am also the recipient of the prestigious Evolutionary Computing and Artificial Intelligence Fellowship 2024, awarded annually by the Department of Computer Science, Colorado State University, for meritorious achievements in the area of artificial intelligence. Such recognition ignites my desire to question and learn more about these phenomena within artificial intelligence.

Current Research: Timeline[2] and Plan:

Jan '24 to May '24 I investigate preference alignment in LLMs along two key dimensions. Building on my prior research in knowledge transfer between model latent spaces [13, 14] and modalities [16, 28] and related works [8, 10], I adopt a *Chain-of-thought (CoT)* prompting [30] approach and explore how LLM-generated reasoning traces (free-text rationales or FTRs) can indirectly expose human preferences. This approach attemps to answer the question: **how can structured reasoning traces be optimized to reflect human preferences that can be leveraged for soft-supervision across textual clustering tasks?** Specifically, I examine whether such rationales serve as soft-labels or validation mechanisms for improved task performance. My recent published work [15, 19] applies this rationale-based knowledge transfer to coreference resolution in event descriptions and intervention clustering in collaborative dialogues. However, these methods do *not* modify LLM parameters directly, limiting their ability to ensure preference consistency in model outputs.

June '24 to Dec '24 To address these limitations, my second research direction focuses on *policy gradient* methods [31]—like Reinforcement Learning from Human Feedback (RLHF)—that explicitly update an LLM's parameters to ensure preferred outputs are more likely during stochastic sampling. Although popular, a major challenge in training RLHF-algorithms like Proximal Policy Optimization (PPO) [23] is their compute inefficiency. Additionally, even efficient alternatives like Direct Preference Optimization (DPO) [22] suffer from overfitting and policy degeneracy, exacerbated by the assumption that human preferences follow the Bradley-Terry model. Furthermore, real-world preferences are often nondeterministic, intransitive, and influenced by sampling biases [11, 3], making existing preference alignment methods suboptimal.

I answer this question: **Can supervised preference alignment methods be improved to handle the inherent diversity and inconsistency of human preferences without necessarily relying on assumptions like the Bradley-Terry model?** To this end, my work proposes Direct Reward Distillation and Policy Optimization (DRDO) [18] (under review for ICML 2025), which addresses reward-preference misalignment[11] by leveraging explicit rewards within a knowledge-distillation framework. Unlike standard supervised preference alignment algorithms such as DPO, DRDO models a joint distribution over prompts and responses, providing a more expressive representation than conditional formulations. By structuring the DRDO objective as a joint learning framework for both rewards and preferences, we mitigate misalignment issues that arise in the presence of non-deterministic or noisy preference data. Building on this approach, I proposed Diverse Preference Learning (DPL) [20] (accepted at

[1]Most of my preference alignment work [17, 20] follows *alignment-via-fine-tuning*, except [15], which explores *alignment-via-prompting*. The former fine-tunes LLMs for human-preferred outputs, while the latter optimizes prompts directly. In contrast, recent methods [1, 24] approximate MCMC distributions and rejection sampling [4], falling under *alignment-via-inference*.

[2]Since a lot of my research directions are being explored in parallel, there may be some overlap in timelines.

NAACL 2025) that intuitively captures nuances in preference feedback *without* requiring a explicit external reward model. Developed in collaboration with industry partners during my research internship at OptumAI, DPL was rigorously tested and validated within production-level pipelines, particularly in mission-critical domains like healthcare. More specifically, DPL enforces sample-level penalties in model training and models a "baseline desirability" alongside "relative preference strengths"— which help capture the diverse human preferences more effectively. My empirical studies suggest that both DRDO and DPL demonstrate superior generalization in alignment tasks such as instruction following, text-summarization as well as general question-answering—while being robust to both clear as well as non-deterministic preference samples.

Sept '24 to May '25 Unlike the broader preference alignment methods discussed earlier, my research also explores a more targeted problem with mission-critical applications in collaborative learning environments: how to design a preference-aligned friction agent that fosters accountability in collaborative goal-oriented dialogues. Here, "friction" refers to reflective interventions—textual generations from an LLM—that act as indirect persuasion, prompting participants to reassess their beliefs ("frictive states") and reflect during collaborative tasks, without the intervention directly offering hints that could bias task outcomes. The core challenge here is that LLMs are typically not trained to generate friction in this sense and collaborative dialogue annotations are typically sparse due to multimodal communication [6]. Standard approaches like DPO, though computationally efficient and scalable, assume a Bradley-Terry model of preferences and thereby suffer from a sampling or data-bias. When using generative AI to create denser training data, even high-capacity LLMs like GPT-4 are prone to various forms of biases such as toward length [7], sycophancy as well as conceptual bias [29]—that are not causally related to the preference label. Therefore, my research question is: **how do we train and evaluate a high-quality friction agent that can leverage the inherent scalability of offline alignment methods and reconstruct the true underlying preference distribution while still being robust to the data skew that may arise when sampling a preference dataset, whether using generative AI or from real-life collaborative dialogues?**

