1 Vision

Agents take in user instructions and ground them as commands executed on machines. Recent advancements in agents are driven by foundational models (LLMs) trained on internet-scale data [1, 2]. These agents can take in naturalistic instructions such as language, drawings [3], and examples [4], and output machine instructions for code generation [5, 6], robotics [7], and creative tasks [8, 9]. With increasing labor costs and an aging demographic, agents are projected to play a key role in our society [10] as human collaborators.

However, unlike humans who learn through interactions, LLM-based agents improve primarily through reengineering [11] and larger datasets [12, 13]. Consequently, end-users struggle to improve the capabilities of these agents through interaction [14], but must rely on companies such as OpenAI to create the next version of GPT. Why is it so difficult to create agents that improve through interactions? I identify two challenges:

(1) Data Gap. Typical instruction datasets are text-only, lacking an embodied environment, specific tasks, and evaluation metrics to verify execution success.

(2) Theory Gap. Existing theories of human instruction are constrained to overly stilted tasks and remain unproven in accounting for the complexity of modern interactive agents.

As a consequence of the lack of (1) data and (2) theory, we are left with training ever larger models, with interaction as an "emergent property" [1, 15], rather than an objective.

My lab's goal is **building agents that learn from human interactions**, with the challenges of data and theory addressed as prerequisites. These challenges are made tractable by the following trends: (a) The increasing ease of collecting human annotations via crowdsource; (b) The increasing generalities of computational cognitive science in explaining human data, and (c) the increasing commodification of (code generating) foundational models that readily integrate multi-modal datasets. I center my research aims around these observations, they are:

Aim1. Dataset Curation To establish "north star" **human instruction datasets** to evaluate and build interactive agents [1].

Aim2. Cognitive Modeling To develop domain general, computational cognitive theories of human communication that account for the curated datasets.

Aim3. Interactive Systems To build interactive agents instantiated with these cognitive theories of human communication.

Together, these objectives will push AI systems beyond their current limitations of relying on significant expert efforts to improve. Instead, they will grow organically from user interactions, becoming capable collaborators within specific domains (Figure 1).

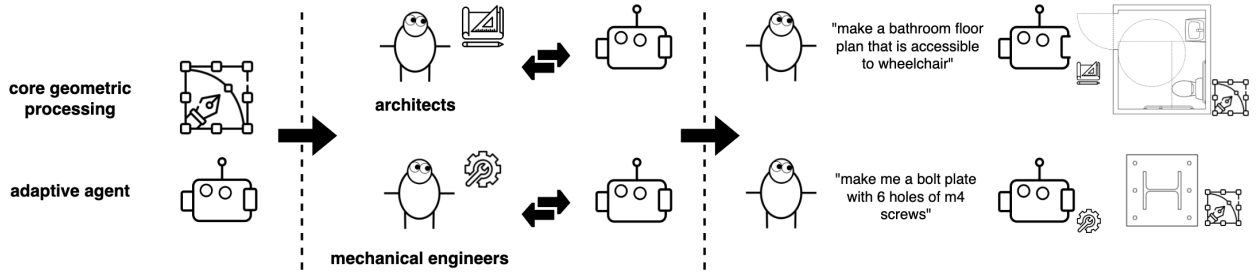

Figure 1: Aspirational example of building agents that learn from human interactions to solve domain specific tasks. Rather than engineering two separate systems for architects and mechanical engineers, the developers build only a core API for geometric processing and an interactive agent, who learns domain-specific capabilities from interactions.

[1] that are enduring despite the advancement of LLM agents (e.g. GPT-x)

2 Aims

Our goal is to discover general cognitive principles that underpins human communication and use these insights to build interactive systems, rather than developing bespoke systems. To this end, we propose the following interrelated aims: (1) curate datasets, (2) develop theories, and (3) build interactive systems (Figure 2).

Aim 1: Dataset Curation Machine learning is driven by datasets, for instance, the success of deep learning was spurred largely by ImageNet [16]. Yet, high-quality and large-scale human instruction-following datasets are scarce. This lack results in a proliferation of agents that seem impressive in demos but lack benchmarks to quantify their usefulness. we intend to build instruction datasets with full specification of the underlying environment, task, and clear evaluation metric for execution success, similar to the pioneer works of [17] but at a larger scale. By controlling these parameters, we can (1) better model the *situation* and *tasks* under which the instructions are given, and (2) provide a framework of automatic evaluation of instruction following agents via execution. Overall, these datasets should expose fundamental phenomena of human communication, rather than for build any specific agent.

Aim 2: Cognitive Modeling Existing cognitive models of human communication [18, 19] tends to focus on the communication of references (i.e. *this* object rather than *that* one), rather than on instructions (i.e. to perform actions). This is partly due to the lack of large-scale, well controlled datasets to develop these models on top of. With Aim1, we are in a unique position to model the collected datasets using Bayesian models [20, 21, 22]. In addition to fitting the data by giving probabilities, these models are also *generative*, making it possible for sample these models to obtain instructions and action sequences, giving a concrete mechanism for building instruction following agents. Overall, this aim will provide a computational account of human instruction, and is applicable across multiple domains.

