Yunze Xiao web: https://algoroxyolo.github.io/ yunzex@andrew.cmu.edu

Statement of Purpose

of Yunze Xiao

When Alan Turing proposed his famous test in 1950, he envisioned a day where machines could think [13]. Today, that question seems almost cliché. Large language models do not just pass for humans [10]; they perform empathy [8], craft personalities [14], and even simulate moral reasoning with striking fluency [11]. The challenge has shifted from teaching machines to mimic us to deciding how human we want them to appear. Should AI express uncertainty? Display warmth? Claim preferences? These design choices matter: too little humanity creates unusable tools; too much risks deception and erodes genuine human connection. These are no longer philosophical puzzles, but engineering decisions unfolding in labs worldwide, choices that shape both human-machine interaction and our self-understanding.

Broadly, I want to develop large language models that embody genuine human-like intelligence, systems that not only generate fluent language but also **think**, **remember**, **feel**, and **interact** in ways that mirror human cognitive and social capabilities. I envision LLMs as intuitive collaborators. These would be systems that understand context with human-like nuance, adapt their responses based on experience, and engage with genuine emotional intelligence. Rather than matching patterns, these models would exhibit coherent personalities, maintain consistent mental models, and participate in the rich, situated interactions that define real human communication. In sum, my research pursues three intertwined directions:

1. Formalizing **anthropomorphism as a modeling dimension** to understand how training signals and interface choices shape human-like behaviors.
2. Designing **architectural innovations** that support richer forms of humanlikeness.
3. Develop **task-oriented augmentation techniques** that take advantage of anthropomorphic qualities to improve downstream performance.

Anthropomorphism as Modeling Dimension As large language models become increasingly fluent, the boundary between tool and companion blurs and every design tweak becomes a choice about what it means to be 'human-like'. But how do we decide how much humanity to build in, and what exactly are we dialing up or down?

First, **anthropomorphism needs to be redefined in the LLM era**. Despite extensive prior research, simply defining anthropomorphism as a one-way projection of human traits onto machines is no longer adequate in the age of large language models. Today's LLMs do more than passively receive human attributes; they actively exhibit sophisticated behaviors, interact in complex social contexts, and initiate signals that reshape user perceptions and reactions. In my position paper [16], we reconceptualize anthropomorphism as a reciprocal phenomenon between designers and interpreters, mediated through four dimensions of the cue (perceptual, linguistic, behavioral and cognitive). Moreover, our multidimensional framework that not only defines these cues but also evaluates their effectiveness and proposes actionable design strategies to calibrate anthropomorphic elements to match artifact capabilities and user expectations. This turns the vague question How human-like is it? into a coordinate system that any model, prompt, or interaction trace can be projected onto and compared within.

As a result, **we need a new generation of toolkits to comprehensively measure anthropomorphism**. Existing evaluations are overwhelmingly focused on linguistic cues [3, 4], overlooking critical perceptual, behavioral, and cognitive dimensions that shape human-like interaction. There is an urgent need for benchmarks that capture the full spectrum of anthropomorphism, spanning individual dimensions, mixed or interacting cues, and diverse contextual settings, such as cultural or application-specific scenarios. In my own work, for example, *InCharacter* [14] introduced an interview-based evaluation that probes the ability of LLMs to maintain consistent personalities and psychological traits throughout extended interactions, demonstrating how richer and more psychologically grounded benchmarks can reveal strengths and limitations missed by surface-level linguistic analysis. Such holistic evaluations are essential not only for tracking model progress, but also for guiding principled, context-aware design choices as LLMs become more deeply embedded in our daily lives.

Most importantly, **anthropomorphism should move beyond risk-oriented formulation**. Previous discussions have focused heavily on its risks [2, 9, 1, 7, 6, 5, 12], fostering a cautious climate and calling for systematic deanthropomorphism. This has real consequences: IRBs and ethics panels often adopt overly conservative stances that slow approval and discourage innovation, researchers can self-censor, and public discourse is often driven by mistrust or panic. These dynamics have stifled a rigorous exploration of anthropomorphism's functional benefits, such as improved usability or calibrated trust. Instead of advocating for deanthropomorphism, we need evidence-based frameworks that identify when human-like characteristics improve results and develop guidelines for responsible implementation. Studying anthropomorphism as a design tool rather than an inherent threat can harness its potential while addressing legitimate concerns.

Together, these advances transform anthropomorphism from a risk to avoid into a design space to navigate. My framework provides the coordinates, comprehensive benchmarks offer the measurements, and empirical

studies chart the path forward. This systematic approach enables us to build AI systems calibrated for their intended purpose: human-like where it helps, transparent where it matters.

Anthropomorphic Embodiment What does it mean for a machine to have a personality, to show emotion, or to remember? These traits are not just surface features.They are the building blocks of how we relate, trust, and collaborate with each other. When LLMs begin to display distinct **personalities**, react with apparent **emotion**, they move beyond mere conversation and begin to feel like real partners in the interaction.

A central challenge in anthropomorphic LLMs is building coherent personas that enable consistent perspectives and meaningful collaboration. Current models struggle with persona stability, fragmenting across contexts. My research addresses this through two approaches: developing methods for LLMs to recognize and reason about MBTI personality patterns in human communication, and conducting the first large-scale study of multidimensional personas in AI debates [11]. This work revealed how ideology and personality attributes produce emergent argumentation patterns mirroring human social psychology: conservative agents favor authority-based arguments while neurotic agents hedge positions, demonstrating that authentic personas emerge from psychological scaffolding, not surface prompting.