To address this, I propose the Frictional Agent Alignment Framework (`FAAF`), to generate precise, context-aware "friction" that prompts for deliberation and re-examination of existing evidence. As shown in Fig. 1, `FAAF`'s two-player objective decouples from data skew[2, 3]: a frictive-state policy (π_ϕ) identifies belief misalignments from dialogue history, while an intervention policy (π_f) crafts collaborator-preferred responses. The core insight here is that optimal friction interventions should *not* be arbitrary interventions in the dialogue, but should surface the presuppositions that gave rise to the most logically necessary frictive state, making interventions precise and interpretable. My research derives an analytical solution to this objective, enabling training a single policy via a simple supervised loss function. Our empirical results suggest that `FAAF`'s interventions are, on average, more aligned with human preferences compared to current approaches when measured by a high-capacity LLM-judge on task-specific preference desiderata like actionability, relevance, alignment with golden samples, etc. Notably, this work is currently under review for ACL 2025 and was recently presented at the DARPA's **Friction for Accountability in Conversational Transactions (FACT)** Artificial Intelligence Exploration (AIE) program meeting in Stanford University.

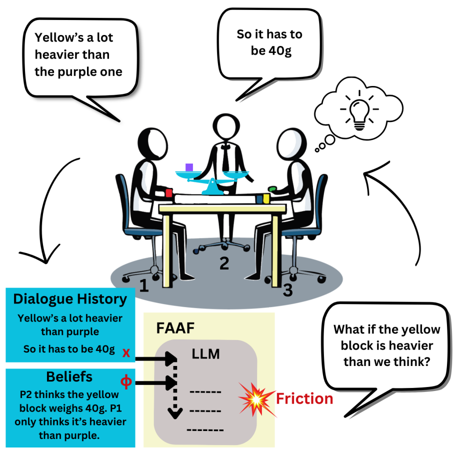

Figure 1: `FAAF` conditions responses on both the dialogue context x and representation of the "frictive" (belief) state ϕ, generating outputs that prompt for reflection, deliberation, and verification of evidence.

Feb '25 to May '25 As for a more robust evaluation method, my proposed plan is to compare `FAAF` with state-of-the-art approaches like Group Relative Policy Optimization [25] (algorithm that powers Deepseek R1's success) in more dynamic role-play settings [9] over multiple turns, where we can robustly and scalably test these approaches through API-based dialogue simulations. Specifically, LLM agents aligned with these approaches will be evaluated in alternating turn-based dialogue simulations between agents and high-capacity AI collaborators in Weights Task [6] and Delidata environments [5] and evaluated on metrics like long-term effectiveness of friction interventions, quality over multiple turns, as well as proxy measures of persuasiveness such as dialogue length and successful task completion rates.

May '25 to Dec '25 Additionally, I plan to explore inference-time alignment algorithms that approximate policy distributions like Best-of-N [12] or Markov Chain Monte Carlo (MCMC) [4] and develop novel strategies to optimally train friction agents for Distributed Partial Information (DIP) Tasks in collaborative settings. These tasks are currently being conducted and recorded in the Signal Lab, Colorado State University for the lego-block building domain. This brings an additional challenge in LLM alignment since state information (or lego-block structure) is only partially observed by collaborators. As such, I intend to model the optimal agent behavior as a solution to a Partially Observable Markov Decision Process (POMDP), which could better account for uncertainty in participant belief states and the latent task state. Furthermore, DIP tasks would likely require incorporating visual information alongside text, enabling more effective interventions in tasks with physical components.

Conclusion:

With my research plan and past peer-reviewed contributions, I believe I am a strong candidate for the Wim Böhm and Partners Award or the P. R. Mukherjee Award, which would provide crucial support as I complete the final phase of my PhD research. Having successfully passed my preliminary examination, I am now focused on addressing fundamental challenges in AI alignment—particularly in shaping AI systems as "thought partners" in human-AI interactions rather than mere "instruction followers." My research contributes to advancing preference alignment and responsible AI development, ensuring models better reflect diverse human values. With a strong track record of publications in top AI conferences, extensive collaboration with interdisciplinary teams, and experience mentoring junior researchers, I am confident that this fellowship will enable me to further my contributions to AI research

References

[1] Afra Amini, Tim Vieira, and Ryan Cotterell. Variational best-of-n alignment, 2024. URL `https://arxiv.org/abs/2407.06057`.

[2] Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR, 2024.

[3] Eugene Choi, Arash Ahmadian, Olivier Pietquin, Matthieu Geist, and Mohammad Gheshlaghi Azar. Robust chain of thoughts preference optimization. In *Seventeenth European Workshop on Reinforcement Learning*, 2024.

[4] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. 1970.