Aim 3: Interactive Systems Current methods for building agents are often short-lived and bespoke, relying heavily on black-box, closed-source models. In contrast, we intend to build systems around the theories developed in Aim2. Specifically, models of how humans generate instructions can be used to guide synthetic data generation [23, 24], to ensure the synthetic data (and consequently, the model trained on it) is "human-like". These models can also be used at inference time, by performing theory of mind reasoning on why a human give a certain instructions, the agent has a high probability of inferring the user's true intent [25, 26]. Overall, this aim will build and evaluate instruction following agents in realistic domains (e.g. interactive CAD modeling), and make our work visible to the broader community.

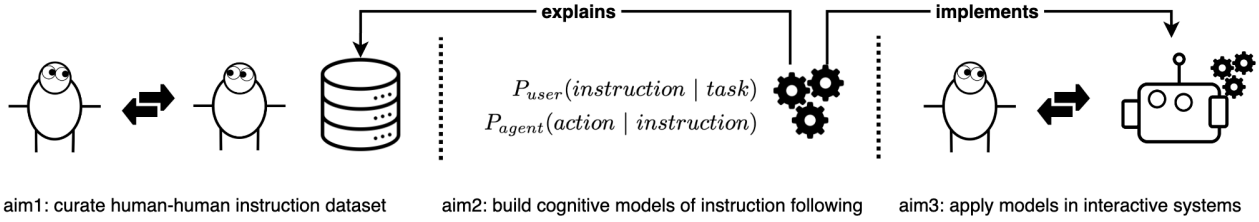

Figure 2: The 3 aims: curating dataset, developing models, and building interactive systems.

References

[1] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[2] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024.

[3] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[4] Wen-Ding Li and Kevin Ellis. Is programming by example solved by llms?, 2024. URL https://arxiv.org/abs/2406.08316.

[5] Frank F Xu, Uri Alon, Graham Neubig, and Vincent Josua Hellendoorn. A systematic evaluation of large language models of code. In *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming*, pages 1–10, 2022.

[6] Lingli Wang, Ni Huang, Yili Hong, Luning Liu, Xunhua Guo, and Guoqing Chen. Voice-based ai in call center customer service: A natural field experiment. *Production and Operations Management*, 32(4): 1002–1018, 2023.

[7] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9493–9500. IEEE, 2023.

[8] Ka-Hei Hui, Aditya Sanghi, Arianna Rampini, Kamal Rahimi Malekshan, Zhengzhe Liu, Hooman Shayani, and Chi-Wing Fu. Make-a-shape: a ten-million-scale 3d shape model. In *Forty-first International Conference on Machine Learning*, 2024.

[9] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

[10] Andrzej Cichocki and Alexander P Kuleshov. Future trends for human-ai collaboration: A comprehensive taxonomy of ai/agi using multiple intelligences and learning styles. *Computational Intelligence and Neuroscience*, 2021(1):8893795, 2021.

[11] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*, 2024.

[12] Deepak Soekhoe, Peter Van Der Putten, and Aske Plaat. On the impact of data set size in transfer learning using deep neural networks. In *Advances in Intelligent Data Analysis XV: 15th International Symposium, IDA 2016, Stockholm, Sweden, October 13-15, 2016, Proceedings 15*, pages 50–60. Springer, 2016.

[13] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[14] Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of llm agents. *arXiv preprint arXiv:2402.17753*, 2024.

[15] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[17] Alane Suhr, Claudia Yan, Charlotte Schluger, Stanley Yu, Hadi Khader, Marwa Mouallem, Iris Zhang, and Yoav Artzi. Executing instructions in situated collaborative interactions. *arXiv preprint arXiv:1910.03655*, 2019.

[18] Robert XD Hawkins, Noah D Goodman, and Robert L Goldstone. The emergence of social norms and conventions. *Trends in cognitive sciences*, 23(2):158–169, 2019.

[19] Herbert H Clark and Deanna Wilkes-Gibbs. Referring as a collaborative process. *Cognition*, 22(1):1–39, 1986.

[20] Joshua Tenenbaum. Bayesian modeling of human concept learning. *Advances in neural information processing systems*, 11, 1998.

[21] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253, 2017.

[22] Noah D Goodman and Michael C Frank. Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, 20(11):818–829, 2016.

[23] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023.

[24] Panos Achlioptas, Ian Huang, Minhyuk Sung, Sergey Tulyakov, and Leonidas Guibas. Shapetalk: A language dataset and framework for 3d shape edits and deformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12685–12694, 2023.

[25] Jacob Andreas and Dan Klein. Reasoning about pragmatics with neural listeners and speakers. *arXiv preprint arXiv:1604.00562*, 2016.

[26] Daniel Fried, Jacob Andreas, and Dan Klein. Unified pragmatic models for generating and following instructions. *arXiv preprint arXiv:1711.04987*, 2017.