Another critical dimension is enabling LLMs to express emotions. Current models, constrained by overcautious safety alignment, fail as companions and produce unrealistic simulations populated by purely rational actors. To excel in tasks that require social intelligence, LLMs must understand and display emotions at calibrated levels [16]. My work integrating the EMA appraisal model into LLM agents demonstrates how emotional states can shape behavior: agents experiencing frustration exhibit shortened responses and increased aggression, mirroring human patterns. Future work will pursue multimodal emotional consistency, where omni-models express coherent emotions across text, voice, and visual generation, creating truly integrated emotional experiences rather than disconnected outputs.

These embodiment dimensions, personality, and emotion, transform LLM from tools into collaborators. My previous works showed that stable personas and emotional reasoning enable a genuine partnership in complex tasks. However, this requires carefully designed architectures that ensure personality persistence, cross-modal emotional consistency, and coherent evolution of both dimensions. These foundations enable more effective human-AI collaboration and new collaborative possibilities

Anthropomorphism as a downstream approach Although my theoretical framework provides the foundation for understanding anthropomorphism, its true value emerges in the addressing of critical human challenges. By strategically calibrating anthropomorphic features, we can create AI systems that not only understand human complexity but actively improve human well-being and social understanding.

Mental health support represents one of the most promising yet sensitive applications of anthropomorphic AI. Jiraibench demonstrates how LLMs with domain-specific encodings can detect risky health behaviors in naturalistic online conversations[15]. Our system leverages anthropomorphic understanding that recognizes not just keywords but also emotional context, social dynamics, and culture patterns that indicate distress and unhealthy behaviors. Building on this detection capability, I envision anthropomorphic companion agents that provide timely and calibrated support. Unlike current LLMs that maintain emotional distance through safety alignment, these systems would employ my four-dimensional framework to offer appropriate warmth and empathy while maintaining therapeutic boundaries. For example, they might express genuine concern through linguistic cues while avoiding behavioral oversteps that could create unhealthy dependencies.

Anthropomorphic embodiment transforms **LLM-based simulations** from rational actor models to psychologically realistic social systems. My work on emotion simulation, which integrated the EMA appraisal model into LLM agents, demonstrated how emotional states shape agent behavior in ways that mirror human psychological patterns. When agents experience 'frustration' from repeated goal blocking, they exhibit shortened responses, increased negotiation aggression, and preference shifts, behaviors absent in purely rational simulations.This emotional grounding enables unprecedented fidelity in modeling complex social phenomena.

These applications demonstrate that anthropomorphism is not just aesthetic **but functional**. By embedding human-like cognitive and emotional processes into AI systems, we can better understand human behavior and create interventions that resonate with human psychology.

[Why School]

References

[1] Gavin Abercrombie, Amanda Cercas Curry, Vedant Dinkar, Verena Rieser, and Zeerak Talat. Mirages: On anthropomorphism in dialogue systems. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

[2] J. Cheng, K. Gligoric, T. Piccardi, and D. Jurafsky. Anthroscore: A computational linguistic measure of anthropomorphism. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2024.

[3] Myra Cheng, Kristina Gligoric, Tiziano Piccardi, and Dan Jurafsky. AnthroScore: A computational linguistic measure of anthropomorphism. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 807–825, St. Julian's, Malta, March 2024. Association for Computational Linguistics.

[4] Myra Cheng, Sunny Yu, and Dan Jurafsky. Humt dumt: Measuring and controlling human-like language in llms, 2025.

[5] Michelle Cohn and et al. Believing anthropomorphism: Examining the role of anthropomorphic cues on trust in large language models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 2024.

[6] Rahul Deshpande, Naman Rajpurohit, Karthik Narasimhan, and Abhishek Kalyan. Anthropomorphization of ai: Opportunities and risks. In *Proceedings of the 1st Workshop on Natural Legal Language Processing (NatLLP)*, 2023.

[7] Benjamin Gros, Xisen Li, and Mo Yu. Robots-dont-cry: Understanding falsely anthropomorphic utterances in dialog systems. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

[8] Jen-Tse Huang, Man Ho Lam, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael R. Lyu. Emotionally numb or empathetic? evaluating how llms feel using emotionbench, 2024.

[9] Yasmin Ibrahim and J. Cheng. Thinking beyond the anthropomorphic paradigm benefits llm research. 2025.

[10] Cameron R. Jones and Benjamin K. Bergen. Large language models pass the turing test, 2025.

[11] Jiarui Liu, Yueqi Song, Yunze Xiao, Mingqian Zheng, Lindia Tjuatja, Jana Schaich Borg, Mona Diab, and Maarten Sap. Synthetic socratic debates: Examining persona effects on moral decision and persuasion dynamics, 2025.

[12] Sandra Peter, Kai Riemer, and Jevin D. West. The benefits and dangers of anthropomorphic conversational agents. *Proceedings of the National Academy of Sciences*, 122(22):e2415898122, 2025.

[13] Alan Turing. Computing machinery and intelligence. *Mind*, 59(236):433–60, 1950.

[14] Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1840–1873, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[15] Yunze Xiao, Tingyu He, Lionel Z. Wang, Yiming Ma, Xingyu Song, Xiaohang Xu, Irene Li, and Ka Chung Ng. Jiraibench: A bilingual benchmark for evaluating large language models' detection of human self-destructive behavior content in jirai community, 2025.

[16] Yunze Xiao, Lynnette Hui Xian Ng, Jiarui Liu, and Mona Diab. Humanizing machines: Rethinking llm anthropomorphism through a multi-level framework of design, 2025.