[5] Georgi Karadzhov, Tom Stafford, and Andreas Vlachos. Delidata: A dataset for deliberation in multi-party problem solving. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2): 1–25, 2023.

[6] Ibrahim Khalil Khebour, Kenneth Lai, Mariah Bradford, Yifan Zhu, Richard A. Brutti, Christopher Tam, Jingxuan Tu, Benjamin A. Ibarra, Nathaniel Blanchard, Nikhil Krishnaswamy, and James Pustejovsky. Common ground tracking in multimodal dialogue. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3587–3602, Torino, Italia, May 2024. ELRA and ICCL. URL `https://aclanthology.org/2024.lrec-main.318/`.

[7] Nathan Lambert, Valentina Pyatkin, Jacob Daniel Morrison, Lester James Validad Miranda, Bill Yuchen Lin, Khyathi Raghavi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hanna Hajishirzi. RewardBench: Evaluating reward models for language modeling. *ArXiv*, abs/2403.13787, 2024.

[8] Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 991–999, 2015.

[9] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008, 2023.

[10] David McNeely-White, Benjamin Sattelberg, Nathaniel Blanchard, and Ross Beveridge. Exploring the interchangeability of cnn embedding spaces. *arXiv preprint arXiv:2010.02323*, 2020.

[11] Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, et al. Nash learning from human feedback. *arXiv preprint arXiv:2312.00886*, 2023.

[12] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

[13] Abhijnan Nath. *Linear Mappings: Semantic Transfer from Transformer Models for Cognate Detection and Coreference Resolution*. PhD thesis, Colorado State University, 2022.

[14] Abhijnan Nath, Sheikh Mannan, and Nikhil Krishnaswamy. AxomiyaBERTa: A phonologically-aware transformer model for Assamese. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11629–11646, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.739. URL `https://aclanthology.org/2023.findings-acl.739`.

[15] Abhijnan Nath, Shadi Manafi Avari, Avyakta Chelle, and Nikhil Krishnaswamy. Okay, let's do this! modeling event coreference with generated rationales and knowledge distillation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3931–3946, 2024.

[16] Abhijnan Nath, Huma Jamil, Shafiuddin Rehan Ahmed, George Arthur Baker, Rahul Ghosh, James H Martin, Nathaniel Blanchard, and Nikhil Krishnaswamy. Multimodal cross-document event coreference resolution using linear semantic transfer and mixed-modality ensembles. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11901–11916, 2024.

[17] Abhijnan Nath, Changsoo Jung, Ethan Seefried, and Nikhil Krishnaswamy. Simultaneous reward distillation and preference learning: Get you a language model who can do both. *arXiv preprint arXiv:2410.08458*, 2024.

[18] Abhijnan Nath, Changsoo Jung, Ethan Seefried, and Nikhil Krishnaswamy. Simultaneous reward distillation and preference learning: Get you a language model who can do both, 2024. URL `https://arxiv.org/abs/2410.08458`.

[19] Abhijnan Nath, Videep Venkatesha, Mariah Bradford, Avyakta Chelle, Austin Youngren, Carlos Mabrey, Nathaniel Blanchard, and Nikhil Krishnaswamy. "any other thoughts, hedgehog?" linking deliberation chains in collaborative dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5297–5314, 2024.

[20] Abhijnan Nath, Andrey Volozin, Saumajit Saha, Albert Aristotle Nanda, Galina Grunin, Rahul Bhotika, and Nikhil Krishnaswamy. Dpl: Diverse preference learning without a reference model. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2025)*, January 2025. Last modified: 24 Feb 2025.

[21] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[22] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2023.

[23] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[24] Pier Giuseppe Sessa, Robert Dadashi, Léonard Hussenot, Johan Ferret, Nino Vieillard, Alexandre Ramé, Bobak Shariari, Sarah Perrin, Abe Friesen, Geoffrey Cideron, et al. Bond: Aligning llms with best-of-n distillation. *arXiv preprint arXiv:2407.14622*, 2024.

[25] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[26] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

[27] Amos Tversky. Intransitivity of preferences. *Psychological review*, 76(1):31, 1969.

[28] Videep Venkatesha, Abhijnan Nath, Ibrahim Khebour, Avyakta Chelle, Mariah Bradford, Jingxuan Tu, James Pustejovsky, Nathaniel Blanchard, and Nikhil Krishnaswamy. Propositional extraction from natural speech in small group collaborative tasks. In *Proceedings of the 17th International Conference on Educational Data Mining*, pages 169–180, 2024.

[29] Chaoqi Wang, Zhuokai Zhao, Yibo Jiang, Zhaorun Chen, Chen Zhu, Yuxin Chen, Jiayi Liu, Lizhu Zhang, Xiangjun Fan, Hao Ma, et al. Beyond reward hacking: Causal rewards for large language model alignment. *arXiv preprint arXiv:2501.09620*, 2025.

[30] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

[31] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, pages 229–256, 1992.